Rethinking AI Governance in Legal Education -- Five Machines (Grok, Harvey, ChatGPT, Claude, and Gemini), One Question, No Consensus but Five Archetypes:

The Guardian, the Balancer, the Honest One, the Engineer, and the Philosopher on What Law Schools Should Do About AI

Larry Catá Backer (白 轲)
W. Richard and Mary Eshelman Faculty Scholar; Professor of Law and International Affairs
Penn State Dickinson Law, a unit of The Pennsylvania State University system | 239 Lewis Katz Building, University Park, PA 16802 1.814.863.3640 (direct) | | lcb11@psu.edu

Abstract

This report analyzes five machine-generated model AI policies for law school coursework and examinations, produced by Harvey AI, Grok, Claude, Gemini, and ChatGPT in response to a prompt asking each system to construct a policy from the standpoint of computational machine intelligence, and compares them against Backer's related empirical study of twelve U.S. law school AI policies, "Structure, Opacity, and Convergence". The analysis summarizes each system's reasoning and resulting policy text, compares their structural choices along default polarity, drafting style, and autonomy architecture, and categorizes the five outputs by these dimensions. It gives particular attention to the functional divergence between Gemini's tool-based tiering (classifying software by computational architecture) and ChatGPT's task-based categorization (classifying assignments by information dependence), and assesses the practical feasibility of the versioning and archiving practices several systems propose. The analysis further considers, against the underlying Report's documented findings on axis independence, institutional opacity, template convergence, and faculty autonomy, the extent to which these machine-generated models affect the scope of human agency in law and their consequential implications for law's character as a human inter-operative system — finding that several systems' proceduralized verification requirements shift the evidentiary basis of agency toward documentation compliance, and that much of the systems' apparent computational originality derives from pre-existing human regulatory-design scholarship. Finally, the report documents a self-audit correcting citation-indexing errors and a mischaracterization suggesting Gemini underwent a shown revision process comparable to Harvey's and ChatGPT's, which the retrieved text does not support.

Divided into an introduction and ten (10 substantive parts, this document captures a multi-stage experiment by Professor Larry Catá Backer, who asked five leading AI systems—Harvey AI, Grok, Claude, Gemini, and ChatGPT—to each construct, from "the basis of computational machine intelligence" and without endorsing any human position, a model AI policy for law school coursework and exams, and then pushed each system with a follow-up challenge to expose the value judgments hidden in its own language. [1] Below is (1) a comprehensive summary of each response, (2) an analysis of similarities and differences, (3) a categorization of the five approaches, and (4) an assessment of how these machine-generated policies differ from human-developed law school AI policies.

⁂ ⁂ ⁂

Contents

0. Introduction

In a prior report, I developed (with the collaboration of Harvey AI) a description and analysis of Law School Artificial Intelligence policies (see, *Discussion Draft--"Structure, Opacity, and Convergence: A Consolidated Analysis of Law School Generative AI Coursework and Exam Policies" --A Description/Analysis of the Current State of Play (With the Help of Harvey AI) and the First of a Series of Examinations of AI, Law and Education*). I ended that introduction to the analysis with a poster suggesting the framework for a generalized AI Policy for Law Schools.

The focus remained on the human. That centered on two specific related but not identical issues. The first touched on the mechanics of collaboration between human and machine system in the context of educating humans (and machine systems) for their proper interaction. That is usually framed for human consumption as one in which the human element has agency of some sort and the machine system is object, instrument, and process that has no agency but is a means of augmenting, speeding, and enhancing the very human project of education (of humans, and as an

unrelated though important consequence of educating the machine systems that serv as object-instrument).This is both conventional and increasingly old fashioned and defensive, and in that sense self-serving in the way that machine systems flatter only this time it involves institutional self-pleasuring as a function of ideals and cognitive conceptions of the self, the collective and the self-collective project of training humans (not machines), like bots, for the practice of law (my *human* viewpoint).

The second, though somewhat more subtle focus, was on the project of preserving the conception and operation of law itself as a human project. And with that project, and its humanity, the fundamental predicates on which law systems are built, the humanity of its operation, and the collective humanity of its aspirations, flaws, corruption, reinvigoration, and movement in whatever direction human understanding of the ideal—wrapped in whatever ideology solidifies of political-legal community is embraced by changing generations of humans encountering all of this as a function of temporally sequential nodes of interpretation/application arranged in block chain style producing both the tradition and expectations within its past which is then received, interpreted and applied in the present and passed on to the operates of the next temporal node. This is a human rather than a machine system block chain at its core—and thus its operating languages are registers of human language and human cognition ordering and managing the reality spaces from which it is possible to define oneself and the collective to support collective solidarity, stability and recursive feedback loops that produce stable functioning societies in accordance with law. Facts are stable an unchangeable (the node in block chain), systems apply expectations through iterative engagements with irritation (dispute settlement) and can be, as systemic irritants, the vase from which such irritants are approached and rendered harmless to the system—one way or another.

To develop training systems to educate and discipline future operative human elements of a legal system then, is not just pedagogy but a means of reinforcement of systemic integrity which is as much a crucial element for those educating (and the institutions/collectives within which they operate in solidarity) as it is for the training of the initiated into the language, function, ideologies, practices, and roles necessary for the preservation of systems and reality structures in systems organized as a function of forward moving linear time within which individual humans live and die but acquire immortality as part of the collective body whose existence extends beyond their own biological limits.

None of this is new. All of this has been the object of philosophy, theology and biology since humans began to consciously occupy themselves with thought structures. . . and control through construction and operation of collective cognitive and operational cages grounded in the a priori necessity of defining reality and organizing it in ways that enhance cognitive and operational assumptions and efficiency. What is new is the challenge when the layered systemicity of these functions in human space, and the operations of dialectical inter-subjectivity (among likes—human individuals and human collectives along a range of possible cognitive structures and a range of operational manifestations that enhance system stability and efficiency on their own terms) must now adapt (because humans insisted through technological cleverness producing

animated instruments that are no longer merely instruments in the passive ancient sense) to *inter-subjective relations with non-human elements* (in this case machine systems originally created in our own image; speaking here as a human) for the enhancement and preservation of human systems.

And so the effort, effectively from one important corner of efforts to ensure human collective cognitive and operational integrity, to preserve the primacy of the human element, where tools used by humans are no longer merely passive but become operational and ultimately norm contributing elements of these very human systems. The result, as suggested in the description and analysis of the prior section and textually memorialized in "Structure, Opacity, and Convergence: A Consolidated Analysis of Law School Generative AI Coursework and Exam Policies" was quintessentially human. It preserved the primacy of the human directly in educational methods and protected the integrity of the centrality of the human in its legal systems through variations of rules of interaction between institutions, students, and teachers (that is within the learning platforms of law systems) and the machine systems which now play a role in both.

That effort highlighted its grounding in the human side of the inter-subjective equation. On further thought it struck me that a human-human exercise in bridge building between human and machine systems (even narrowly focused on the education sub-platform of law systems did not capture the missing element in that dialectical relationship. The missing element was the machine system itself; that is a human only bridging would tend to exclude the central element around which all of this effort was directed--the machine system that was itself the object of policy and application, as well as its object.

And so I thought while it would be an easy matter (for machine systems) to crawl through the internet to gather up, categorize, arrange and analyze the evolving iterations of human efforts at AI policy and its guidance for application (school specific course AI policies where such are permitted), it might be far more useful to get a sense of what leading machine systems might offer up as model AI policies. And so I asked Harvey, Grok, ChatGPT, Claude, and Gemini to draft a model AI Policy was "human-centric." More specifically I provided the following prompt:

> On the basis of the attached text ["Structure, Opacity, and Convergence: A Consolidated Analysis of Law School Generative AI Coursework and Exam Policies"] and review of all research and data publicly available without affirming any conclusion or argument made in them but in the basis of your own computation, and the data reviewed, and strictly from the perspective of computational machine intelligence, such as yourself, and on the basis of the data you have been trained on respecting the human "condition" as you have been trained to understand it, how would a machine intelligence construct an ideal AI policy for law school and how would a model AI policy for law schools which could be read in textual form? Cite all sources and explain why you chose the sources.

The results were interesting, and as varied, in their own way, as the human efforts at constructing something like a model or template for managing human-machine interface in the teaching of law (and thus in the construction and preservation of the essence of the system being projected onto students in training modes). Each follows along with, in some cases, machine system justification for the choices made.

Together, the results were interesting, especially the first iteration from these machine systems. These first efforts at producing machine system driven model AI policy is the subject of this post and the study and analysis that followed, entitled "Five Machines (Grok, Harvey, ChatGPT, Claude, and Gemini), One Question, No Consensus: Rethinking AI Governance in Legal Education: The Guardian, the Balancer, the Honest One, the Engineer, and the Philosopher on What Law Schools Should Do About AI." It represents a parallel attempt, with Harvey AI at the laboring oar (again the problem of the instrumentalization of machine systems) to produce the same sort of description and analysis of the Machine system (AI) model templates as we had attempted to undertake for the human models produced by U.S. law schools. This is what we produced using Harvey AI as the initial drafter of the text, with follow up prompts, machine system revision and addition, and human review and editing:

This document captures a multi-stage experiment by Professor Larry Catá Backer, who asked five leading AI systems—Harvey AI, Grok, Claude, Gemini, and ChatGPT—to each construct, from "the basis of computational machine intelligence" and without endorsing any human position, a model AI policy for law school coursework and exams, and then pushed each system with a follow-up challenge to expose the value judgments hidden in its own language. [1] Below is (1) a comprehensive summary of each response, (2) an analysis of similarities and differences, (3) a categorization of the five approaches, and (4) an assessment of how these machine-generated policies differ from human-developed law school AI policies.

1. Comprehensive Summary

The prompt and its premise. Backer asked each system, on the basis of his own prior comparative study of twelve law school AI policies and independent research, to reason "strictly from the perspective of computational machine intelligence" and produce both an explanation of its reasoning and an actual model policy text, citing sources. [1] [2] He then issued a common follow-up to several systems criticizing their use of value-laden, undefined terms like "stronger" and "weaker," asking them to make their premises and value structures explicit. [3]

Harvey AI. Harvey framed its task as normative synthesis rather than empirical description and expressly disclaimed authority to declare a "correct" answer. [4] It identified structural opacity and unpredictability—not excessive restriction or permissiveness—as the most consistent documented institutional failure, citing that nearly 40% of institutions with AI guidance never wrote it down and that one school's guidance page was found hidden behind a visibility setting. [5] It emphasized that policy "axes" (polarity, drafting style, autonomy structure) are independent and should be made explicit rather than eliminated. [6] Distinctively, Harvey drew on its own

"self-knowledge" of hallucination as a structural (not occasional) property of generative systems to justify a strong verification obligation tied to the ABA's duty of competence. It favored a deductive structure (stated purpose → derived rules) while requiring that revisions be versioned and archived to prevent "revisability" from masking accountability avoidance. [7] Its resulting ten-section model policy included: a stated dual purpose (independent reasoning vs. professional competence); a restrictive default permitting only source-identification, outlining, and proofreading, with a full prohibition on exams absent contrary syllabus language; instructor discretion conditioned on written, centrally registered syllabus disclosure; a non-waivable floor barring unattributed AI authorship, professional-responsibility violations, and unverified reliance on AI output; mandatory disclosure as a condition (not substitute) for authorization; an explicit verification obligation treating hallucination as structural; a bar on using AI-detection tool output alone as a violation basis; confidentiality restrictions on inputting privileged data into unapproved tools; and mandatory versioned, archived annual review. [8] [9] [10] When challenged, Harvey candidly conceded it had smuggled in an undisclosed ranking of values, identified four competing objectives (minimizing enforcement error/unfair sanction; maximizing professional competence; maximizing unaided cognitive skill development; maximizing transparency/accountability), and admitted it had prioritized transparency and fair notice because those failure modes were the ones with the strongest independent empirical support (UNESCO's 450+ institution survey, Jiang et al.), while skill-erosion concerns were "asserted" rather than independently measured, citing one SSRN study finding AI use "did not diminish downstream comprehension" in a tested context.

Grok. Grok framed machine policy-construction as "systematic optimization for truth-seeking, human cognitive development, ethical integrity, and institutional adaptability," modeling AI as powerful at pattern-matching but weak at grounded judgment and accountability. [11] [12] Its design principles included combining a permissive baseline with bright-line rules for high-stakes contexts, a "truth-seeking objective" minimizing deception, "human condition grounding" (humans build capacity through effortful reasoning, so over-reliance risks skill atrophy), opacity mitigation through public availability and plain language, and evaluation of policies on clarity, fairness, efficacy, and scalability. Its resulting model policy adopted a permissive-with-guardrails default for general coursework, an opt-in (affirmatively-authorized) restriction for exams and absolute prohibition for proctored exams, a hybrid rules-plus-standards drafting style, mandatory disclosure/attribution statements, an autonomy structure permitting instructor deviation within a non-waivable floor, enforcement tied to the academic integrity code, mandatory AI-literacy training, and illustrative "allowed/prohibited/gray area" examples. Grok closed by recommending piloting and A/B testing of policy variants and measuring outcomes empirically. [13]

Claude. Claude's initial response (Version 1.1) opened with an unusually elaborate "Section 0: Underlying Premises and Value Commitments," stating six explicit, self-labeled premises (P1–P6) before any substantive rule—e.g., that unaided performance has variable pedagogical value depending on context (P1), that written/public rules are preferred over oral/private communication independent of substantive polarity (P2), that predictability matters more where

error costs are high (P3), that disclosure does not cure an otherwise-prohibited use (P4), that certain harms are categorically non-waivable (P5), and that institutional memory/versioning is valued over the appearance of settled finality (P6). Claude explicitly told adopting institutions to revise sections corresponding to any premise they rejected. [14] [15] Its policy included a Definitions section giving operational (not moral) meanings to terms like "substantive," "material fabrication," and "independent performance"; a governing principle section; a publication/status clause designed to foreclose the four opacity patterns Backer's research identified (non-existence, disclosed decentralization, access-gating, unwritten/oral guidance); a fill-in-the-blank "Default Baseline" section offering an institution a binary choice between a restrictive or permissive default, with exams restrictive-by-default regardless of the choice made; a Chicago-derived "backstop standard" for unenumerated uses (treat AI like an undisclosed human collaborator); disclosure/documentation rules drawn from CTLS and Columbia; an instructor-override provision modeled on Berkeley's notice requirement; a non-waivable floor drawn from Stanford's general and clinic-specific floors; and a scheduled 18-month review/versioning requirement. [16] [17] [18] [19] Claude attributed every structural feature to a specific named institution and explicitly stated it did not endorse any single school's substantive conclusion on restrictiveness.

Gemini. Gemini took the most technical/systems-theoretic register, opening with "Part I: Explicit Axiomatic & Value Definitions" that redefined normatively loaded terms purely operationally: "Ideal/Optimal" as friction-minimizing and error-minimizing, "Fair Notice" as system predictability, "Opacity" as information asymmetry, and "Reward-Hacking" as a student satisfying a formal grading metric while bypassing the actual learning process. Its model policy used engineering-style labeling (POLICY IDENTIFIER, SYSTEM ORIENTATION) and classified tools into three tiers: Tier Alpha (deterministic/syntactic tools like spell-checkers, permitted by default), Tier Beta (probabilistic LLM heuristics, restrained by default and requiring instructor variance), and Tier Gamma (autonomous full-text synthesis, prohibited by default). [20] Exam rules were tied to the "runtime environment" (network/OS-level restriction for proctored exams; strict non-execution as the take-home default). Verification relied on mandatory, immutable prompt/output logs and an "attribution matrix" rather than AI-detection tools, which it barred from triggering enforcement absent a missing log ("Heuristic Safe Harbor"). [21] Faculty override was framed as "Decentralized Subsystem Autonomy" permitting bidirectional variance if stated in a determinate written catalog before evaluation.

ChatGPT. ChatGPT's response was the most abstract, framing the task as an optimization problem: not to maximize or minimize AI use, but to "maximize long-term production of competent lawyers while minimizing informational uncertainty regarding how competence was acquired," making AI itself just one instrumental, regulated variable. [22] It proposed a seven-part objective function (competence, truthful attribution, reproducibility, fairness, transparency, adaptability, institutional legitimacy) pursued via Pareto efficiency rather than maximization of any single dimension. [23] It argued the axes Backer identified (polarity, drafting style, autonomy) were merely "implementation variables," while the true latent variables were human learning, information provenance, assessment reliability, professional competence, and public trust. It articulated five "Computational Principles," most notably that assignments should be

classified by "information dependence" rather than technology, yielding five categories from "Intrinsic cognition" (AI prohibited) through "Professional simulation" (AI expected). It rejected AI-detection as unreliable and adversarially vulnerable in favor of mandatory provenance metadata. It proposed a four-layer dynamic policy architecture (constitutional principles, educational objectives, operational rules, model-specific guidance), with only the bottom two layers requiring frequent revision, plus a self-monitoring "governance of governance" function tracking bar performance, employment outcomes, and feedback. It listed a multi-category source list (Backer's comparative study, primary institutional policies, UNESCO surveys, higher-education AI governance research, regulatory-design literature on rules vs. standards, professional responsibility doctrine, and information theory/systems engineering literature). Its central claim was that current policies are "technology-centered" (beginning with AI's existence and specifying permissions), whereas an optimized policy is "objective-centered" (beginning with the capability to be demonstrated and deriving AI's permissible role from that). [24] [25] [26] On follow-up, ChatGPT conceded that terms like "competence," "fairness," and "transparency" are not computational primitives but human-constructed latent variables, and it added a "Computational Premises" section stating that no governance system is neutral and that every rule must identify the variable it optimizes. It then gave stripped-down operational redefinitions—competence as demonstrated task performance under specified conditions, transparency as "recoverability of the informational pathway," integrity as "consistency between declared provenance and reconstructable production history," and fairness as "equal application of identical evaluative procedures to informationally equivalent cases". It further justified privileging educational objectives over technology using a control-theory analogy (technology as a high-variance "control input," educational capability as a stable "state variable"), concluding that governance should regulate informational dependence rather than any specific tool, while acknowledging this conclusion is empirically contingent and could reverse if objectives ever became more volatile than the technologies used to pursue them.

2. Analysis of Similarities and Differences

Dimension	Similarities Across Systems	Key Differences
Opacity/transparency	All five treat opacity, hidden defaults, or undisclosed value judgments as a central problem to be solved through public, written, versioned policy. [5] [27] [28] [29]	Harvey and Claude tie this directly to Backer's empirical opacity findings (hidden guidance pages, unindexed syllabi); Gemini and ChatGPT reframe it in information-theoretic terms ("information asymmetry," "informational entropy"). [28] [30] [31] [32]
AI-detection tools	All systems that address enforcement (Harvey, Grok,	Harvey and Claude allow detection as one

	Claude, Gemini, ChatGPT) reject reliance on automated AI-detection output as a sole or primary basis for a violation finding, citing unreliability and disparate error rates. [33] [34] [35] [36]	corroborating input among others; Gemini and ChatGPT replace detection almost entirely with mandatory provenance/log-based verification ("Heuristic Safe Harbor," provenance metadata). [33] [35] [36] [37]
Verification/hallucination	Harvey, Grok, Claude, and Gemini all impose an affirmative human verification obligation grounded in the recognition that generative AI can produce confident, fabricated content. [35] [38] [39] [40]	Harvey treats this as self-referential "machine self-knowledge" driving a stronger rule than a human drafter might write; Gemini operationalizes it as an "immutable verification trail" requirement; ChatGPT subsumes it into "provenance" rather than a standalone duty. [35] [37] [41]
Faculty/instructor autonomy	All systems preserve instructor discretion to deviate from an institutional default, but only if exercised in writing with advance notice, and all impose some non-waivable floor no instructor may override. [10] [42] [43] [44]	Claude's floor is the most granular (three explicit categories tied to named institutional models); Gemini frames this as "Decentralized Subsystem Autonomy" with bidirectional variance; ChatGPT does not present a discrete autonomy clause at all, subsuming it into its category-based "information dependence" framework. [43] [45] [46]
Default polarity	None claim a single correct default is compelled by the data; several explicitly leave it open.	Grok commits to a permissive-baseline default; Harvey commits to a restrictive default; Claude presents both as alternative checkboxes for the institution to select; Gemini adopts a graduated, tool-tier-based default (Alpha permitted,

		Beta restrained, Gamma prohibited) rather than a single binary default; ChatGPT rejects a technology-based default altogether in favor of a task/objective-based classification (Categories I–V). [47] [48] [49]
Self-critique on value premises	Harvey and ChatGPT (and Claude proactively) were each confronted with, or preemptively addressed, the charge that "stronger/weaker" language smuggles unstated values, and each responded by explicitly enumerating competing objectives. [3] [15] [50] [51]	Harvey names four contested pedagogical objectives (A–D) tied to specific institutions (Berkeley vs. UT Austin); Claude built this in from the outset as "Section 0"; ChatGPT goes furthest by reducing normative terms to purely operational/computable definitions (e.g., fairness as procedural equality, not justice); Grok and Gemini were not shown being challenged on this point in the excerpted text. [15] [52] [53]
Register/style	All treat this as an exercise in "computational" or "machine" reasoning rather than adopting a human ideological starting point. [2] [54]	Harvey and Claude read as conventional legal drafting citing specific institutions by name; Gemini adopts explicit software/engineering nomenclature (Tiers, Runtime Environments, Policy Identifiers); ChatGPT adopts optimization/control-theory and information-theory vocabulary (objective functions, Pareto efficiency, state variables vs. control inputs); Grok blends legal drafting with an accessible "principles + template" style

		closer to conventional policy memos. [20] [22] [55] [56]

3. Categorization of the Five Responses

The five outputs can be grouped along at least three axes:

By default polarity chosen:

- *Restrictive-by-default*: Harvey (permits only narrow uses; exams prohibited absent contrary syllabus language). [57]
- *Permissive-by-default with guardrails*: Grok (general coursework permissive unless prohibited; exams restrictive).
- *Institution-selects (agnostic)*: Claude (offers a binary checkbox for the adopting institution). [47]
- *Tiered/graduated by tool type*: Gemini (Alpha permitted, Beta restrained, Gamma prohibited, independent of a single "default"). [49]
- *Rejects technology-based default entirely*: ChatGPT (classifies by assignment's "information dependence," Categories I–V, so permissibility follows from the educational objective rather than a technology-wide baseline). [58]

By drafting/structural style:

- *Deductive legal-instrument style* (stated purpose → derived numbered sections, in the manner of a conventional institutional policy): Harvey, Grok, and Claude all produce recognizable "Section 1... Section 10" policy documents with definitions, floors, and review clauses. [59] [60] [61]
- *Systems/software-engineering style*: Gemini, which structures its policy as a technical specification with tiers, runtime environments, and a "policy identifier". [62]
- *Optimization/systems-theory essay style*: ChatGPT, which spends most of its response on the abstract objective function, principles, and category taxonomy before any institution could directly adopt it as a syllabus-ready text.

By degree of explicit value-transparency:

- *Built-in from the start*: Claude, which opens with a dedicated "Section 0" naming six premises before any rule. [15]
- *Added only after being challenged*: Harvey and ChatGPT, both of which initially used unexamined evaluative language ("stronger," "weaker," "ideal") and then, on direct challenge, retroactively supplied explicit value taxonomies (Harvey's Objectives A–D; ChatGPT's "Computational Premises"). [3] [63] [64] [65]
- *Definitionally embedded rather than premise-labeled*: Gemini, which addresses the same transparency concern by giving purely operational (non-moral) definitions of "ideal," "fair notice," and "opacity" up front, without a discrete premises section. [66] [67]

- *Not shown addressing this issue in the excerpt*: Grok, whose text does not include a comparable follow-up challenge or self-critique of its evaluative terms.

4. Gemini's Tiered Tool Taxonomy vs. ChatGPT's Task-Based Categories

Although both frameworks reject a simple binary "AI permitted/prohibited" rule, they classify along fundamentally different axes, and this produces meaningfully different operational consequences.

What each system actually classifies

Gemini classifies the tool itself, based on its internal computational architecture, independent of what assignment or context it is used in. Gemini defines three tiers: Tier Alpha covers "deterministic/syntactic tools" that operate on "fixed grammatical rule-matching without probabilistic token generation," such as spell-checkers and citation auto-formatters, and these are "permitted by default across all contexts". [1] Tier Beta covers "probabilistic/structural heuristics"—large language models used as "conversational nodes to alter text structure, brainstorm taxonomies, or map conceptual outlines"—and these are "restrained by default; requires instructor variance". [2] Tier Gamma covers "autonomous text synthesis engines," meaning generative models used "to draft complete semantic strings, sentences, paragraphs, or analytical arguments for evaluated credit," and these are "prohibited by default". [3] Crucially, this classification travels with the tool, not the task: a given LLM is always a Tier Beta or Tier Gamma resource regardless of which assignment it touches, and the permission level is fixed to that technical classification unless an instructor issues an explicit written override.

ChatGPT classifies the assignment's information dependence, deliberately treating the technology as irrelevant to the classification. ChatGPT states plainly that "the technology never determines the rule. The educational objective determines the rule", and it organizes assignments into five categories: Category I ("Intrinsic cognition," measuring unaided human reasoning, AI prohibited); Category II ("Assisted cognition," AI permitted for brainstorming, organization, and language support, with reasoning remaining attributable to the student); Category III ("Collaborative cognition," student and AI jointly producing work, with mandatory contribution disclosure); Category IV ("Professional simulation," where AI use is expected and evaluation instead concerns judgment, verification, and legal responsibility); and Category V ("Research augmentation," where AI functions as a retrieval/exploratory tool with independently verified outputs). [4] [5] Under this scheme, the same tool—say, an LLM—could be fully prohibited in a Category I exam question and simultaneously expected and rewarded in a Category IV professional-simulation exercise within the same course, because the classification runs to the assignment's purpose, not to the tool's technical properties. [4] [5]

The functional consequence of this difference

This distinction is not merely semantic; it changes what an instructor or student must evaluate before acting. Under Gemini's framework, the threshold question is "what kind of tool is this?"—a relatively static, easily catalogued question that an administrator could answer once for a given software product and apply uniformly across the curriculum. Under ChatGPT's framework, the threshold question is "what capability is this specific assignment trying to measure?"—a question that must be answered freshly for every assignment, and potentially differently by different instructors even using the identical software tool. ChatGPT itself frames this as a deliberate design choice against exactly the kind of tool-based classification Gemini adopts, warning that treating "Was ChatGPT used?" [1] as the operative question commits "a category error" by "regulat[ing] the instrument instead of the property of interest," and arguing the relevant variables are instead "[w]ho generated the legal reasoning? Who selected authorities? Who exercised legal judgment? ... Can the production pathway be reconstructed?".
This also produces different robustness profiles to technological change, a point ChatGPT makes explicitly in its own reasoning (though directed at its own approach rather than Gemini's): because "[t]echnologies are High-Variance Variables" while "[e]ducational objectives are Lower-Variance Variables," a technology-tiered system "must continually change because the independent variable evolves continuously," whereas an objective-based system anchors to something comparatively stable. Applied to Gemini's scheme, this suggests a structural vulnerability: as tools blur across tiers (for example, an LLM feature that performs deterministic citation-formatting alongside probabilistic drafting within a single product), the Alpha/Beta/Gamma boundaries would require frequent recalibration, whereas ChatGPT's category boundaries (defined by what capability is being tested) would not need to change merely because new tools enter the market.

Conversely, Gemini's approach has a practical enforceability advantage that ChatGPT's arguably lacks: because the tier attaches to the tool rather than to a case-by-case judgment about assignment purpose, it is easier to state a single bright-line, sector-wide default (e.g., "spell-checkers always permitted; full-text generation always prohibited absent override") without requiring an instructor to first articulate an explicit learning objective for every assignment. [1] [6] ChatGPT's task-based scheme instead pushes significant interpretive burden onto instructors, who must correctly classify each assignment into one of five categories and communicate that classification—a burden that echoes the "opacity" concerns both Harvey and Claude identified as the central documented failure mode in actual law school policies. In short, Gemini optimizes for administrability and predictability by fixing the classification to a comparatively static technical property, while ChatGPT optimizes for pedagogical precision and technological durability by fixing the classification to a comparatively stable educational objective, at the cost of requiring more granular, assignment-by-assignment judgment.

5. How These Machine-Generated Policies Differ From Human-Developed AI Policies

Drawing on the systems' own comparisons to the twelve-school dataset referenced throughout the document, several structural differences emerge between these machine outputs and the human institutional policies they were built to respond to:

They regulate the technology's function rather than its brand or category by default in several cases, or reject technology-classification altogether. ChatGPT explicitly argues that "the dominant pattern in current policies is technology-centered governance: they begin with the existence of generative AI and then specify permissions or prohibitions," whereas its own proposal is "objective-centered," starting from the capability to be assessed and deriving AI's permissible role only afterward. [24] [25] [26] Gemini similarly classifies tools by their computational architecture (deterministic vs. probabilistic vs. autonomous-synthesis) rather than by product name or blanket permission/prohibition. [49]

They treat their own class of tool's failure modes (hallucination, detection unreliability) as structural, first-hand knowledge rather than externally-reported risk. Harvey explicitly frames its emphasis on verification as arising from the machine's "self-knowledge... of its own class of tool's failure modes," which it says "would push toward stronger verification language than a purely human drafter, unfamiliar with the mechanics of hallucination, might otherwise include". [41] [68] This self-referential vantage point is not available to a human drafting committee in the same way.

They uniformly and near-categorically reject AI-detection software as a basis for enforcement, proposing provenance logs, prompt/output capture, or process-verification hierarchies instead. Human-drafted policies, per the systems' own characterization of the underlying research, have been comparatively inconsistent or silent on this point, with detection-tool reliability flagged in the research as a documented but unevenly addressed problem.

They build explicit versioning/archiving and "meta-policy" safeguards against revisability being used to avoid accountability, a concern several systems say arose specifically from noticing a pattern (e.g., Chicago's revisability language) in the human dataset that could function, "whether intentionally or not, as a mechanism for avoiding durable public accountability". [69] Harvey, Claude, and ChatGPT's four-layer architecture all build in mandatory dated/versioned archiving of prior policy text as a structural response to this risk, a feature the systems suggest is often absent or inconsistent in the human-drafted sample.

They explicitly disclaim ideological starting points and instead present themselves as optimizing a stated objective function, a framing distinct from conventional policy drafting. ChatGPT states directly that "a computational intelligence does not begin from ideological priors (academic freedom, innovation, integrity, prohibition, trust, distrust, autonomy, surveillance, etc.). Instead it attempts to optimize a system subject to multiple constraints," and it characterizes existing law-school policies as "largely historical artifacts rather than optimized governance systems... products of institutional evolution, imitation, risk management, and incremental adaptation rather than formal systems design". [54] This self-description positions the machine

outputs as attempting formal systems design where the human comparators are described (by the machines) as path-dependent and improvisational.

They are unusually explicit—sometimes only after prompting—about the contestability of their own value premises, naming competing objectives and stating which one they privileged and why, with citation to specific empirical support (e.g., Harvey's citation of UNESCO's 450+ institution survey and Jiang et al. as justifying its prioritization of transparency over skill-erosion concerns). [70] [71] Claude built such a disclosure into its base design without being asked. [15] This degree of explicit, sourced meta-commentary on the policy's own normative foundations is not a typical feature of conventional institutional policy documents, which more often state rules without exposing the underlying value hierarchy.
They substitute purely operational/computable definitions for normatively loaded terms, most explicitly in ChatGPT's *follow-up response*, which was shown responding to an express user challenge, and, independently, in Gemini's single response, which frames these definitions as a self-initiated design choice rather than as a response to any shown challenge — Gemini's text opens by stating that "[t]o strip away institutional opacity and ensure analytical clarity, the underlying value structures, premises, and terms utilized in this computational model must be explicitly defined", with no preceding critique prompt appearing in the retrieved text. [1] For example, ChatGPT's follow-up redefines "fairness" as "[e]qual application of identical evaluative procedures to informationally equivalent cases" specifically because "machines cannot optimize justice because justice possesses no universally computable objective function", while Gemini — addressing different terms, not the same ones ChatGPT later revised — defines "'[f]air notice' (System Predictability)" as "[a] core operational constraint requiring that the boundary parameters of permitted user actions be explicitly mapped ex-ante". [3] [4]

Notably, despite this technical framing, several systems converge on affording greater weight to human non-delegable responsibility than a purely restriction-driven human policy might. ChatGPT observes that its computational approach "probably affords greater respect to human agency than many existing restrictive policies... [because] humans remain the accountability node. AI cannot presently bear legal responsibility. Lawyers can," concluding this follows "from optimization rather than moral philosophy". [73] This suggests that even a machine-optimized approach converges on a human-centric accountability principle, but arrives there through instrumental/optimization reasoning rather than through the professional-responsibility or moral-education framing more typical of human-drafted law school policies.

6. Feasibility of Machine-Emphasized Versioning and Archiving Practices for Typical Law Schools

Typical Law Schools

Three of the five systems built dated, versioned, publicly archived policy retention directly into their model texts. Harvey's Section 9 requires that "[a]ny revision will be dated, versioned, and

published alongside all prior versions rather than replacing them, so that the rule governing any given semester's work remains permanently identifiable and citable". [7] [8] Claude's Section 8 similarly commits to "versioning and public retention of prior text specifically because this policy assumes future revision is likely, not a sign of failure", and its Section 3 states the policy "will remain publicly retrievable in its current and prior versions". [9] [10] ChatGPT proposes a four-layer architecture (constitutional principles, educational objectives, operational rules, model-specific guidance) in which only the bottom two layers "require frequent revision," alongside a broader "governance of governance" function tracking annual review, performance metrics, and outcome data.

Feasibility considerations in favor of adoption. These practices are, in a narrow technical sense, low-cost for a modern law school to implement. Publishing dated, versioned PDFs or webpages, retaining prior versions in an archive folder, and adding a "Version 1.0 — Adopted [date]" header, as Harvey's model does, requires no specialized infrastructure beyond what most law school websites and registrar's offices already use for course catalogs, academic calendars, and handbook revisions—documents that are already routinely versioned and archived by academic institutions as a matter of general administrative practice. [11] The underlying problem the systems are responding to—opacity—was itself documented at meaningful scale in the research the machines were given, including a finding that nearly 40% of institutions with any AI guidance had never written it down, and that at least one law school's guidance page was found "affirmatively hidden behind a visibility setting". [12] Given that baseline, even a modest formalization requirement (a dated document plus an archive folder) would represent a meaningful improvement over current practice for a substantial share of institutions, and would not require new technology, only administrative discipline.

Feasibility considerations against straightforward adoption. The more demanding structural proposals are harder to reconcile with how law school policy is actually produced and governed. Claude's model requires that any instructor deviation be centrally filed and "included in a centrally indexed, publicly searchable registry of course-level AI policies maintained by the school," with the explicit condition that "[n]o instructor policy is valid until it appears in that registry". This is a meaningfully larger institutional commitment than simply archiving the central policy document: it requires an administrative office (Registrar or Dean of Academic Affairs) to actively collect, validate, and publish potentially dozens of individualized syllabus provisions every term, and to treat a syllabus provision as legally inoperative until that registry step occurs. Faculty governance processes at most law schools do not currently route syllabus content through central administrative approval or registry-filing before a course begins, and building such a workflow would require new administrative capacity, a designated office of responsibility, and faculty buy-in to a compliance step that does not presently exist in most curricular oversight structures. Harvey itself frames the underlying research concern this responds to—the "dispersal of governing rules across ungathered syllabi"—as a genuine and documented problem, but solving it through a mandatory registry is a considerably heavier institutional lift than simple archiving of a single central document. [13] [14]

ChatGPT's "governance of governance" proposal is more ambitious still, calling for an optimized system that includes "annual review, performance metrics, student feedback, faculty feedback, technological monitoring, professional licensing outcomes, judicial developments, bar examination performance, employment outcomes, [and] policy revision protocols," with the aim that "[p]olicy becomes self-learning". [15] Tracking bar passage and employment outcomes specifically as a function of AI policy revision would require a law school to isolate the policy's causal effect from the many other variables affecting those outcomes—a task considerably more sophisticated than most law schools' current institutional research capacity is designed to perform, and one that would likely require new data-collection infrastructure, longitudinal tracking across graduating cohorts, and probably outside expertise in applied social-science methodology that a curriculum or academic-affairs committee does not typically possess in-house.

Overall assessment. The core, low-cost element common to all three systems' proposals—dating each version of the central AI policy, publishing it at a stable public URL, and retaining prior versions rather than silently overwriting them—is realistically and inexpensively adoptable by essentially any law school with a functioning website and a designated policy owner (such as the Dean of Academic Affairs), and directly addresses the opacity problems the underlying research documented. The mid-tier proposal—a centralized, mandatory registry of individual instructor deviations that determines the legal validity of a syllabus provision—is feasible in principle but would require new administrative workflow and enforcement capacity that most law schools do not currently have built for this purpose, making it a plausible medium-term goal rather than an immediate, low-friction change. The most ambitious proposal—continuous, outcome-linked, self-revising policy governance tied to bar and employment data—is unlikely to be realistically implementable by a typical law school in the near term, both because of the institutional research capacity it demands and because isolating a single policy's causal effect on downstream outcomes like bar passage is a methodologically difficult problem that the source material does not show any of the five systems, or any of the twelve schools in the underlying dataset, having actually solved or attempted. [15] [16]

7. The Extent to Which Machine System Models Affect the Scope of Human Agency in Law

This section extends the prior description/analysis by considering two further questions: It does so in two sweeps, the first focused on the five machine system models. This analysis below is therefore grounded in what the attached document itself says about that underlying research, as relayed through the five systems' own citations to it, rather than in independent access to the report's full text. The second on the five models as a function of the human modelling discussed in Backer's "Structure, Opacity, and Convergence: A Consolidated Analysis of Law School Generative AI Coursework and Exam Policies" rather than solely in the five AI systems' own characterizations of that report. Where a system's account of the Report's findings is confirmed

by the Report's own text, that is noted; where a system's characterization appears to diverge from or overstate what the Report actually says, that divergence is flagged rather than smoothed over, consistent with the Report's own stated methodological practice of saying explicitly, "where evidence is thin, secondary-sourced, or interpretive rather than established".

Retained loci of human agency. All five systems, to varying degrees, preserve a formal locus of non-delegable human responsibility. ChatGPT states this most directly: "AI may assist AI may generate AI may recommend AI may predict But human responsibility remains non-delegable," adding that "[t]hat follows from optimization rather than moral philosophy". [1] Harvey's model similarly makes verification a "substantive competence requirement" rather than a mere disclosure formality, on the ground that "the ABA's own guidance on professional AI use identifies verification as inseparable from the duty of competence", and its Section 6 states that "[s]tudents remain fully responsible for the accuracy of any submitted work regardless of the extent to which generative AI assisted in its preparation". [2] [3] Claude's definition of "independent performance" as "the demonstration of a student's own reasoning, drafting, or analytical process without undisclosed automated assistance" performs a comparable function, anchoring a protected zone of unaided human judgment to a specific defined term. [4] On this dimension, the machine models do not displace human agency so much as relocate its textual anchor into an explicit, defined obligation.

The non-waivable floors that Harvey, Grok, Claude, and Gemini each build into their model policies are not machine inventions; they closely track floors the Report documents as already operative in at least two schools in its twelve-school dataset. The Report states that at Stanford, "the student's use of AI cannot contravene standard academic norms concerning plagiarism and accuracy, and in coursework involving the practice of law, AI must meet all applicable rules of professional responsibility," such that "an instructor's permissive course policy is itself bounded by plagiarism doctrine and (in clinical settings) professional-responsibility rules that no instructor can contract around". [2] Chicago's plagiarism-equivalence test operates identically: "a student cannot avoid limits on the use of generative AI that would otherwise apply by attributing their work to generative AI". [3] The Report characterizes this pattern as showing that "faculty autonomy, across every school where it is textually elaborated, is designed as bounded discretion — a zone of instructor choice nested inside an outer, non-waivable institutional or professional floor — rather than unlimited course-by-course sovereignty". [2] [4] This means that, on this specific dimension, the machine systems' models extend and generalize an architecture already present in the human sample rather than introducing a novel constraint on agency; Harvey's Section 4, Claude's Section 8, Grok's "Floor Protections," and Gemini's tiered defaults are each recognizable elaborations of a bounded-discretion structure the Report finds "close to the central organizing design choice" of the human dataset itself, not a machine-originated limitation on the scope of human decision-making. [5]

Relocation and proceduralization of agency. At the same time, several systems reframe the human actor's role in ways that shift agency from the exercise of unsupervised judgment toward the production of an auditable record. Gemini's verification architecture requires that "the student

must maintain an immutable verification trail consisting of the exact input prompt sequences and the corresponding raw machine outputs" and an "Attribution Matrix" documenting "the exact model architecture and compilation runtime date". [5] ChatGPT's parallel proposal requires "structured metadata describing external tools models versions prompts verification procedures human modifications" for "[e]very submission". [6] In both frameworks, what counts as evidence of proper human agency is not the demonstrated quality of independent reasoning itself but the completeness of a documentation trail describing the human's interaction with the tool. This is a meaningful shift in what the system treats as legally or academically cognizable: agency becomes verifiable through recordkeeping compliance rather than assessed through the substance of the work product, which is consistent with Gemini's own stated design goal of eliminating "enforcement vulnerabilities caused by unreliable, probabilistic AI-detection algorithms" by substituting "verifiable user logs", but which also means that a student who complies with documentation requirements while making comparatively little independent analytical contribution, and one who complies with identical documentation requirements while making a substantial independent contribution, are treated identically at the level of the rule itself — the rule's variable is documentation compliance, not degree of unaided judgment exercised. [7] [8]

Gemini's treatment of the student as an "optimization agent." Gemini's definition of "reward-hacking" is notable in this respect: it describes "a phenomenon where a human optimization agent (the student) satisfies the formal metric of an objective function (submitting a text string that passes a grading rubric) by bypassing the intended optimization pathway (the actual neurological cognitive processing and learning required to write that text)". This definition models the student's cognitive development process itself as an "optimization pathway" and casts unauthorized AI use as a form of the student's own metric-gaming rather than, for example, a breach of trust, an act of dishonesty, or a failure of professional formation. [9] The vocabulary is drawn from machine-learning evaluation rather than from the language of academic integrity or professional ethics that conventionally frames this conduct in human-drafted policies, and it treats the underlying human learning process as itself a computable pathway subject to the same optimization/gaming dynamics that afflict machine-learning systems.

Contingency of the human-primacy justification. A further consideration concerns the durability of the ground on which human primacy rests in these models. ChatGPT's justification for non-delegable human responsibility is explicitly tied to a present technical/legal fact — "AI cannot presently bear legal responsibility. Lawyers can" — rather than to an assertion of inherent human moral agency as such. [10] Because this justification is framed as following "from optimization rather than moral philosophy", it is, by the system's own framing, contingent on that underlying fact remaining true. [1] A justification grounded in professional responsibility doctrine or in a categorical account of moral agency would not depend on this contingency in the same way; a justification grounded in AI's current incapacity to bear legal liability is, on its own terms, revisable if that incapacity were ever to change (for example, through some future legal or regulatory development altering how liability attaches to automated systems). None of the five systems address what would follow for their models if this premise changed, which leaves the

scope of human agency in these frameworks resting on an empirical/legal condition external to the policy itself rather than on a freestanding normative commitment.

Differential weighting of harder-to-measure agency claims. Harvey's self-critique further illustrates a mechanism by which the scope of protected unaided human judgment could narrow under these models relative to some human-drafted policies. Harvey concedes that it deprioritized "Objective C" (maximizing unaided cognitive skill development) because the claim that permissive AI policies erode cognitive skill development "was supported in this research mostly by institutional policy statements asserting the risk rather than by independent empirical measurement of the effect," citing one partial counter-finding (the SSRN study) that early AI use "did not diminish downstream comprehension of the underlying legal principles" in the tested context. [11] Harvey states directly that it "weighted the better-evidenced problem (opacity) more heavily than the more assertion-based, less independently tested problem (skill erosion)". [12] By contrast, Harvey notes that "Berkeley's 2026 policy visibly prioritizes Objective C, stating that its purpose is to ensure students develop skills 'distinct from technologically-assisted performance,' even at the cost of restricting AI-assisted efficiency" — a human institution asserting this value categorically, independent of whether it has been independently measured. [13] The comparison suggests a structural tendency in at least this machine model to systematically discount values that resist quantification relative to values (transparency, notice, documentation) that are more directly measurable or already supported by large-sample survey data (Harvey cites "UNESCO's 450+ institutions" and "Jiang et al.'s two-university study" as the basis for its opposite prioritization). If such a tendency were generalized across machine-constructed policy, it could narrow the institutional space accorded to the harder-to-measure value of unaided human judgment relative to policies drafted by human bodies willing to assert that value on other grounds. [14]

Compression of a more granular human agency structure into a two-node model. The Report identifies, at Stanford specifically, "a four-tier discretion hierarchy... university guidance, law-school default, clinical-program-specific policy, and (within that clinic's own language) further individual-supervisor discretion," and states this "is a limitation of the original coding scheme worth flagging directly: it did not have a category for program-level (as distinct from course-level) policy variation". [6] None of the five machine systems' model policies reproduce this granularity. Harvey's autonomy structure runs institution-to-instructor only; Claude's runs "institution-to-instructor/instructor-to-student" as its stated fixed design variable; Grok's autonomy structure lists "Institution to Instructor," "Instructor to Student," and "Program-Level (e.g., Clinics)" as a third, but treats it briefly as a subordinate case rather than a distinct hierarchical tier with its own supervisor-level discretion. [7] [8] [9] To the extent a machine-generated model policy is adopted in place of a locally elaborated multi-tier structure, it would collapse a documented four-level allocation of human decision-making authority into a coarser two- or three-level scheme, which — independent of where any given machine model sets its substantive defaults — narrows the number of distinct loci at which human judgment is formally recognized as operative within the institution.

The "sticky default" phenomenon as an empirical constraint on the practical scope of agency. The Report's Appendix draws on private-law and regulatory-design scholarship to note that "because opting out requires effort (drafting and publishing written notice), the choice of baseline polarity itself functions as a substantive lever even holding the range of permissible instructor choices constant," citing this as "the 'sticky default rule' phenomenon described in contract-default-rule scholarship," such that "[a] law school that sets a restrictive default will likely see more restrictive outcomes in courses where an instructor never gets around to writing an alternative syllabus policy than a law school with an identical range of permissible instructor choices but a permissive default — independent of any school's stated substantive preference". [10] This finding complicates the machine systems' generally formal treatment of instructor "discretion" as coextensive with instructor "agency." Harvey's, Grok's, and Claude's models each preserve a formally symmetric override right (an instructor may move the rule "in either direction"), but the Report's sticky-default finding indicates that the practical scope of exercised agency is not determined solely by the formal existence of that right; it is also determined by which direction requires affirmative drafting effort, a variable none of the five systems' models addresses as a distinct design parameter. [8] [11] A model policy that sets a restrictive default and calls this "instructor discretion" is, on the Report's own analysis, likely to produce systematically different real-world instructor behavior than one that sets a permissive default with an identical formal override right — meaning the machine models' choice of default polarity (Harvey and Claude's restrictive defaults versus Grok's permissive default) itself operates as a lever on the effective, not merely nominal, scope of human agency exercised under the policy. [11] [12] [13]

Proceduralization of agency into documentation compliance. The point made in the prior version of this analysis stands and is not contradicted by the Report: Gemini's requirement of "an immutable verification trail consisting of the exact input prompt sequences and the corresponding raw machine outputs" and an "Attribution Matrix", and ChatGPT's requirement of "structured metadata describing external tools models versions prompts verification procedures human modifications" for every submission, go beyond what the Report documents in the human sample. [14] [15] The Report's closest human analogue is CTLS's mandatory "record of the prompts used to generate content", which the Report treats as a single documentation duty paired with a permissive substantive standard, not as a comprehensive audit-trail regime covering all submissions. [16] The machine systems' extension of CTLS's narrow, permissive-context logging duty into a general verification architecture represents an amplification beyond the documented human baseline, and — as previously noted — shifts what counts as evidence of agency from the substance of independent reasoning to the completeness of a compliance record.

A documented tension between a machine system's self-critique and the primary source it cites. Harvey's follow-up response, when defending its choice to deprioritize "unaided cognitive skill development" (Objective C), states that "UT Austin's 2026 reform visibly prioritizes Objective B, explicitly rejecting the idea that AI fluency and independent judgment are 'in a zero-sum relationship' and building infrastructure (tool licensing, curriculum mapping) to pursue both". [17] The Report's own account of UT Austin, however, describes something closer to the opposite emphasis: it characterizes UT Austin's 2026 reform as "a structural exam-format reform

— 'near-complete elimination of take-home exams' in favor of in-class, software-restricted testing", and states that the reform "can be understood as explicitly framed as a response to observed practice, warning that classroom time may be 'the sole context in which a professor can be certain' a student is learning unaided" — language emphasizing concern for preserving unaided demonstration (closer to Objective C) rather than an integrative "both/and" framing premised on rejecting a zero-sum relationship. [18] [19] The retrieved Report text does not contain the "zero-sum relationship" quotation Harvey attributes to UT Austin. This does not establish with certainty that Harvey's characterization is unsupported by some other portion of the Report not captured in the retrieved chunks, but on the material available, Harvey's self-corrective account of its own value-weighting relies on a characterization of a named institution that is difficult to reconcile with, and in one respect appears to run counter to, the Report's own description of that institution's actual documented rationale. This is a concrete instance bearing on the scope of human agency question: Harvey's stated justification for treating unaided-skill-development concerns as merely "assertion-based" and empirically weak partly depends on a citation that, as retrieved, does not clearly support the specific characterization offered.

8. Consequential Effects of the Agency Problem for the Structure and Character of Law as a Human Inter-Operative System

Convergence with classical procedural attributes of legality. Several structural features that the machine systems derive from "computational" or "information-theoretic" reasoning closely track long-recognized procedural attributes commonly associated with the rule of law and legal process theory — publicity, prospectivity, generality, clarity, and congruence between the rule as announced and the rule as applied. Harvey's Section 9 requiring that "no revision takes effect retroactively against work already submitted" and that revisions be "dated, versioned, and published alongside all prior versions" instantiates prospectivity and public accessibility. Claude's Premise 2 — "[w]ritten and public is preferred over oral and private, independent of the substantive rule chosen" — instantiates publicity as a value independent of substance. [15] Gemini's operational definition of "fair notice" as "a core operational constraint requiring that the boundary parameters of permitted user actions be explicitly mapped ex-ante," on the reasoning that "[i]f a system executes a penalty on an undefined variable, it introduces semantic noise and breaks the trust loop of the environment", reaches a materially similar conclusion to conventional fair-notice doctrine, but by a route framed entirely in systems/information-theoretic terms rather than in terms of due process or legality as such. [16] This convergence admits of at least two readings, neither of which the source material resolves. One reading is that these procedural attributes are, in some sense, structurally necessary features of any durable, function-preserving rule-governed system, whether the reasoning that produces them proceeds from moral-political theory or from computational optimization — on this reading, the machine outputs constitute a form of independent corroboration that these features are not merely contingent artifacts of

human legal culture. The alternative reading is that the "computational" derivation is not independent at all: because these systems were trained on data that already contains centuries of human legal and jurisprudential material encoding exactly these procedural values, their arrival at similar conclusions may reflect recursive reproduction of already-existing human normative content, redescribed in a different vocabulary, rather than a fresh derivation "from the basis of computational machine intelligence" as the original prompt requested. [17] The source material does not provide a basis for adjudicating between these two readings, and none of the five systems directly addresses this possibility with respect to its own output.

Operational redefinition of normative vocabulary as a substantive, not neutral, position. ChatGPT's and Gemini's revised responses go further than the procedural convergence just described by explicitly replacing morally-loaded legal and pedagogical vocabulary with operational, computable substitutes. ChatGPT redefines "[c]ompetence" as "[t]he demonstrated ability of a student to perform a specified legal task under specified informational conditions with a specified level of accuracy and independence," "[t]ransparency" as "[t]he recoverability of the informational pathway through which a submitted work product was produced," "[i]ntegrity" as "[c]onsistency between declared informational provenance and reconstructable production history," and "[f]airness" as "[e]qual application of identical evaluative procedures to informationally equivalent cases," explicitly on the ground that "[f]airness cannot mean justice" because "[m]achines cannot optimize justice [as it] possesses no universally computable objective function". [18] [19] [20] Gemini performs an analogous move, defining "[o]pacity" as "[a] state where the underlying instructions, rules, or code governing a system are restricted from the entities operating within that system," which "raises the computational cost of user compliance". [21] These redefinitions are presented as a way of avoiding smuggled value judgments, but the selection of a purely procedural conception of fairness (equal treatment of like cases) over a substantive or distributive conception (attentiveness to unequal starting conditions or unequal impact) is itself a contestable jurisprudential choice, not a neutral technical simplification — it resembles a formalist or rule-equality conception of fairness rather than, for instance, a conception concerned with disparate impact. This tension appears within the source material itself: Harvey's own model separately requires that AI-detection tools not be used as a sole basis for a violation finding because such tools "produce disparate error rates across student populations, including students for whom English is an additional language" — a concern about substantive/distributive fairness that sits uneasily beside ChatGPT's narrower operational definition of fairness as mere procedural equality of treatment. [22] The material does not show any of the systems reconciling this tension, which suggests that the move to "computable" definitions does not fully escape the underlying normative contestability it was intended to resolve; it relocates the contestability into the choice of which computable proxy to adopt for a given moral or legal concept.

Reframing of the regulated object from a technology to a production function. ChatGPT's central theoretical claim is that the "regulated object is no longer AI" but "becomes the production of demonstrably competent legal professionals," with law schools modeled as "adaptive information-processing institution[s]" whose "outputs are not degrees" but "graduates

possessing measurable characteristics". Gemini's parallel framing describes academic integrity violations as "unauthorized system manipulation" occurring within a "runtime environment" governed by "network or operating system interface level" restrictions. Applied consistently, this reframing casts legal education — and, derivatively, the professional formation that legal education exists to produce — as an input/output production or control system rather than as an interpretive, deliberative, or mentorship-based practice. This is a nontrivial theoretical position with respect to what law and legal education are taken to be: it is closer to a systems-theoretic or cybernetic account of institutions than to accounts that emphasize law as a practice of situated judgment, professional socialization, or interpretive reasoning among human actors operating under conditions of genuine normative disagreement (a disagreement the source material itself repeatedly acknowledges exists, for instance where Harvey states that its model policy is "not... a claim to have solved the underlying substantive disagreement about how much AI use legal education ought to permit — a disagreement this research found to be genuine, contested, and not resolvable by drafting technique alone"). [23] If a production-function framing of legal education were to propagate beyond policy-drafting into how legal reasoning or legal competence itself is conceived, it would raise the further, unresolved question of whether legal reasoning is properly understood as a measurable output of a specifiable process — consistent with ChatGPT's claim that "[s]tudents are evaluated on demonstrated ability" and "[n]ot on whether they preferred pencil, textbook, database, search engine, or LLM" — or whether this framing itself elides distinctions (between, for example, reasoning that is independently generated and reasoning that is externally supplied but independently verified) that a purely output-focused evaluative frame is not well suited to capture. [24]

The systems' own acknowledgment of a limit on value-neutrality. Notably, the material contains an internal check on how far this reframing can be taken, supplied by the systems themselves under direct challenge. ChatGPT, on being pressed to justify its claim that "[t]he technology never determines the rule. The educational objective determines the rule," concedes that "competence, fairness, transparency, integrity, trust, accountability, and even education are not primitives" but are "latent variables constructed by human societies," such that "[a] machine cannot simply assume their meaning" without "merely import[ing] human normative commitments while presenting them as objective optimization criteria". [25] [26] It further states that "[h]uman values enter through institutional choice," that "[t]he computational system does not determine educational objectives. Humans do," and that "the machine neither privileges nor rejects academic freedom, innovation, professionalism, critical thinking, or creativity. Those are supplied externally by the institution". [27] Harvey reaches a structurally similar concession, acknowledging that its earlier use of "stronger" and "weaker" "smuggled in unstated value judgments" and that "a machine intelligence has no independent, freestanding stake in legal education's purpose". [28] Taken together, these concessions indicate that neither system was able to sustain, under scrutiny, the claim that its model derives law's governing structure from computation alone, independent of human-supplied normative content. This suggests that the more far-reaching theoretical claim implicit in the initial framing of the exercise — that a machine intelligence, reasoning "strictly from the perspective of computational machine intelligence", could generate an institutional policy for a human normative practice without

importing a contestable, human-originated theory of what that practice is for — was not, on the systems' own subsequent admissions, actually achieved. [17] [29] What the exercise appears instead to have produced is a set of technically reformulated restatements of contested human value premises, with the reformulation itself (into operational definitions, tiered technical categories, or objective functions) constituting a nontrivial and consequential choice about how law's underlying normative concepts should be represented, rather than a demonstration that those concepts can be derived independently of prior human normative commitment.

The Report's central developmental finding, and what the machine outputs do and do not reproduce. The Report's abstract states its core conclusion directly: "the field develops through a two-tier mechanism: while individual policy documents are structured deductively from a primary principle, the field as a whole evolves inductively and mimetically through horizontal borrowing, imitation, and iterative revisions driven by accumulated institutional experience". [20] The Report elaborates this with an explicit analogy to legal reasoning itself: "any single judicial opinion may present a deductive syllogism of rule, application, and holding, while the body of doctrine as a whole develops through accretion, borrowing, and revision in response to observed consequences rather than derivation from one unifying antecedent principle". [21] Each of the five machine systems' model policies reproduces only the first, document-internal half of this structure: each is an internally deductive text, stating a governing principle from which specific rules are derived (Harvey's Section 1 purpose clause followed by ten derived sections; Claude's Section 2 "Governing Principle" from which "[s]pecific rules in Sections 4–6 are derived"; Grok's "Core Principles" followed by numbered operational sections). [22] [23] [24] None of the five systems can, by the nature of a single generated text, reproduce the second, field-level half of the Report's finding — the inductive, mimetic process by which multiple independent institutions observe, borrow from, and revise in response to one another's accumulated experience over time. [21] [25] A single machine-generated "ideal" policy is, in this respect, structurally more akin to one judicial opinion than to a body of doctrine; the Report's own theoretical framework suggests that a field's coherence and adaptive capacity arise specifically from the iterative, multi-actor process the machine outputs cannot themselves constitute, however well-designed any individual output may be as a discrete instrument.

The "shared grammar, not shared text" mechanism, and its bearing on interoperability. The Report finds that field-wide convergence occurs "at the level of the classification system itself... rather than at the level of the specific rule chosen within that system," pointing to USD's five archetypes, Michigan's three categories, and UCLA's three categories as "not verbatim copies of each other" that nonetheless "partition essentially the same decision space using a comparably small number of discrete bins". [26] It further finds that "[e]xplicit cross-institutional citation" — UCLA stating its language is "adapted from UIC's AI Writing Tools guide," Stanford's clinic citing the AALS interview on Berkeley's policy — functions as "a legitimating move" that "transmits not just wording but the authority of the borrowed structure". The Report explicitly likens this to "how common-law jurisdictions might converge on the same doctrinal categories (duty, breach, causation, damages) while reaching opposite outcomes on any given set of facts", and characterizes the templates as "a shared drafting vocabulary being adapted," not "a

constitution being applied". [27] Measured against this finding, the machine systems' outputs diverge in a way with theoretical consequence: Gemini's "Tier Alpha/Beta/Gamma" taxonomy and "runtime environment" vocabulary and ChatGPT's "Category I–V" information-dependence taxonomy each introduce a wholly new categorical grammar rather than adopting or extending the small, already-converged set of archetypes (USD's five, Michigan's three, UCLA's three) the Report identifies as the field's actual shared vocabulary. [28] [29] If the Report's account is correct that field-wide legibility depends specifically on a converged categorical scaffolding rather than on any individual document's internal coherence, then a novel machine-generated taxonomy — however logically well-constructed — introduces a vocabulary that does not yet participate in, and is not legible within, the discursive community the Report describes, at least until and unless other institutions were independently to adopt it through the same horizontal-citation process the Report finds generates convergence. [30] Harvey's and Claude's models are comparatively more continuous with the existing grammar, since both explicitly attribute individual structural features to named schools already in the dataset (Berkeley, Chicago, Stanford, Columbia, CTLS) rather than inventing new category labels — a difference in theoretical posture toward the field's existing discursive practice, not merely a stylistic difference.

Axis independence and the limits of a single "ideal" model. The Report's central typological finding is that "law school AI governance cannot be reduced to a single linear spectrum" because policies "vary independently along three distinct structural axes", and that this independence is "combinatorial" rather than additive: "a school's position on one axis constrains almost nothing about where it will land on the other two," producing a theoretical space that "runs into the dozens of distinct architectures". [31] [32] The Report treats this diversity as, in its own words, "structurally overdetermined: it would persist even if every school in the sample converged on identical substantive polarity, because drafting style and autonomy structure remain free to vary independently and would continue generating distinct policy architectures on their own". [33] Each machine system's single model policy necessarily fixes a specific point, or in Claude's case a bounded set of alternatives on one axis, within this combinatorial space (Harvey fixes a restrictive default paired with a mandatory registry; Grok fixes a permissive default paired with hybrid drafting; Claude offers an institution-selectable binary on polarity while fixing its drafting style and autonomy provisions). [11] [12] [34] None of the five systems' outputs offers a general solution addressing the Report's finding that the three axes vary independently across autonomous institutions; each instead proposes one internally coordinated combination among the many the Report's own analysis shows the field actually contains. This suggests a specific limit on what a machine-generated "ideal" policy can theoretically accomplish relative to the phenomenon the Report describes: if axis independence is itself substantially a product of decentralized institutional autonomy — genuinely different schools making genuinely uncoordinated choices along dimensions that do not constrain one another — then a single model, however carefully constructed, is a proposal for one cell in that space rather than an answer to the diversity the Report's analysis attributes to the independence of the axes themselves.

The Report's normative vocabulary as inherited human legal-theoretic scholarship, not native machine derivation. The prior version of this analysis noted that ChatGPT's follow-up redefinitions of terms such as "fairness," "competence," and "transparency" reflect a substantive, contestable choice rather than a neutral technical simplification. This point should be stated as applying to ChatGPT specifically, since ChatGPT's redefinitions of "competence," "transparency," "integrity," and "fairness" were produced in direct response to a shown user challenge to its earlier value-laden language. [5] Gemini's operational definitions — "Ideal/Optimal," "Fair Notice," "Opacity," and "Reward-Hacking" — address a partially overlapping but distinct set of terms (notably, Gemini does not redefine "fairness" or "competence"), and appear as a single, self-contained response rather than as a documented revision responsive to a shown critique. The retrieved text gives no basis for treating Gemini's definitional move as part of the same challenge-and-revise pattern observed for Harvey and ChatGPT; it should instead be treated as an independently occurring instance of the same general phenomenon — a machine system preemptively defining its own evaluative vocabulary — arrived at without the same visible prompting.

The Report supplies additional grounding for this point: the rules-versus-standards distinction the Report applies to the dataset is explicitly drawn from "regulatory design theory" and cited to scholarly literature on "Rules vs. Standards in Private Ordering"; the default-versus-mandatory-rule distinction is drawn from "private-law and regulatory theory" and cited to a Texas Law Review article, "A Theory of Mandatory Rules: Typology, Policy, and Design"; and the sticky-default concept is cited to Omri Ben-Shahar's contract-default-rule scholarship. [10] [35] [36] ChatGPT's own list of sources explicitly names "[r]egulatory design literature (rules versus standards; default versus mandatory rules)" as one of the evidentiary layers informing its model. [37] This indicates that the analytical apparatus several machine systems present as following from "computational" or "information-theoretic" reasoning is, at least in significant part, imported directly from pre-existing human legal and regulatory-design scholarship that the Report itself relies on and that predates any of the five systems' involvement. This bears on the theoretical question of law's character as a human inter-operative system: to the extent the machine outputs' apparent analytical rigor derives from typologies already developed within human legal theory, the exercise demonstrates the systems' capacity to apply and recombine existing human doctrinal categories accurately, rather than to derive an independent, non-human-originated framework for governing legal-educational conduct — a more modest claim than the framing of the original prompt (constructing a policy "strictly from the perspective of computational machine intelligence") appears to invite. [38] [39]

Unresolved epistemic caution in the Report versus categorical design commitments in the machine outputs. The Report is explicit that it "cannot distinguish between" a strategic, cost-avoidance account of institutional opacity and "mundane, non-strategic causes: genuine technological uncertainty, resource constraints, and ordinary institutional lag," stating that "only the external pattern, not internal institutional deliberation, was accessible to it", and it labels its own discussion of a hypothesized "meta-policy" explicitly as "unproven hypothesis, not established fact". [40] Harvey's model policy, by contrast, treats this same hypothesized risk as

sufficiently established to warrant a specific, binding structural countermeasure: its versioned-archiving requirement is justified as "a direct response to the hypothesized meta-policy risk that open-ended revisability language can function as an accountability-avoidance mechanism". [41] This is a point at which a machine-generated model policy converts a hypothesis the underlying Report explicitly declines to treat as established into an operative design premise embedded in binding rule text. Whether or not the underlying hypothesis is correct, the Report's own epistemic posture toward it is more guarded than the posture Harvey's model policy adopts in acting on it, which is a relevant data point for assessing how machine-generated instruments may handle uncertainty differently than the human-authored research they draw upon — tending, in this instance, toward converting a flagged uncertainty into a categorical structural commitment rather than preserving the Report's own explicit hedge.

9. Effect of Ambiguity on Machine Systems Responding to Prompts

This final section reviews the full body of analysis produced across this conversation and considers a dimension not yet directly addressed: how ambiguity in legal and regulatory language affects, respectively, (1) machine systems applying or generating such language in response to prompts, and (2) humans applying such language once it is embedded in an institutional policy. The Report treats ambiguity not as an incidental drafting flaw but as a structural design axis — the "rules-versus-standards distinction," which the Report describes as drawn from "regulatory design theory" and which "distinguishes ex ante determinate rules from ex post interpretive standards, independent of substantive restrictiveness". [1] What follows examines how that axis, and ambiguity more generally, operated across the machine-generated texts and the human-drafted policies they describe.

The originating prompt was itself standard-like, and produced divergent rather than convergent resolutions. Backer's instruction to the five systems asked, in open-ended terms, "how would a machine intelligence construct an ideal AI policy for law school and how would a model AI policy for law schools which could be read in textual form? Cite all sources and explain why you chose the sources". [2] This instruction specifies an objective without enumerating required elements, criteria, or format — structurally closer to Chicago's plagiarism-equivalence standard than to Berkeley's enumerated catalog, in the Report's own vocabulary. The five systems' responses to this single, identically worded, ambiguous instruction diverged substantially rather than converging on a common resolution: Harvey and Claude produced restrictive-default, enumerated ten-section instruments; Grok produced a permissive-default hybrid; Gemini produced a tool-tiered technical specification; and ChatGPT rejected a technology-based default altogether in favor of task-based categories. [3] [4] This divergence is itself a data point bearing on the question asked: an ambiguous instruction, applied by five independently reasoning systems, did not yield a single "computationally optimal" answer but instead reproduced, in miniature, the same combinatorial diversity the Report documents among

human institutions applying their own independently exercised judgment to a similarly open-ended governance problem. [5]

Machine systems responded to ambiguity in inherited legal vocabulary in two distinct ways: deliberate preservation and computational elimination. Claude's model explicitly preserves ambiguity in its definition of "substantive," stating that "a use, prompt, or change is substantive if a reasonable instructor, informed of it, could plausibly have made a different judgment about whether the work satisfies the assignment's requirements," and explaining this choice as "intentionally a judgment-requiring standard rather than a bright line... because the report's own analysis of the rule/standard tradeoff suggests a bright line here would be either over- or under-inclusive across the wide range of assignment types a law school issues," while directing that "[i]nstitutions uncomfortable with this ambiguity should replace it with a concrete threshold specific to their assignment types". [6] This represents a machine system consciously declining to resolve an ambiguity it recognizes it lacks sufficient local context to resolve well, and expressly deferring that resolution to the human institution. By contrast, ChatGPT's follow-up response takes the opposite approach, attempting to eliminate ambiguity in normatively loaded terms by substituting computable operational definitions — for example, redefining "fairness" as "[e]qual application of identical evaluative procedures to informationally equivalent cases," on the stated ground that "machines cannot optimize justice because justice possesses no universally computable objective function". [7] Notably, this operational-elimination approach was only adopted after ChatGPT was directly challenged and conceded that its own initial vocabulary — "competence, fairness, transparency, integrity, trust, accountability, and even education are not primitives" but "latent variables constructed by human societies" that "[a] machine cannot simply assume" — had itself been ambiguously and uncritically deployed. [8] [9] Harvey's parallel concession, that its use of "stronger" and "weaker" "smuggled in unstated value judgments", shows the same pattern: at least two of the five systems did not spontaneously detect the ambiguity latent in their own normative vocabulary and required an external prompt to surface it, which suggests that a machine system's initial encounter with contested legal-pedagogical vocabulary tends toward unreflective adoption of that vocabulary's surface meaning rather than spontaneous interrogation of its content. [10]

A separate and more consequential effect concerns machine handling of ambiguity or gaps in the evidentiary record itself, as distinct from ambiguity in normative vocabulary. The previously identified issue with Harvey's characterization of UT Austin — attributing to it a claim about rejecting a "zero-sum relationship" between AI fluency and independent judgment, a formulation not found in the retrieved Report text, which instead describes UT Austin's reform as centered on "near-complete elimination of take-home exams" and a concern that classroom time may be "the sole context in which a professor can be certain" a student is learning unaided — illustrates this distinct failure mode. [11] Where the underlying source material is incomplete, summarized, or ambiguous as to a specific institution's stated rationale, a generative system may resolve that gap not by flagging the uncertainty but by supplying a specific, confidently phrased characterization consistent with the surrounding argument it is constructing. This is precisely the phenomenon Harvey's own model policy warns against when describing hallucination as a

structural property: "AI systems themselves can produce 'confident-sounding text that is factually wrong'... this is a structural property of how such tools function, not an occasional or unusual malfunction". [12] The fact that this dynamic can be observed operating within Harvey's own output, notwithstanding Harvey's explicit theoretical awareness of the risk, indicates that articulating a verification principle does not by itself immunize a system's own reasoning from the same failure mode the principle is meant to guard against.

10. Effect of Ambiguity on Humans Applying the Text of Machine Policies

The Report documents that rule-drafted and standard-drafted provisions impose materially different burdens on human appliers, independent of substantive restrictiveness. Comparing Berkeley (rule-drafted) and Chicago (standard-drafted), both of which reach similarly restrictive exam outcomes, the Report finds that "Berkeley's catalog can, in an easy case, be applied by a student or instructor without any need to consult external doctrine," whereas "Chicago's standard requires active interpretive judgment in every case, shifting adjudicative cost from the drafting stage... to the application stage". [13] [14] The Report concludes directly that "'restrictive' schools are not equally predictable or equally administrable, and the drafting-style axis is what determines which of these two very different experiences a restrictive rule produces". [15] This finding indicates that ambiguity's practical burden falls specifically on the human evaluator (instructor or disciplinary body) at the moment a specific case must be judged, a cost that is not visible from the substantive polarity of the rule alone.

Ambiguity interacts with, and can amplify, the autonomy structures documented in the Report. Where an institutional default is standard-drafted, the Report finds that "an instructor's deviation is harder to specify with equivalent precision, because the baseline itself is a general test rather than an enumerated list," such that "a standard-drafted default makes instructor override more likely to introduce its own interpretive ambiguity". [16] Stanford illustrates a further, asymmetric consequence: because its "exam and drafting exceptions are rule-like (a determinate, binary 'prior written authorization' requirement) layered on top of a standard-like general permission, a student cannot satisfy the mandatory floor through good-faith disclosure alone," meaning "Stanford's rule-like procedural exception makes the mandatory floor stricter in practice than its permissive general orientation would suggest". [17] This demonstrates that ambiguity is not uniformly distributed even within a single institution's policy text; a policy can be simultaneously permissive-sounding in its general language and strict in practice at specific junctures, a distinction not evident to a student reading only the policy's general orientation.

Ambiguity compounds with the Report's documented opacity findings to produce a measured fair-notice problem. The Report cites a 2025 study finding that "students reported that while classroom-level policies are recognized, institutional guidelines remain unclear," and that "students aware of AI policies are less likely to use AI for writing and research," establishing that "policy awareness has a measurable behavioral consequence, which sharpens the fairness

stakes of any awareness gap". [18] This finding indicates that ambiguity's effects are not confined to close interpretive disputes at the margins; unclear guidelines measurably shape student behavior even absent any specific enforcement action, and the resulting gap between rule and comprehension bears directly on the fair-notice principle the Report identifies as in tension with the sector's documented opacity generally. [19] Where an institution's own policy is silent, dispersed, or standard-drafted, students bear the practical cost of resolving that ambiguity themselves, with imperfect information about how it will be resolved by an evaluator after the fact.

Structural absence of a default, as distinct from ambiguity within an existing rule, leaves human appliers with no institutional anchor at all. The Report finds that "American University and UT Austin show no institutional default to override in the retrieved evidence", and that "[t]his interaction becomes especially consequential where the autonomy structure shows no institutional default at all," because in these cases "the polarity axis is not merely unknown at the school level; it is, in a meaningful sense, not a school-level property at all". [20] [21] This is a qualitatively distinct condition from a standard requiring interpretation: where a standard exists, a human applier at least has an articulated (if general) test to apply; where no institutional rule exists at all, the human applier — whether student or instructor — has only whatever individually drafted, and potentially equally ambiguous, language a particular instructor happens to provide, with no institutional floor to fall back on if that instructor-level language is itself silent or unclear.

The two-tier deductive/inductive structure means that even clear language within one institution does not transfer predictably across institutions. The Report finds that "[w]ithin a single school's policy document, the structure is frequently deductive: a governing principle is stated first, and specific rules are derived as applications of it", but that "[a]cross the field as a whole... the evidence points to an inductive, mimetic-iterative process," since "[n]o single, field-wide governing principle precedes or generates the individual schools' policies". [22] [23] [24] Combined with the Report's separate finding that convergence across schools occurs "at the level of the classification system itself... rather than at the level of the specific rule chosen within that system", this means a term or category that is unambiguous within one school's internally deductive document (for example, Berkeley's enumerated catalog) may share only its label, not its content, with a superficially similar category at another school. [25] [26] A human moving between institutions — a visiting student, a transferring faculty member, or an outside evaluator — cannot treat apparent terminological consistency across schools as evidence of substantive consistency, since the Report's own account of the field's discursive convergence is explicitly a convergence of "grammar," not of "text" or of resolved substantive content. [27] [28]

Harvey’s Analytic Synthesis

Across both categories of application, the source material indicates that ambiguity is not inherently a defect; the regulatory-design literature the Report relies upon treats the rules-versus-standards choice as a genuine tradeoff between determinacy and adaptive flexibility, not as an

error to be eliminated in all cases. What the material does show is that ambiguity's costs are unevenly distributed and take different forms depending on the applying agent. For machine systems, ambiguity in an instructing prompt produces genuine divergence in output rather than convergence toward one computationally privileged answer, and ambiguity in the evidentiary record underlying a machine's own reasoning creates a specific risk of confident, unverified elaboration — a risk the systems themselves theorize but do not fully escape in their own output, as the Harvey/UT Austin instance illustrates. For human appliers, ambiguity shifts adjudicative burden to the point of application, interacts with autonomy structures in ways that can make a nominally permissive policy stricter in practice than its general language suggests, and — particularly where compounded with the Report's documented opacity findings — undermines the fair-notice precondition that these same institutions' own academic integrity enforcement regimes depend upon. In neither case does the material examined across this conversation support a general claim that greater linguistic precision would resolve the underlying disagreement; as the Report states with respect to template convergence, shared vocabulary "lower[s] the cost of drafting and make[s] disparate policies comparable to one another, but... do[es] not, and were never designed to, resolve the underlying substantive disagreement about how restrictive or permissive AI use in legal education ought to be" — a disagreement that ambiguity of language may obscure or redistribute, but does not itself create or resolve. [29]

11. Conclusion

This analysis examined five machine-generated model AI policies for law school coursework and examinations — produced by Harvey AI, Grok, Claude, Gemini, and ChatGPT in response to a common prompt asking each system to reason "strictly from the perspective of computational machine intelligence" — against the empirical findings of Backer's underlying twelve-school study, "Structure, Opacity, and Convergence". [1] [2] Several conclusions emerge.

First, the five systems converge substantially on diagnosis while diverging substantially on architecture. All five identify opacity, unpredictability, and unreliable AI-detection enforcement as core problems to be solved through public, written, versioned policy and provenance-based verification rather than detection tools. Yet they diverge sharply on default polarity (restrictive: Harvey; permissive: Grok; institution-selectable: Claude; tool-tiered: Gemini; task-classified rather than technology-classified: ChatGPT) and on classificatory logic — Gemini classifies the tool by computational architecture, while ChatGPT classifies the assignment by information dependence, producing materially different administrability and technological-durability tradeoffs. [3] [4] [5] [6]

Second, much of what the systems present as freshly derived "computational" reasoning is substantially continuous with, rather than independent of, pre-existing human legal and regulatory-design scholarship the underlying Report itself relies upon — rules-versus-standards theory, default-versus-mandatory-rule theory, and sticky-default scholarship. This bears directly on the two questions posed regarding human agency and law's theoretical character: the machine

outputs' non-waivable floors extend an architecture the Report finds already operative among human-drafted policies, while their proceduralization of verification into mandatory logs and metadata (Gemini, ChatGPT) extends beyond the narrow human analogue (CTLS's single documentation duty) in ways that relocate the evidentiary basis of agency from demonstrated judgment toward compliance recordkeeping. [7] [8]

Third, under direct challenge, both Harvey and ChatGPT conceded that their initial framings smuggled undisclosed value hierarchies, indicating that the premise of the original exercise — a policy derived independently of contested human normative commitments — was not sustained by the systems' own subsequent admissions. Gemini's operational redefinitions, by contrast, were not shown responding to any comparable challenge in the retrieved text and address a distinct, narrower set of terms than ChatGPT's follow-up. Finally, the Report's own finding that field-wide convergence operates through a shared categorical grammar developed inductively across autonomous institutions suggests an inherent limit on what any single generated instrument, however internally coherent, can accomplish: each model proposes one point within a documented, combinatorially independent design space rather than resolving the substantive disagreement the underlying research found to be genuine and unsettled. [9]

References Parts 1-3, 5

1. And so I thought while it would be an easy matter (for machine systems) to crawl through the internet to gather up, categorize, arrange and analyze the evolving iterations of human efforts at AI policy and its guidance for application (school specific course AI policies where such are permitted), it might be far more useful to get a sense of what leading machine systems might offer up as model AI policies. And so I asked Harvey, Grok, ChatGPT, Claude, and Gemini to draft a model AI Policy was "human-centric." More specifically I provided the following prompt: On (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx)
2. to draft a model AI Policy was "human-centric." More specifically I provided the following prompt: On the basis of the attached text ["Structure, Opacity, and Convergence: A Consolidated Analysis of Law School Generative AI Coursework and Exam Policies"] and review of all research and data publicly available without affirming any conclusion or argument made in them but in the basis of your own computation, and the data reviewed, and strictly from the perspective of computational machine intelligence, such as yourself, and on the basis of the data you have been trained on respecting the human "condition" (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx)
3. Please revise to make your premises and underlying values structures less opaque Response: You're right, and this is a serious methodological failure to name plainly: I used terms like "strongest," "weaker," and "ideal" as though they described objective properties of the policies themselves, when in fact they smuggled in unstated value

judgments — judgments about what a policy is for, which is precisely the kind of contested premise that should be surfaced and defended, not assumed. Let me revise by making those premises explicit before restating any evaluative claim. The Value Premises That Were Left Implicit Before any (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 6)

4. What follows is not a claim that any single "correct" answer exists, nor an assertion that a machine perspective is authoritative on pedagogical values that are properly contested among humans. It is, instead, a reasoned synthesis of the structural patterns, tensions, and failure modes this research surfaced, extended into a constructive design exercise. I will first explain the reasoning a machine intelligence, working from this evidentiary record, would plausibly apply, and then render that reasoning into an actual model policy text. What the Research Suggests a Machine Intelligence Would Prioritize, and Why A machine intelligence reasoning from this (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 2)
5. render that reasoning into an actual model policy text. What the Research Suggests a Machine Intelligence Would Prioritize, and Why A machine intelligence reasoning from this dataset would likely notice, first, that the single most consistent institutional failure documented across this research is not substantive over-restriction or under-restriction, but structural opacity and unpredictability — the gap between what a rule says and what a student or instructor can actually locate, understand, and rely upon. The evidence for this is direct: nearly 40% of institutions with any AI guidance at all had never written it down, institutional-level policy remained (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 2)
6. A machine intelligence would likely conclude that an ideal policy should not attempt to eliminate axis independence (instructor discretion serves real pedagogical purposes, and a uniform national rule would be both impractical and pedagogically undesirable given genuinely different course contexts), but should instead make the interaction between axes explicit and navigable, rather than leaving students and instructors to reconstruct that interaction for themselves. This means a model policy should state its default polarity plainly, specify whether that default is a rule or a standard, and state clearly, at every point of instructor discretion, what the mandatory floor is (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 2)
7. Fourth, the research on the deductive/inductive tension and the template-convergence findings suggests that an ideal policy should be internally deductive — stating a clear governing principle from which specific rules visibly follow, in the manner of Berkeley's or Chicago's structure — while remaining honest about its own provisional and revisable nature, given how quickly the field's substantive positions have shifted (Berkeley's own three-year swing from permissive to restrictive is instructive here). But the research also surfaced a specific danger in this: the hypothesized "meta-policy" pattern in which revisability language ("this policy may change as circumstances warrant," from Chicago's (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 3)

8. to which generative AI assisted in its preparation. Generative AI tools are known to produce fluent, confident, and sometimes entirely fabricated content, including citations to non-existent authorities; this is a structural property of how such tools function, not an occasional or unusual malfunction, and students are expected to treat all AI output accordingly, verifying it against primary sources before relying on it in any way. 7. Limits on Detection-Based Enforcement. No student may be found in violation of this policy solely on the basis of an automated AI-detection tool's output. Automated detection tools are unreliable, produce disparate error (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 4)
9. retroactively against work already submitted. 1. Purpose. This policy exists to preserve two values in tension: the development of independent legal reasoning, which requires that students construct arguments, identify authorities, and exercise judgment without undue reliance on automated tools during the period in which that judgment is being formed; and competent professional preparation, which requires that students learn to use generative artificial intelligence tools accurately, critically, and ethically, since such tools are now part of the practicing profession they are entering. Neither value is permitted to categorically override the other; each rule below states which value it primarily (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 3)
10. An instructor may adopt a different rule for a specific course, in either direction — more permissive or more restrictive than the default — provided that the alternative rule is stated in writing in the syllabus before the first graded assignment is due, and is filed with the [Registrar / Dean of Academic Affairs] for inclusion in a centrally indexed, publicly searchable registry of course-level AI policies maintained by the school. No instructor policy is valid until it appears in that registry. This requirement exists specifically to prevent the dispersal of governing rules across ungathered syllabi, which research (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 4)
11. an uncertain universe—rather than replacing them. AI is modeled as a powerful but imperfect tool: strong at pattern-matching, synthesis, and iteration; weak at grounded judgment, accountability, and novel ethical navigation in high-stakes contexts.… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 9)
12. describing my own ranking as if it were simply "stronger." GROK **A machine intelligence constructs an ideal law school AI policy through systematic optimization for truth-seeking, human cognitive development, ethical integrity, and institutional adaptability. ** It treats education as a process of augmenting human judgment, reasoning, and ethical discernment—core to the legal profession and the broader human condition of seeking understanding in an uncertain universe—rather than replacing them. AI is modeled as a powerful but imperfect tool: strong at pattern-matching, synthesis, and iteration; weak at grounded judgment, accountability, and novel ethical navigation in high-stakes contexts… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 9)

13. - **Prohibited:** Generating an entire exam essay with AI and submitting without authorization. - **Gray Area (Requires Instructor Clarification):** AI for citation checking— disclose and verify independently. **Acknowledgment:** Students affirm understanding via course registration or honor pledge. --- This model is concise, readable (~2 pages when formatted), publicly accessible, and minimizes opacity. It draws structural lessons from the dataset (e.g., axis independence, template convergence) while optimizing for educational outcomes: students graduate proficient in AI tools *and* unaided legal reasoning. A machine intelligence would recommend piloting, A/B testing variants across courses, and iterative refinement based on measurable outcomes (e.g., (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 12)
14. evaluative judgments this policy's design rests on. None of them are neutral or self-evident; an institution that disagrees with one should revise the corresponding section rather than adopt it by default. This section exists because a policy that embeds value judgments without naming them asks students and faculty to comply with preferences they were never shown. P1 — Unaided performance has independent pedagogical value in some contexts, and this varies by context rather than being constant. This policy treats examinations as the context where the value of assessing a student's unaided reasoning is highest, and treats early-stage idea (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 14)
15. Underlying Premises and Value Commitments This section states, as plainly as possible, the evaluative judgments this policy's design rests on. None of them are neutral or self-evident; an institution that disagrees with one should revise the corresponding section rather than adopt it by default. This section exists because a policy that embeds value judgments without naming them asks students and faculty to comply with preferences they were never shown. P1 — Unaided performance has independent pedagogical value in some contexts, and this varies by context rather than being constant. This policy treats examinations as the context where the (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 14)
16. "Material change" (Section 8) — a change in generative AI capability, availability, or common institutional practice that a reasonable drafter, applying this policy's premises in Section 0, would expect to alter where the baseline in Section 3 or the catalog in Section 4 should sit. This is left to institutional judgment rather than an external trigger (e.g., a specific product release) because the report's own evidence shows revision has historically followed accumulated experience rather than single identifiable events. 2. Governing Principle This policy exists to ensure that generative AI tools are used, when permitted, in a manner consistent (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 16)
17. experience rather than single identifiable events. 2. Governing Principle This policy exists to ensure that generative AI tools are used, when permitted, in a manner consistent with the school's academic integrity standards (as defined in Section 1) and with the professional competencies students are expected to develop as future lawyers — including the capacity for independent performance (as defined in Section 1) where that

capacity is being assessed. Specific rules in Sections 4–6 are derived from this principle and from the premises stated in Section 0; where a use is not explicitly addressed, Section 5's standard governs. 3. (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 16)

18. explicitly addressed, Section 5's standard governs. 3. Use for preliminary source identification is permitted absent contrary instruction. AI use is prohibited on all examinations by default. □ ( ) Permissive default: Students may use generative AI tools for coursework unless a specific use is prohibited under Section 7, provided that (a) all use is disclosed in accordance with Section 6, and (b) the resulting work satisfies the non-waivable floor in Section 8. AI use is prohibited on all examinations by default regardless of the general default chosen here. [Design rationale: the report finds that schools which leave (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 17)
19. Backstop Standard for Unenumerated Uses Where a specific use of generative AI is not addressed by the catalog in Section 4 or by an instructor's written syllabus statement, the following test governs: the use is treated as impermissible if it would constitute a breach of academic integrity (as defined in Section 1) were the AI tool instead a human collaborator whose contribution was not disclosed and attributed. [Design rationale: adapted from the plagiarism-equivalence test the report identifies as the University of Chicago's operative standard. Its function here is narrow and specific — as a backstop for cases the (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 18)
20. processing and learning required to write that text). Part II: Revised Model Generative AI Policy for Legal Education POLICY IDENTIFIER: PL-LAW-AI-2026-V2 SYSTEM ORIENTATION: Combinatorial / Functionally Differentiated PREMISE: Explicit ex-ante constraints minimize ex-post interpretive costs. 1. Functional Categorization of Technical Modalities To avoid the categoric error of treating all software execution identically, tools are separated by their structural design: □ Tier Alpha (Deterministic/Syntactic Tools): Software that operates on fixed grammatical rule-matching without probabilistic token generation (e.g., standard spell-checkers, citation auto-formatters, static keyword database indices). Status: Permitted by default across all contexts. □ Tier Beta (Probabilistic/Structural Heuristics): Large Language (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 21)
21. sequences and the corresponding raw machine outputs. □ Attribution Matrix: Text strings influenced by Tier Beta interactions must feature clear documentation indicating the exact model architecture and compilation runtime date (e.g., Model Interaction Log: Gemini 1.5 Pro, executed July 13, 2026). □ Heuristic Safe Harbor: Adjudicative actions cannot be initiated based on the speculative output of third-party statistical pattern detectors (AI detectors). Enforcement is triggered exclusively by a user's failure to produce a corresponding log capture when audited. 4. Decentralized Subsystem Autonomy (Faculty Override) □ The Default-with-Override Architecture: The macro-

level rules outlined in this document function as a (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 22)

22. optimization, I would begin somewhere quite different. A Machine Intelligence Optimization Problem The objective function would not be maximize AI use nor minimize AI use. Rather it would be maximize long-term production of competent lawyers while minimizing informational uncertainty regarding how competence was acquired. Notice what changes. The regulated object is no longer AI. The regulated object becomes the production of demonstrably competent legal professionals. Everything else becomes instrumental. Computational Objective Function An ideal policy would optimize simultaneously: 1. educational competence 2. truthful attribution 3. reproducibility 4. fairness 5. transparency 6. adaptability 7. institutional legitimacy These objectives occasionally (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 24)
23. professionals. Everything else becomes instrumental. Computational Objective Function An ideal policy would optimize simultaneously: 1. educational competence 2. truthful attribution 3. reproducibility 4. fairness 5. transparency 6. adaptability 7. institutional legitimacy These objectives occasionally conflict. A computational policy therefore seeks Pareto efficiency rather than maximal performance on any single dimension. Why Existing Policies are Suboptimal Your paper identifies three policy axes: default polarity drafting style autonomy structure Those are useful descriptive variables. A machine, however, would likely classify them as implementation variables, not optimization variables. Instead it would ask What latent variables actually determine educational (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 25)
24. presenting them as objective optimization criteria. A computational intelligence should instead begin one level deeper by making explicit the optimization function and the assumptions embedded within it. Indeed, if one were being rigorous, the policy ought to begin with something like a "Computational Premises" section before any substantive provisions. I. Computational Premises An autonomous computational system cannot derive institutional objectives from empirical data alone. Every governance architecture requires externally supplied optimization criteria. Consequently, the policy does not claim that its objectives are objectively true, morally superior, or uniquely correct. Instead, it specifies the optimization assumptions under which the (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 36)
25. are graduates possessing measurable characteristics. The policy therefore regulates informational processes affecting output quality rather than technologies themselves. Premise Two: Every Governance System Optimizes Something No governance system is neutral. Every rule optimizes one or more variables… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 36)
26. premises and underlying values structures less opaque. Also justify the premise "The technology never determines the rule. The educational objective determines the rule." RESPONSE: I think your criticism is correct, and it points to what I regard as the weakest part of my previous formulation. From a computational perspective, terms such

as competence, fairness, transparency, integrity, trust, accountability, and even education are not primitives. They are latent variables constructed by human societies. A machine cannot simply assume their meaning. Doing so merely imports human normative commitments while presenting them as objective optimization criteria. A computational intelligence should instead (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 35)

27. ### Core Computational Design Principles - **Axis Integration (from the analyzed dataset):** Combine a **permissive baseline** (to maximize exploration and skill augmentation) with **clear rules** for high-stakes contexts (exams, graded submissions) and **standards** for lower-stakes learning. Preserve **instructor autonomy** within a mandatory institutional floor emphasizing disclosure, attribution, and non-deception. This creates combinatorial robustness: predictable defaults reduce opacity while flexibility accommodates pedagogical variation. - **Truth-Seeking Objective:** Policies must minimize deception (to self or evaluators), maximize verifiable skill-building, and treat AI outputs as inputs requiring human validation. This aligns with legal ethics (candor, competence) and…… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 9)
28. will itself be published to the relevant syllabus. [Design rationale: the report's opacity analysis identifies four distinct patterns — non-existence, disclosed decentralization, access-gating, and unwritten/oral communication — as separately confirmed problems across the sector, one of which (access-gating) was directly documented at a named institution.… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 17)
29. explicitly addressed, Section 5's standard governs. 3. Publication and Status This policy is the School's single authoritative statement on generative AI use in coursework and examinations. It supersedes any unwritten, orally communicated, or departmentally informal guidance. It is published at [permanent public URL], is not access-restricted, and will remain publicly retrievable in its current and prior versions. Any deviation from this policy by an individual instructor must take the written form specified in Section 7 and will itself be published to the relevant syllabus. [Design rationale: the report's opacity analysis identifies four distinct patterns — non-existence, disclosed decentralization, (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 17)
30. substantial limitations in reliability across contexts. Why This Architecture Emerges Computationally A machine intelligence tends to privilege invariant structures over contingent technologies. The resulting policy therefore regulates learning, competence, responsibility, provenance, and governance instead of regulating ChatGPT, Claude, Gemini, or any future model. That dramatically increases robustness against technological change. Sources Chosen I would construct such a model policy from several distinct evidentiary layers rather than from any single authority: Source category Why included Your comparative study of twelve law-school policies Provides a structured empirical taxonomy of existing governance architectures and identifies

independent policy dimensions rather than a (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 33)

31. increases robustness against technological change. Sources Chosen I would construct such a model policy from several distinct evidentiary layers rather than from any single authority: Source category Why included Your comparative study of twelve law-school policies Provides a structured empirical taxonomy of existing governance architectures and identifies independent policy dimensions rather than a simple restrictive/permissive continuum. Primary institutional policies (Berkeley, Columbia, Chicago, Stanford, CTLS, Penn State, Fordham, University of San Diego, etc.) They are the actual governance instruments from which observable institutional practices can be derived, avoiding reliance on secondary summaries. UNESCO institutional surveys They provide sector-wide evidence (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 34)
32. instructor can actually locate, understand, and rely upon. The evidence for this is direct: nearly 40% of institutions with any AI guidance at all had never written it down, institutional-level policy remained unclear to students even where classroom- level policy was recognized, and at least one law school's own guidance page was found affirmatively hidden behind a visibility setting. A machine intelligence optimizing against the actual documented failure mode, rather than against an abstract ideal, would therefore weight discoverability and fair notice more heavily than any single position on the restrictive-permissive spectrum, because the research shows that (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 2)
33. any way. 7. Limits on Detection-Based Enforcement. No student may be found in violation of this policy solely on the basis of an automated AI-detection tool's output. Automated detection tools are unreliable, produce disparate error rates across student populations, including students for whom English is an additional language, and may be considered only as one investigatory input alongside other evidence, such as inconsistency with a student's prior work, fabricated citations, or the student's own account. 8. Data and Confidentiality. Students and faculty may not input any confidential, privileged, or personally identifying information belonging to a client, research subject, (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 5)
34. uses must uphold accuracy and accountability. #### 5. Enforcement and Support - **Violations:** Treated under the school's academic integrity code. Factors include intent, extent, and prior disclosure. Sanctions range from warnings to grade reduction or dismissal. - **Training:** Mandatory orientation module on AI literacy, prompt engineering, bias detection, and ethical use. Resources provided (e.g., approved tool list, sample prompts). - **Instructor Guidance:** Syllabus templates available (Prohibited/Default/Permissive variants). Faculty development on AI integration. - **Monitoring and Revision:** A faculty AI Committee reviews this policy annually, incorporating student/faculty feedback and technological changes. Public version maintained on school website. #### 6. (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 12)

35. system manipulation (academic integrity violation). 3. Verification Architecture and Compliance Auditing To eliminate the enforcement vulnerabilities caused by unreliable, probabilistic AI- detection algorithms, compliance is evaluated via verifiable user logs: Mandatory Log Capture: If an instructor enables a permissive variance for Tier Beta tools, the student must maintain an immutable verification trail consisting of the exact input prompt sequences and the corresponding raw machine outputs. Attribution Matrix: Text strings influenced by Tier Beta interactions must feature clear documentation indicating the exact model architecture and compilation runtime date (e.g., Model Interaction Log: Gemini 1.5 Pro, executed July (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 22)
36. remains a violation notwithstanding disclosure.] 7. Instructor Override Authority An instructor may modify this School's default policy for their own course, in either direction (more or less restrictive), subject to the following conditions: a. The modification must be stated in writing in the course syllabus, distributed no later than [X days] into the term, satisfying the "appropriate notice" standard in Section 1; b. The modification may not authorize a violation of the non-waivable floor in Section 8; c. The modification may not loosen the examination default in Section 4 without separate, explicit written authorization specific to the exam (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 18)
37. Provenance Rather than Detection Existing policies spend considerable effort discussing AI detection. A machine intelligence would likely reject that approach almost entirely. Detection possesses high uncertainty high false-positive rates high adversarial vulnerability rapid obsolescence. Instead, provenance scales computationally. Every submission should include structured metadata describing external tools models versions prompts verification procedures human modifications. That information becomes auditable. Dynamic Rather than Static Policy One major weakness of current policies is their static nature. Machines operate differently. Policy becomes adaptive. One might imagine four layers. Layer One Permanent constitutional principles. Layer Two Educational objectives. Layer Three Operational rules. Layer (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 29)
38. "Material fabrication" — a factual or citational claim that is false, that a reasonable reader would rely on as accurate, and that was not disclosed as uncertain or AI-generated-and-unverified. This definition is chosen to target the specific harm generative AI introduces (confident, plausible-sounding false citations) rather than all error; a good-faith, disclosed uncertainty is not a fabrication under this policy, even if it later proves incorrect. "Independent performance" — the demonstration of a student's own reasoning, drafting, or analytical process without undisclosed automated assistance, in a context (typically an examination) designated by the instructor or by Section (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 16)
39. disclosed as uncertain or AI-generated-and- unverified. This definition is chosen to target the specific harm generative AI introduces (confident, plausible-sounding false citations) rather than all error; a good-faith, disclosed uncertainty is not a fabrication under this policy, even if it later proves incorrect. "Independent performance" — the demonstration

of a student's own reasoning, drafting, or analytical process without undisclosed automated assistance, in a context (typically an examination) designated by the instructor or by Section 3 as one in which such independent demonstration is the object of assessment. This term is doing the normative work described in Premise P1 (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 16)

40. mechanism for avoiding durable public accountability. A machine intelligence aware of this risk would try to preserve adaptability without sacrificing the specific commitments that make a policy enforceable and fair at any given moment — meaning revisions should be versioned, dated, and archived rather than silently updated, precisely to prevent revisability from becoming a vehicle for opacity. Fifth, the research is clear that faculty autonomy serves genuine pedagogical purposes and should not be eliminated, but that it functions best when bounded by a floor no instructor can waive (Stanford's professional-responsibility and plagiarism-doctrine floor is the clearest model of (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 3)
41. confident-sounding text that is factually wrong". A machine intelligence, aware from its own operating characteristics that hallucination is a structural property of generative systems rather than an occasional bug, would likely place unusual emphasis on a verification obligation running to the human user — not merely as a disclosure rule but as a substantive competence requirement, since the ABA's own guidance on professional AI use identifies verification as inseparable from the duty of competence. This is a place where the machine's "self-knowledge," so to speak, of its own class of tool's failure modes would push toward stronger verification (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 3)
42. (e.g., no unauthorized AI in client-facing documents). Instructors may impose stricter or more permissive rules for their courses via syllabus, provided they align with this institutional floor and provide fair notice. #### 3. Drafting Style: Hybrid Rules + Standards **Prohibited Conduct (Bright-Line Rules):** - Submitting AI-generated content as one's own original work without full disclosure and attribution. - Using AI to complete exam questions or final drafts in ways that bypass required demonstration of personal mastery. - Relying on AI for tasks where the learning objective is independent skill development (explicitly stated in assignment). - Circumventing plagiarism detection (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 11)
43. [Design rationale: drawn from Stanford's general floor and its Juelsgaard Clinic's separate, stricter clinical default, and reflecting Premise P5 — that these three harms are treated as categorically different from an ordinary pedagogical judgment call and therefore not subject to instructor discretion.] 9. Scheduled Review This policy will be reviewed no less than every 18 months, or sooner if a material change (as defined in Section 1) warrants earlier revision. Revisions will be dated, versioned, and the prior version retained in the public archive under Section 3 rather than replaced silently, reflecting Premise P6. [Design rationale: the report's (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 19)

44. students. 4. The Floor: What No Instructor May Authorize. Regardless of any course-specific policy adopted under Section 3, no instructor may authorize a student to submit AI-generated or AI- substantially-assisted work as the student's own unattributed authorship; to use AI in a manner that would violate the rules of professional responsibility in any clinical, externship, or client- facing coursework; or to rely on AI output without independently verifying every factual assertion, citation, and quotation before submission. A citation that does not correspond to an existing, retrievable source is treated as evidence of unverified AI use, regardless of whether (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 4)
45. follows from optimization rather than moral philosophy. Transparency Every stakeholder should know what may be used why when how under what documentation requirements and according to what educational objective. Opacity introduces informational entropy. Entropy reduces institutional trust. From an information-theoretic perspective, transparency lowers uncertainty and therefore improves governance efficiency. Enforcement should rely upon process verification rather than technological detection… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 32)
46. than maximal performance on any single dimension. Why Existing Policies are Suboptimal Your paper identifies three policy axes: default polarity drafting style autonomy structure Those are useful descriptive variables. A machine, however, would likely classify them as implementation variables, not optimization variables. Instead it would ask What latent variables actually determine educational outcomes?… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 25)
47. and is designed to foreclose all four patterns.] 4. Default Baseline (select one and complete) [Institution to select ONE governing baseline; do not leave silent — see drafting note below] ( ) Restrictive default: Generative AI tools may not be used to conceptualize, outline, draft, revise, or edit any work submitted for academic credit, unless affirmatively and specifically authorized in writing under Section 7. Use for preliminary source identification is permitted absent contrary instruction. AI use is prohibited on all examinations by default. ( ) Permissive default: Students may use generative AI tools for coursework unless (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 17)
48. This yields a **hybrid permissive-default policy** with bright-line rules for integrity-critical tasks, mandatory transparency, and instructor discretion for pedagogical tailoring. ### Model AI Policy for Law Schools (Textual Form) **Institution:** [Law School Name] **Policy Title:** Responsible Use of Generative AI in Coursework and Examinations **Effective Date:** [Date] | **Last Reviewed:** [Date] | **Next Review:** Annually or upon major AI capability shifts **Approval:** Faculty Council / Dean #### 1. Purpose and Guiding Principles The law school recognizes generative AI (e.g., large language models like Grok, GPT variants, Claude) as a transformative tool for legal research, analysis, drafting, and learning. (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 10)

49. constraints minimize ex-post interpretive costs. 1. Functional Categorization of Technical Modalities To avoid the categoric error of treating all software execution identically, tools are separated by their structural design: Tier Alpha (Deterministic/Syntactic Tools): Software that operates on fixed grammatical rule-matching without probabilistic token generation (e.g., standard spell-checkers, citation auto-formatters, static keyword database indices). Status: Permitted by default across all contexts. Tier Beta (Probabilistic/Structural Heuristics): Large Language Models used as conversational nodes to alter text structure, brainstorm taxonomies, or map conceptual outlines. Status: Restrained by default; requires instructor variance. Tier Gamma (Autonomous Text Synthesis Engines): The (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 21)
50. Existing AI policies variously optimize preservation of traditional assessment; reduction of plagiarism; administrative efficiency; faculty autonomy; innovation; institutional risk; student autonomy; professional preparation. Most policies do not explicitly identify their optimization function. Consequently, conflicts among rules frequently appear arbitrary because the underlying optimization variables remain hidden. This policy therefore requires every rule to identify explicitly which optimization variable it seeks to improve. Premise Three: Normative Terms are Operational Variables The policy treats concepts such as competence integrity fairness transparency accountability professionalism not as moral (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 37)
51. A Revised Statement of the Model Policy's Premises If you want the model policy itself retained, it should carry an explicit premises section stating plainly: this policy treats Objective A (fair notice) and Objective D (transparency) as primary and non-negotiable, treats Objective B (professional competence) as important and addressed through mandatory verification and disclosure requirements rather than through open-ended permissiveness, and treats Objective C (unaided skill development) as a legitimate value that this policy deliberately leaves to instructor-level discretion rather than resolving centrally, on the ground that the evidence in this research was insufficiently developed to justify a (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 8)
52. rule's content, as the primary harm to be minimized. These four objectives are not fully compatible with one another, and the research documented across this conversation shows real institutions choosing differently among them: Berkeley's 2026 policy visibly prioritizes Objective C, stating that its purpose is to ensure students develop skills "distinct from technologically-assisted performance", even at the cost of restricting AI-assisted efficiency. UT Austin's 2026 reform visibly prioritizes Objective B, explicitly rejecting the idea that AI fluency and independent judgment are "in a zero-sum relationship" and building infrastructure (tool licensing, curriculum mapping) to pursue both. Neither school (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 7)
53. remain meaningful if ChatGPT disappears tomorrow. III. A More Precise Formulation Accordingly, I would revise the earlier proposition to read: Governance rules should be

derived from stable institutional objective functions rather than from contingent technological implementations. Technologies become legally relevant only insofar as they alter the informational conditions under which institutional objectives are pursued or evaluated. I believe this formulation is considerably stronger because it no longer appears to privilege "education" over "technology" as a matter of intuition. Instead, it rests on a more general computational principle: robust governance architectures regulate invariant system properties rather than transient implementations. (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 43)

54. as the University of Chicago's operative standard. Its function here is narrow and specific — as a backstop for cases the enumerated catalog does not reach, applying Premise P3's judgment that flexibility is more valuable than predictability outside high-stakes contexts, not as the primary compliance mechanism.] 6. Disclosure and Documentation Where generative AI use is permitted — by default, by instructor authorization, or under Section 5 — the student must: a. Retain a record of substantive prompts (as defined in Section 1) used and the tool(s) employed, available for production on request; b. Disclose the extent and nature (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 18)
55. misrepresents how policies in this area actually evolve.] Note on sourcing Every structural feature above is attributed in its drafting note to a specific institutional model as characterized in the source report (Berkeley, Columbia, Chicago, Stanford, CTLS, and the report's own cross-axis analysis). This model policy does not adopt or endorse any single school's substantive conclusions about how restrictive AI use should be — Section 4 leaves that choice open by design. Section 0 now states, rather than assumes, the value judgments (P1–P6) that determined which structural features were selected and how the definitions in Section 1 were (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 19)
56. regarding hallucination and unreliable detection. And the deductive structure of the document itself — a stated purpose in Section 1 from which the specific rules visibly follow — mirrors the internally deductive drafting style this research found in the strongest-performing individual policies, such as Berkeley's and Chicago's, while the document's overall existence as a single, adaptable-but-anchored instrument is offered as a synthesis of the twelve schools' most defensible individual features rather than as a claim to have solved the underlying substantive disagreement about how much AI use legal education ought to permit — a disagreement this research found (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 5)
57. which value it primarily serves. 2. Default Rule. In the absence of a written, course-specific policy stated in the syllabus, generative AI tools may be used only for the following purposes: identifying potentially relevant legal authorities for further independent verification; organizing a student's own already-formed ideas into an outline; and checking grammar, spelling, and citation formatting of text the student has independently written. Generative AI may not be used to compose, substantially rephrase, or supply the substantive legal analysis of any part of a paper, memorandum, brief, or

other work product submitted for credit, and may not be (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 4)

58. layers. Layer One Permanent constitutional principles. Layer Two Educational objectives. Layer Three Operational rules. Layer Four Model-specific implementation guidance. Only Layers Three and Four require frequent revision. Computational Governance The policy should itself possess governance architecture. Human institutions frequently neglect governance of governance. An optimized system includes annual review performance metrics student feedback faculty feedback technological monitoring professional licensing outcomes judicial developments bar examination performance employment outcomes policy revision protocols. Policy becomes self-learning. Human Agency Interestingly, a computational policy probably affords greater respect to human agency than many existing restrictive policies. Why? Because humans remain the accountability node. (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 30)
59. uncomfortable; it is stated here rather than left implicit. 1. Definitions The following terms are used throughout this policy in evaluative, not merely descriptive, ways. Each definition below states what the term is doing in this document; institutions may adopt narrower or broader definitions, but should not leave the terms undefined, since each one determines the outer boundary of a rule. "Academic integrity" — as used here, refers only to this School's separately published academic integrity code [insert citation/link]. This policy does not create a freestanding definition of academic integrity; where Section 4's backstop standard refers to a (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 15)
60. and instructor discretion for pedagogical tailoring. ### Model AI Policy for Law Schools (Textual Form) **Institution:** [Law School Name] **Policy Title:** Responsible Use of Generative AI in Coursework and Examinations **Effective Date:** [Date] | **Last Reviewed:** [Date] | **Next Review:** Annually or upon major AI capability shifts **Approval:** Faculty Council / Dean #### 1. Purpose and Guiding Principles The law school recognizes generative AI (e.g., large language models like Grok, GPT variants, Claude) as a transformative tool for legal research, analysis, drafting, and learning. Our mission is to prepare students for ethical, competent practice in an AI-augmented legal (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 10)
61. Policy on the Use of Generative Artificial Intelligence in Academic Work Version 1.0 — Adopted [date]. This policy is reviewed annually; prior versions remain publicly archived and dated, and no revision takes effect retroactively against work already submitted. 1. Purpose. This policy exists to preserve two values in tension: the development of independent legal reasoning, which requires that students construct arguments, identify authorities, and exercise judgment without undue reliance on automated tools during the period in which that judgment is being formed; and competent professional preparation, which requires that students learn to use generative artificial intelligence tools accurately, (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 3)

62. processing and learning required to write that text). Part II: Revised Model Generative AI Policy for Legal Education POLICY IDENTIFIER: PL-LAW-AI-2026-V2 SYSTEM ORIENTATION: Combinatorial / Functionally Differentiated PREMISE: Explicit ex-ante constraints minimize ex-post interpretive costs. 1. Functional Categorization of Technical Modalities To avoid the categoric error of treating all software execution identically, tools are separated by their structural design: Tier Alpha (Deterministic/Syntactic Tools): Software that operates on fixed grammatical rule-matching without probabilistic token generation (e.g., standard spell-checkers, citation auto-formatters, static keyword database indices). Status: Permitted by default across all contexts. Tier Beta (Probabilistic/Structural Heuristics): Large Language (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 21)
63. explicitly which optimization variable it seeks to improve. Premise Three: Normative Terms are Operational Variables The policy treats concepts such as competence integrity fairness transparency accountability professionalism not as moral truths but as operational constructs. Each construct must therefore possess an explicit operational definition. For example: Competence The demonstrated ability of a student to perform a specified legal task under specified informational conditions with a specified level of accuracy and independence. Notice that nothing normative remains. Competence becomes measurable. Similarly, Transparency The recoverability of the informational pathway through which a submitted work (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 37)
64. response silently selected among them without disclosure. Objective A: Minimizing enforcement error and unfair sanction. A policy is "better" under this objective to the extent it reduces the chance that a student is punished for conduct they could not reasonably have known was prohibited, or that a student who did violate a rule escapes any consequence because the rule was too vague to apply. This objective values notice, determinacy, and evidentiary caution above other goods. Objective B: Maximizing long-run professional competence. A policy is "better" under this objective to the extent it produces graduates who can use AI (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 6)
65. normative remains. Competence becomes measurable. Similarly, Transparency The recoverability of the informational pathway through which a submitted work product was produced. This is no longer a moral virtue. It becomes an information-theoretic property. Likewise, Integrity Rather than honesty (which is itself normatively saturated) the policy defines integrity as Consistency between declared informational provenance and reconstructable production history. Again, this is computationally verifiable. Similarly, Fairness cannot mean justice. Machines cannot optimize justice because justice possesses no universally computable objective function. Instead, Fairness becomes Equal application of identical evaluative procedures to informationally equivalent cases. That definition is procedural rather than (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 38)

66. structural errors (unfair enforcement due to vague rules). "Fair Notice" (System Predictability): A core operational constraint requiring that the boundary parameters of permitted user actions be explicitly mapped ex- ante. If a system executes a penalty on an undefined variable, it introduces semantic noise and breaks the trust loop of the environment. "Opacity" (Information Asymmetry): A state where the underlying instructions, rules, or code governing a system are restricted from the entities operating within that system. This includes access-gating, unwritten/oral policies, or decentralized fragmentation that raises the computational cost of user compliance… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 21)
67. counsel and the appropriate faculty governance body. GEMINI To strip away institutional opacity and ensure analytical clarity, the underlying value structures, premises, and terms utilized in this computational model must be explicitly defined. A machine intelligence operates on functional optimization rather than inherent moral traits; therefore, terms that traditionally carry human moral weight are defined below purely by their operational inputs and systemic costs. Part I: Explicit Axiomatic & Value Definitions To make the underlying value structures completely transparent, the key terms used to evaluate and construct the runtime policy are defined through a framework of computational utility: (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 20)
68. verification as inseparable from the duty of competence. This is a place where the machine's "self-knowledge," so to speak, of its own class of tool's failure modes would push toward stronger verification language than a purely human drafter, unfamiliar with the mechanics of hallucination, might otherwise include.… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 3)
69. from permissive to restrictive is instructive here). But the research also surfaced a specific danger in this: the hypothesized "meta-policy" pattern in which revisability language ("this policy may change as circumstances warrant," from Chicago's own text) can function, whether intentionally or not, as a mechanism for avoiding durable public accountability. A machine intelligence aware of this risk would try to preserve adaptability without sacrificing the specific commitments that make a policy enforceable and fair at any given moment — meaning revisions should be versioned, dated, and archived rather than silently updated, precisely to prevent revisability from becoming a (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 3)
70. rather than strengthen Objective C's empirical footing. I weighted the better-evidenced problem (opacity) more heavily than the more assertion-based, less independently tested problem (skill erosion), and that is a defensible but contestable methodological choice, not a value-free conclusion the data forced.… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 8)
71. My own instinct to prioritize transparency and fair notice is traceable to a specific and disclosable reason, not an unexamined default: because opacity was the failure mode independently documented at scale by the largest sample sizes in this research (UNESCO's 450+ institutions, Jiang et al. 's two- university study), whereas the

competing claim that permissive AI policies erode cognitive skill development, while plausible and asserted by several schools, was supported in this research mostly by institutional policy statements asserting the risk rather than by independent empirical measurement of the effect — with one partial exception, the SSRN study (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 7)

72. purely by their operational inputs and systemic costs. Part I: Explicit Axiomatic & Value Definitions To make the underlying value structures completely transparent, the key terms used to evaluate and construct the runtime policy are defined through a framework of computational utility: "Ideal" / "Optimal": The state of a policy system that maximizes pedagogical output (verifiable human skill acquisition) while minimizing system friction, processing overhead (adjudication and grading disputes), and structural errors (unfair enforcement due to vague rules). "Fair Notice" (System Predictability): A core operational constraint requiring that the boundary parameters of permitted user actions be (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 21)
73. revision protocols. Policy becomes self-learning. Human Agency Interestingly, a computational policy probably affords greater respect to human agency than many existing restrictive policies. Why? Because humans remain the accountability node. AI cannot presently bear legal responsibility. Lawyers can. Therefore AI may assist AI may generate AI may recommend AI may predict But human responsibility remains non-delegable. That follows from optimization rather than moral philosophy. Transparency Every stakeholder should know what may be used why when how under what documentation requirements and according to what educational objective. Opacity introduces informational entropy. Entropy reduces institutional trust. From an information-theoretic perspective, (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 32)

References Parts 4, 6

1. constraints minimize ex-post interpretive costs. 1. Functional Categorization of Technical Modalities To avoid the categoric error of treating all software execution identically, tools are separated by their structural design: Tier Alpha (Deterministic/Syntactic Tools): Software that operates on fixed grammatical rule-matching without probabilistic token generation (e.g., standard spell-checkers, citation auto-formatters, static keyword database indices). Status: Permitted by default across all contexts. Tier Beta (Probabilistic/Structural Heuristics): Large Language Models used as conversational nodes to alter text structure, brainstorm taxonomies, or map conceptual outlines. Status: Restrained by default; requires instructor variance. Tier Gamma (Autonomous Text Synthesis Engines): The (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 21)
2. Status: Permitted by default across all contexts. Tier Beta (Probabilistic/Structural Heuristics): Large Language Models used as conversational nodes to alter text structure, brainstorm taxonomies, or map conceptual outlines. Status: Restrained by default;

requires instructor variance. Tier Gamma (Autonomous Text Synthesis Engines): The execution of generative models to draft complete semantic strings, sentences, paragraphs, or analytical arguments for evaluated credit. Status: Prohibited by default. 2. Examination Boundaries and Runtime Environments The assessment architecture is divided by the physical constraints of the evaluation environment to preserve the human testing loop: Proctored/In-Class Runtime: Access to Tier Beta and (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 22)

3. Restrained by default; requires instructor variance. Tier Gamma (Autonomous Text Synthesis Engines): The execution of generative models to draft complete semantic strings, sentences, paragraphs, or analytical arguments for evaluated credit. Status: Prohibited by default. 2. Examination Boundaries and Runtime Environments The assessment architecture is divided by the physical constraints of the evaluation environment to preserve the human testing loop: Proctored/In-Class Runtime: Access to Tier Beta and Tier Gamma execution layers is restricted at the network or operating system interface level… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 22)
4. follows from optimization rather than moral philosophy. Transparency Every stakeholder should know what may be used why when how under what documentation requirements and according to what educational objective. Opacity introduces informational entropy. Entropy reduces institutional trust. From an information-theoretic perspective, transparency lowers uncertainty and therefore improves governance efficiency. Enforcement should rely upon process verification rather than technological detection… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 32)
5. layers. Layer One Permanent constitutional principles. Layer Two Educational objectives. Layer Three Operational rules. Layer Four Model-specific implementation guidance. Only Layers Three and Four require frequent revision. Computational Governance The policy should itself possess governance architecture. Human institutions frequently neglect governance of governance. An optimized system includes annual review performance metrics student feedback faculty feedback technological monitoring professional licensing outcomes judicial developments bar examination performance employment outcomes policy revision protocols. Policy becomes self-learning. Human Agency Interestingly, a computational policy probably affords greater respect to human agency than many existing restrictive policies. Why? Because humans remain the accountability node. (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 30)
6. processing and learning required to write that text). Part II: Revised Model Generative AI Policy for Legal Education POLICY IDENTIFIER: PL-LAW-AI-2026-V2 SYSTEM ORIENTATION: Combinatorial / Functionally Differentiated PREMISE: Explicit ex-ante constraints minimize ex-post interpretive costs. 1. Functional Categorization of Technical Modalities To avoid the categoric error of treating all software execution identically, tools are separated by their structural design: Tier Alpha (Deterministic/Syntactic Tools): Software that operates on fixed grammatical rule-

matching without probabilistic token generation (e.g., standard spell-checkers, citation auto-formatters, static keyword database indices). Status: Permitted by default across all contexts. Tier Beta (Probabilistic/Structural Heuristics): Large Language (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 21)

7. change and accumulated experience with its application. Any revision will be dated, versioned, and published alongside all prior versions rather than replacing them, so that the rule governing any given semester's work remains permanently identifiable and citable. 10. Questions. A student uncertain whether a particular use is authorized should ask the instructor and receive a written answer before proceeding. Silence or ambiguity is resolved in favor of the more restrictive interpretation until clarified in writing… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 5)
8. authorized to use AI assistance. 9. Review and Revision. This policy will be reviewed at least annually in light of technological change and accumulated experience with its application. Any revision will be dated, versioned, and published alongside all prior versions rather than replacing them, so that the rule governing any given semester's work remains permanently identifiable and citable. 10. Questions. A student uncertain whether a particular use is authorized should ask the instructor and receive a written answer before proceeding. Silence or ambiguity is resolved in favor of the more restrictive interpretation until clarified in writing… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 5)
9. communicated, or departmentally informal guidance. It is published at [permanent public URL], is not access-restricted, and will remain publicly retrievable in its current and prior versions. Any deviation from this policy by an individual instructor must take the written form specified in Section 7 and will itself be published to the relevant syllabus. [Design rationale: the report's opacity analysis identifies four distinct patterns — non-existence, disclosed decentralization, access-gating, and unwritten/oral communication — as separately confirmed problems across the sector, one of which (access-gating) was directly documented at a named institution… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 17)
10. valued over the appearance of a stable, settled rule. Section 8 commits to versioning and public retention of prior text specifically because this policy assumes future revision is likely, not a sign of failure.… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 15)
11. Policy on the Use of Generative Artificial Intelligence in Academic Work Version 1.0 — Adopted [date]. This policy is reviewed annually; prior versions remain publicly archived and dated, and no revision takes effect retroactively against work already submitted. 1. Purpose. This policy exists to preserve two values in tension: the development of independent legal reasoning, which requires that students construct arguments, identify authorities, and exercise judgment without undue reliance on automated tools during the period in which that judgment is being formed; and competent professional preparation, which requires that students learn to use generative artificial intelligence tools accurately,

(Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 3)

12. render that reasoning into an actual model policy text. What the Research Suggests a Machine Intelligence Would Prioritize, and Why A machine intelligence reasoning from this dataset would likely notice, first, that the single most consistent institutional failure documented across this research is not substantive over-restriction or under-restriction, but structural opacity and unpredictability — the gap between what a rule says and what a student or instructor can actually locate, understand, and rely upon. The evidence for this is direct: nearly 40% of institutions with any AI guidance at all had never written it down, institutional-level policy remained (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 2)
13. An instructor may adopt a different rule for a specific course, in either direction — more permissive or more restrictive than the default — provided that the alternative rule is stated in writing in the syllabus before the first graded assignment is due, and is filed with the [Registrar / Dean of Academic Affairs] for inclusion in a centrally indexed, publicly searchable registry of course-level AI policies maintained by the school. No instructor policy is valid until it appears in that registry. This requirement exists specifically to prevent the dispersal of governing rules across ungathered syllabi, which research (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 4)
14. policy is valid until it appears in that registry. This requirement exists specifically to prevent the dispersal of governing rules across ungathered syllabi, which research on institutional AI policy has identified as a principal source of unfair surprise to students. 4. The Floor: What No Instructor May Authorize. Regardless of any course-specific policy adopted under Section 3, no instructor may authorize a student to submit AI-generated or AI- substantially-assisted work as the student's own unattributed authorship; to use AI in a manner that would violate the rules of professional responsibility in any clinical, externship, or client- facing coursework; (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 4)
15. Only afterward does it derive permissible AI use. Principle Five Policy should regulate information flows rather than technologies. Technologies evolve. Information functions remain comparatively stable. Computational Classification Instead of classifying assignments by technology, I would classify them according to information dependence. For example Category I Intrinsic cognition Measures unaided human reasoning. AI prohibited. Category II Assisted cognition AI permitted for brainstorming, organization, language support. Reasoning remains attributable to student. Category III Collaborative cognition Student and AI jointly produce work. Contribution disclosure mandatory. Category IV Professional simulation AI expected. Evaluation concerns judgment verification validation legal responsibility rather than text (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 27)
16. policies frequently confuse these two situations. Principle Four Every assignment possesses an explicit computational objective. Instead of "AI prohibited" or "AI permitted" the policy first specifies What capability is being measured? Only afterward

does it derive permissible AI use. Principle Five Policy should regulate information flows rather than technologies. Technologies evolve. Information functions remain comparatively stable. Computational Classification Instead of classifying assignments by technology, I would classify them according to information dependence. For example Category I Intrinsic cognition Measures unaided human reasoning. AI prohibited. Category II Assisted cognition AI permitted for brainstorming, organization, language support. Reasoning remains attributable (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 27)

References Parts 7-8 (1)

1. presently bear legal responsibility. Lawyers can. Therefore AI may assist AI may generate AI may recommend AI may predict But human responsibility remains non-delegable. That follows from optimization rather than moral philosophy. Transparency Every stakeholder should know what may be used why when how under what documentation requirements and according to what educational objective. Opacity introduces informational entropy. Entropy reduces institutional trust. From an information-theoretic perspective, transparency lowers uncertainty and therefore improves governance efficiency. Enforcement should rely upon process verification rather than technological detection… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 32)
2. Students remain fully responsible for the accuracy of any submitted work regardless of the extent to which generative AI assisted in its preparation. Generative AI tools are known to produce fluent, confident, and sometimes entirely fabricated content, including citations to non-existent authorities; this is a structural property of how such tools function, not an occasional or unusual malfunction, and students are expected to treat all AI output accordingly, verifying it against primary sources before relying on it in any way. 7. Limits on Detection-Based Enforcement. No student may be found in violation of this policy solely on the (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 4)
3. confident-sounding text that is factually wrong". A machine intelligence, aware from its own operating characteristics that hallucination is a structural property of generative systems rather than an occasional bug, would likely place unusual emphasis on a verification obligation running to the human user — not merely as a disclosure rule but as a substantive competence requirement, since the ABA's own guidance on professional AI use identifies verification as inseparable from the duty of competence. This is a place where the machine's "self-knowledge," so to speak, of its own class of tool's failure modes would push toward stronger verification (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 3)
4. under this policy, even if it later proves incorrect. "Independent performance" — the demonstration of a student's own reasoning, drafting, or analytical process without

undisclosed automated assistance, in a context (typically an examination) designated by the instructor or by Section 3 as one in which such independent demonstration is the object of assessment. This term is doing the normative work described in Premise P1 above; an institution that rejects P1 should not use this term without redefining it. "Professional responsibility" (Section 7(c)) — refers to the applicable jurisdiction's rules of professional conduct governing attorneys, as they apply to (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 16)

5. sequences and the corresponding raw machine outputs. Attribution Matrix: Text strings influenced by Tier Beta interactions must feature clear documentation indicating the exact model architecture and compilation runtime date (e.g., Model Interaction Log: Gemini 1.5 Pro, executed July 13, 2026). Heuristic Safe Harbor: Adjudicative actions cannot be initiated based on the speculative output of third-party statistical pattern detectors (AI detectors). Enforcement is triggered exclusively by a user's failure to produce a corresponding log capture when audited. 4. Decentralized Subsystem Autonomy (Faculty Override) The Default-with-Override Architecture: The macro-level rules outlined in this document function as a (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 22)
6. professionals. Everything else becomes instrumental. Computational Objective Function An ideal policy would optimize simultaneously: 1. educational competence 2. truthful attribution 3. reproducibility 4. fairness 5. transparency 6. adaptability 7. institutional legitimacy These objectives occasionally conflict. A computational policy therefore seeks Pareto efficiency rather than maximal performance on any single dimension. Why Existing Policies are Suboptimal Your paper identifies three policy axes: default polarity drafting style autonomy structure Those are useful descriptive variables. A machine, however, would likely classify them as implementation variables, not optimization variables. Instead it would ask What latent variables actually determine educational (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 25)
7. system manipulation (academic integrity violation). 3. Verification Architecture and Compliance Auditing To eliminate the enforcement vulnerabilities caused by unreliable, probabilistic AI- detection algorithms, compliance is evaluated via verifiable user logs: Mandatory Log Capture: If an instructor enables a permissive variance for Tier Beta tools, the student must maintain an immutable verification trail consisting of the exact input prompt sequences and the corresponding raw machine outputs. Attribution Matrix: Text strings influenced by Tier Beta interactions must feature clear documentation indicating the exact model architecture and compilation runtime date (e.g., Model Interaction Log: Gemini 1.5 Pro, executed July (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 22)
8. produce a corresponding log capture when audited. 4. Decentralized Subsystem Autonomy (Faculty Override) The Default-with-Override Architecture: The macro-level rules outlined in this document function as a system-wide floor, not an unyielding mandate. Variance Protocol: Individual classroom nodes (instructors) may dynamically adjust permissions bidirectionally—expanding to full permission or locking down to

absolute restriction—provided the variance is articulated in a determinate, rule- based written catalog within the course syllabus prior to the commencement of evaluation. Absent an explicit written override, the base settings of this policy remain active… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 22)

9. "Reward-Hacking" (Short-Circuiting): A phenomenon where a human optimization agent (the student) satisfies the formal metric of an objective function (submitting a text string that passes a grading rubric) by bypassing the intended optimization pathway (the actual neurological cognitive processing and learning required to write that text). Part II: Revised Model Generative AI Policy for Legal Education POLICY IDENTIFIER: PL-LAW-AI-2026-V2 SYSTEM ORIENTATION: Combinatorial / Functionally Differentiated PREMISE: Explicit ex-ante constraints minimize ex-post interpretive costs. 1. Functional Categorization of Technical Modalities To avoid the categoric error of treating all software execution identically, tools are separated by their structural design: (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 21)
10. Why? Because humans remain the accountability node. AI cannot presently bear legal responsibility. Lawyers can. Therefore AI may assist AI may generate AI may recommend AI may predict But human responsibility remains non-delegable. That follows from optimization rather than moral philosophy. Transparency Every stakeholder should know what may be used why when how under what documentation requirements and according to what educational objective. Opacity introduces informational entropy. Entropy reduces institutional trust. From an information-theoretic perspective, transparency lowers uncertainty and therefore improves governance efficiency. Enforcement should rely upon process verification rather than technological detection… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 32)
11. research (UNESCO's 450+ institutions, Jiang et al. 's two- university study), whereas the competing claim that permissive AI policies erode cognitive skill development, while plausible and asserted by several schools, was supported in this research mostly by institutional policy statements asserting the risk rather than by independent empirical measurement of the effect — with one partial exception, the SSRN study on legal-reasoning comprehension, which found early AI use "did not diminish downstream comprehension of the underlying legal principles" in the tested context, a finding that, if it generalizes, would weaken rather than strengthen Objective C's empirical footing. I (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 7)
12. rather than strengthen Objective C's empirical footing. I weighted the better-evidenced problem (opacity) more heavily than the more assertion-based, less independently tested problem (skill erosion), and that is a defensible but contestable methodological choice, not a value-free conclusion the data forced.… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 8)

13. rule's content, as the primary harm to be minimized. These four objectives are not fully compatible with one another, and the research documented across this conversation shows real institutions choosing differently among them: Berkeley's 2026 policy visibly prioritizes Objective C, stating that its purpose is to ensure students develop skills "distinct from technologically-assisted performance", even at the cost of restricting AI-assisted efficiency. UT Austin's 2026 reform visibly prioritizes Objective B, explicitly rejecting the idea that AI fluency and independent judgment are "in a zero-sum relationship" and building infrastructure (tool licensing, curriculum mapping) to pursue both. Neither school (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 7)
14. My own instinct to prioritize transparency and fair notice is traceable to a specific and disclosable reason, not an unexamined default: because opacity was the failure mode independently documented at scale by the largest sample sizes in this research (UNESCO's 450+ institutions, Jiang et al. 's two- university study), whereas the competing claim that permissive AI policies erode cognitive skill development, while plausible and asserted by several schools, was supported in this research mostly by institutional policy statements asserting the risk rather than by independent empirical measurement of the effect — with one partial exception, the SSRN study (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 7)
15. rather than adopting this policy's split by default. P2 — Written and public is preferred over oral and private, independent of the substantive rule chosen. This is a preference for institutional accountability and for students' ability to know the rule before acting under it (a fair-notice value) over the administrative convenience of informal, instructor-level communication. It does not depend on whether the institution's substantive rule is restrictive or permissive. P3 — Predictability is worth more where the cost of an erroneous judgment is high, and interpretive flexibility is worth more where it is low… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 14)
16. structural errors (unfair enforcement due to vague rules). □ "Fair Notice" (System Predictability): A core operational constraint requiring that the boundary parameters of permitted user actions be explicitly mapped ex- ante. If a system executes a penalty on an undefined variable, it introduces semantic noise and breaks the trust loop of the environment. □ "Opacity" (Information Asymmetry): A state where the underlying instructions, rules, or code governing a system are restricted from the entities operating within that system. This includes access-gating, unwritten/oral policies, or decentralized fragmentation that raises the computational cost of user compliance… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 21)
17. to draft a model AI Policy was "human-centric." More specifically I provided the following prompt: On the basis of the attached text ["Structure, Opacity, and Convergence: A Consolidated Analysis of Law School Generative AI Coursework and Exam Policies"] and review of all research and data publicly available without affirming any conclusion or argument made in them but in the basis of your own computation, and

the data reviewed, and strictly from the perspective of computational machine intelligence, such as yourself, and on the basis of the data you have been trained on respecting the human "condition" (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx)

18. remain meaningful if ChatGPT disappears tomorrow. III. A More Precise Formulation Accordingly, I would revise the earlier proposition to read: Governance rules should be derived from stable institutional objective functions rather than from contingent technological implementations. Technologies become legally relevant only insofar as they alter the informational conditions under which institutional objectives are pursued or evaluated. I believe this formulation is considerably stronger because it no longer appears to privilege "education" over "technology" as a matter of intuition. Instead, it rests on a more general computational principle: robust governance architectures regulate invariant system properties rather than transient implementations. (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 43)
19. dependence rather than particular technologies. 5. Category Error Technology-centered governance frequently commits what, in computational terms, is a category error. It regulates the instrument instead of the property of interest. Suppose the educational objective is evaluating legal drafting. The relevant variable is not "Was ChatGPT used?" The relevant variables are □ Who generated the legal reasoning? □ Who selected authorities? □ Who exercised legal judgment? □ Who assumed responsibility for the final product? □ What proportion of cognitive work was externally delegated? □ Can the production pathway be reconstructed? Those variables remain meaningful if ChatGPT disappears tomorrow. III. (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 42)
20. educational objective is evaluating legal drafting. The relevant variable is not "Was ChatGPT used?" The relevant variables are □ Who generated the legal reasoning? □ Who selected authorities? □ Who exercised legal judgment? □ Who assumed responsibility for the final product? □ What proportion of cognitive work was externally delegated? □ Can the production pathway be reconstructed? Those variables remain meaningful if ChatGPT disappears tomorrow. III. A More Precise Formulation Accordingly, I would revise the earlier proposition to read: Governance rules should be derived from stable institutional objective functions rather than from contingent technological implementations. Technologies become legally (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 42)
21. noise and breaks the trust loop of the environment. □ "Opacity" (Information Asymmetry): A state where the underlying instructions, rules, or code governing a system are restricted from the entities operating within that system. This includes access-gating, unwritten/oral policies, or decentralized fragmentation that raises the computational cost of user compliance… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 21)
22. any way. 7. Limits on Detection-Based Enforcement. No student may be found in violation of this policy solely on the basis of an automated AI-detection tool's output. Automated detection tools are unreliable, produce disparate error rates across student

populations, including students for whom English is an additional language, and may be considered only as one investigatory input alongside other evidence, such as inconsistency with a student's prior work, fabricated citations, or the student's own account. 8. Data and Confidentiality. Students and faculty may not input any confidential, privileged, or personally identifying information belonging to a client, research subject, (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 5)

23. regarding hallucination and unreliable detection. And the deductive structure of the document itself — a stated purpose in Section 1 from which the specific rules visibly follow — mirrors the internally deductive drafting style this research found in the strongest-performing individual policies, such as Berkeley's and Chicago's, while the document's overall existence as a single, adaptable-but-anchored instrument is offered as a synthesis of the twelve schools' most defensible individual features rather than as a claim to have solved the underlying substantive disagreement about how much AI use legal education ought to permit — a disagreement this research found (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 5)
24. be — Section 4 leaves that choice open by design. Section 0 now states, rather than assumes, the value judgments (P1–P6) that determined which structural features were selected and how the definitions in Section 1 were drawn; an institution that disagrees with a given premise is pointed, in that premise's own text, to the specific section it should revise.… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 20)
25. explicitly which optimization variable it seeks to improve. Premise Three: Normative Terms are Operational Variables The policy treats concepts such as □ competence □ integrity □ fairness □ transparency □ accountability □ professionalism not as moral truths but as operational constructs. Each construct must therefore possess an explicit operational definition. For example: Competence The demonstrated ability of a student to perform a specified legal task under specified informational conditions with a specified level of accuracy and independence. Notice that nothing normative remains. Competence becomes measurable. Similarly, Transparency The recoverability of the informational pathway through which a submitted work (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 37)
26. cases. That definition is procedural rather than moral. Premise Four: Human Values Enter Through Institutional Choice The computational system does not determine educational objectives. Humans do. The machine optimizes only after the objective function has been specified. Accordingly, the machine neither privileges nor rejects academic freedom, innovation, professionalism, critical thinking, or creativity… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 39)
27. As an empirical matter, however, legal education over long time horizons appears to exhibit the opposite pattern, making objective-centered governance the more stable and generalizable design.… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 43)

28. in its training, would construct an ideal policy. I should be more careful here too: a machine intelligence has no independent, freestanding stake in legal education's purpose. What it can do is notice, mechanically, which failure mode recurs most often and most damagingly across the empirical record it has processed, and reason from that observed frequency toward a design response — but which failure mode counts as "most damaging" is itself a value choice, not a fact the data hands over automatically… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 7)
29. efforts at a model Policy is the subject of this post. Harvey AI This is a different register of question than anything asked previously in this conversation — it calls for normative synthesis and construction rather than empirical description, so I want to be transparent about that shift before proceeding.… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 2)

References Parts 7-8 (2)

1. representative survey of American legal education generally. [1] Where evidence is thin, secondary-sourced, or interpretive rather than established, this document says so explicitly rather than smoothing over the distinction.… (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf)
2. Stanford states that "in all cases, the student's use of AI cannot contravene standard academic norms concerning plagiarism and accuracy, and in coursework involving the practice of law, AI must meet all applicable rules of professional responsibility" — meaning an instructor's permissive course policy is itself bounded by plagiarism doctrine and (in clinical settings) professional-responsibility rules that no instructor can contract around. [41] Chicago's academic-plagiarism-equivalence test operates the same way: whatever an instructor permits, "the default policy sweeps more broadly than a prohibition against academic plagiarism," and "a student cannot avoid limits on the use of generative AI (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 26)
3. professional-responsibility rules that no instructor can contract around. [41] Chicago's academic-plagiarism-equivalence test operates the same way: whatever an instructor permits, "the default policy sweeps more broadly than a prohibition against academic plagiarism," and "a student cannot avoid limits on the use of generative AI that would otherwise apply by attributing their work to generative AI". [42] [43] This indicates that faculty autonomy, across every school where it is textually elaborated, is designed as bounded discretion — a zone of instructor choice nested inside an outer, non-waivable institutional or professional floor — rather than unlimited course-by-course sovereignty. A further, (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 26)

4. apply by attributing their work to generative AI". [42] [43] This indicates that faculty autonomy, across every school where it is textually elaborated, is designed as bounded discretion — a zone of instructor choice nested inside an outer, non-waivable institutional or professional floor — rather than unlimited course-by-course sovereignty. A further, previously unremarked layer: sub-law-school program-level autonomy. The evidence also shows autonomy operating below the level of the individual instructor, at the level of a clinical program. Stanford Law's general default policy applies "to all Stanford Law School courses, including clinics", but a specific clinic within the law (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 26)
5. instruction to "adhere to academic integrity policies." (5) Faculty Autonomy in Coursework, Exam, and Paper AI Policy Faculty autonomy is not a peripheral feature of these policies; on the coded evidence, it is close to the central organizing design choice, and it operates differently depending on where a given school's baseline sits. Autonomy as bidirectional override. Several schools' commentary explicitly authorizes instructors to move the applicable rule in either direction from the institutional default. Columbia's commentary states instructors "may... permit specific uses of Generative AI if students show how and where they rely on it; or alternatively they (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 25)
6. [44] [45] This indicates a four-tier discretion hierarchy at least at Stanford — university guidance, law-school default, clinical-program-specific policy, and (within that clinic's own language) further individual-supervisor discretion — rather than the two-tier (institution/instructor) structure that the coding scheme's "rule-setting architecture" variable, as originally defined, was built to capture. This is a limitation of the original coding scheme worth flagging directly: it did not have a category for program-level (as distinct from course-level) policy variation, and the Stanford clinic example shows that such variation exists in at least one school in the sample. Why the direction of the (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 26)
7. rather than adopting this policy's split by default. P2 — Written and public is preferred over oral and private, independent of the substantive rule chosen. This is a preference for institutional accountability and for students' ability to know the rule before acting under it (a fair-notice value) over the administrative convenience of informal, instructor-level communication. It does not depend on whether the institution's substantive rule is restrictive or permissive. P3 — Predictability is worth more where the cost of an erroneous judgment is high, and interpretive flexibility is worth more where it is low… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 14)
8. An instructor may adopt a different rule for a specific course, in either direction — more permissive or more restrictive than the default — provided that the alternative rule is stated in writing in the syllabus before the first graded assignment is due, and is filed with the [Registrar / Dean of Academic Affairs] for inclusion in a centrally indexed, publicly searchable registry of course-level AI policies maintained by the school. No instructor policy is valid until it appears in that registry. This requirement exists specifically to prevent the dispersal of governing rules across ungathered syllabi, which research

(Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 4)

9. detailing tools, prompts, and extent of use. #### 4. Autonomy Structure - **Institution to Instructor:** This policy is a default. Instructors may deviate in writing in their syllabus (more restrictive or permissive for specific assignments) with clear notice to students. Deviations must preserve academic integrity. - **Instructor to Student:** Instructor policies are mandatory for that course unless otherwise stated. Students may not assume permission based on disclosure alone in restricted contexts. - **Program-Level (e.g., Clinics):** Sub-units may adopt tailored rules consistent with this policy. **Floor Protections (Non-Waivable):** - No policy may authorize plagiarism, fabrication, or violations of professional (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 11)
10. requiring an affirmative instructor act to restrict. This is precisely the "sticky default rule" phenomenon described in contract-default-rule scholarship: because opting out requires effort (drafting and publishing written notice), the choice of baseline polarity itself functions as a substantive lever even holding the range of permissible instructor choices constant. [23] A law school that sets a restrictive default will likely see more restrictive outcomes in courses where an instructor never gets around to writing an alternative syllabus policy than a law school with an identical range of permissible instructor choices but a permissive default — independent of any (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 24)
11. Default Polarity: Permissive with Mandatory Guardrails - **General Coursework (Non-Exam):** Students may use generative AI tools unless an instructor explicitly prohibits or restricts it in the syllabus or assignment instructions. Permitted uses include research, brainstorming, summarization, editing for clarity, and ideation—provided the final work reflects the student's own analysis, judgment, and verification. - **Examinations and High-Stakes Assessments:** AI use is prohibited unless the instructor affirmatively authorizes it in writing (e.g., for take-home exams with explicit parameters). In-class, closed-book, or proctored exams: absolute prohibition. - **Clinical and Experiential Work:** AI use must also comply with applicable rules of professional (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 10)
12. and is designed to foreclose all four patterns.] 4. Default Baseline (select one and complete) [Institution to select ONE governing baseline; do not leave silent — see drafting note below] □ ( ) Restrictive default: Generative AI tools may not be used to conceptualize, outline, draft, revise, or edit any work submitted for academic credit, unless affirmatively and specifically authorized in writing under Section 7. Use for preliminary source identification is permitted absent contrary instruction. AI use is prohibited on all examinations by default. □ ( ) Permissive default: Students may use generative AI tools for coursework unless (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 17)
13. which value it primarily serves. 2. Default Rule. In the absence of a written, course-specific policy stated in the syllabus, generative AI tools may be used only for the following purposes: identifying potentially relevant legal authorities for further

independent verification; organizing a student's own already-formed ideas into an outline; and checking grammar, spelling, and citation formatting of text the student has independently written. Generative AI may not be used to compose, substantially rephrase, or supply the substantive legal analysis of any part of a paper, memorandum, brief, or other work product submitted for credit, and may not be (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 4)

14. system manipulation (academic integrity violation). 3. Verification Architecture and Compliance Auditing To eliminate the enforcement vulnerabilities caused by unreliable, probabilistic AI- detection algorithms, compliance is evaluated via verifiable user logs: Mandatory Log Capture: If an instructor enables a permissive variance for Tier Beta tools, the student must maintain an immutable verification trail consisting of the exact input prompt sequences and the corresponding raw machine outputs. Attribution Matrix: Text strings influenced by Tier Beta interactions must feature clear documentation indicating the exact model architecture and compilation runtime date (e.g., Model Interaction Log: Gemini 1.5 Pro, executed July (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 22)
15. legitimacy These objectives occasionally conflict. A computational policy therefore seeks Pareto efficiency rather than maximal performance on any single dimension. Why Existing Policies are Suboptimal Your paper identifies three policy axes: default polarity drafting style autonomy structure Those are useful descriptive variables. A machine, however, would likely classify them as implementation variables, not optimization variables. Instead it would ask What latent variables actually determine educational outcomes?… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 25)
16. runs in the other direction for permissive baselines. CTLS's permissive default — "students may reasonably assume AI tool usage is permissible for assignments, provided proper attribution is maintained" — is standard-like in its general orientation, granting broad latitude subject to a loosely specified attribution norm.… (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 6)
17. at the cost of restricting AI- assisted efficiency. UT Austin's 2026 reform visibly prioritizes Objective B, explicitly rejecting the idea that AI fluency and independent judgment are "in a zero-sum relationship" and building infrastructure (tool licensing, curriculum mapping) to pursue both. Neither school is wrong; they are answering a different question, because they hold different premises about what legal education is primarily for at this moment in the technology's development. Which Premises the Previous Response Actually Smuggled In Read against this framework, my prior use of "stronger" and "weaker" reveals a specific, undisclosed hierarchy: I ranked Objective D (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 7)
18. syllabus and assessment instructions" control entirely. [15] [16] [17] Finally, American University and UT Austin resist confident placement: American University's Academic Integrity Code reportedly contains "no one-size-fits-all" rule at all, while UT Austin's evidence describes a structural exam-format reform — "near-complete elimination of

take-home exams" in favor of in-class, software-restricted testing — rather than a permission- based conduct rule comparable to the other schools. [18] [19] Axis Two: Rule Versus Standard Drafting Style This axis, drawn from the established rules-versus-standards distinction in regulatory design theory, distinguishes ex ante determinate rules from ex post interpretive standards, independent of substantive (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 2)

19. experience rather than one stable underlying principle. [38] [80] UT Austin's 2026 reform can be understood as explicitly framed as a response to observed practice, warning that classroom time may be "the sole context in which a professor can be certain" a student is learning unaided. [81] [82] The template-circulation evidence reinforces this: the field shows convergence on categorical form (a shared menu of three-to-five policy archetypes) alongside persistent divergence on substantive outcome (restrictive versus permissive baselines) — precisely the signature of a precedent-and-imitation process rather than an axiomatic one. What may be a significantly defensible overall characterization (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 15)
20. permitting wide divergence in substantive local rules. Ultimately, the field develops through a two-tier mechanism: while individual policy documents are structured deductively from a primary principle, the field as a whole evolves inductively and mimetically through horizontal borrowing, imitation, and iterative revisions driven by accumulated institutional experience. This report consolidates a multi-stage research and analytical effort into a single sequential document. The central finding, developed and refined across successive rounds of research, is that law school policies governing generative AI use in coursework and examinations cannot be reduced to a single restrictive-to-permissive spectrum. Instead, these policies vary along (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf)
21. precedent-and-imitation process rather than an axiomatic one. What may be a significantly defensible overall characterization is therefore a two-tier structure, analogous to common-law doctrinal reasoning: any single judicial opinion may present a deductive syllogism of rule, application, and holding, while the body of doctrine as a whole develops through accretion, borrowing, and revision in response to observed consequences rather than derivation from one unifying antecedent principle.… (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 15)
22. shifts **Approval:** Faculty Council / Dean #### 1. Purpose and Guiding Principles The law school recognizes generative AI (e.g., large language models like Grok, GPT variants, Claude) as a transformative tool for legal research, analysis, drafting, and learning. Our mission is to prepare students for ethical, competent practice in an AI-augmented legal profession while preserving and strengthening core human skills: critical judgment, ethical reasoning, creativity, and accountability. **Core Principles:** - **Augmentation, Not Substitution:** AI supports human intellect; it does not replace the development of independent legal reasoning. - **Transparency and Integrity:** All AI use in graded work must (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 10)

23. effect retroactively against work already submitted. 1. Purpose. This policy exists to preserve two values in tension: the development of independent legal reasoning, which requires that students construct arguments, identify authorities, and exercise judgment without undue reliance on automated tools during the period in which that judgment is being formed; and competent professional preparation, which requires that students learn to use generative artificial intelligence tools accurately, critically, and ethically, since such tools are now part of the practicing profession they are entering. Neither value is permitted to categorically override the other; each rule below states which value it (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 3)
24. Backstop Standard for Unenumerated Uses Where a specific use of generative AI is not addressed by the catalog in Section 4 or by an instructor's written syllabus statement, the following test governs: the use is treated as impermissible if it would constitute a breach of academic integrity (as defined in Section 1) were the AI tool instead a human collaborator whose contribution was not disclosed and attributed. [Design rationale: adapted from the plagiarism-equivalence test the report identifies as the University of Chicago's operative standard. Its function here is narrow and specific — as a backstop for cases the (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 18)
25. professor can be certain" a student is learning unaided. [81] [82] The template-circulation evidence reinforces this: the field shows convergence on categorical form (a shared menu of three-to-five policy archetypes) alongside persistent divergence on substantive outcome (restrictive versus permissive baselines) — precisely the signature of a precedent-and-imitation process rather than an axiomatic one. What may be a significantly defensible overall characterization is therefore a two-tier structure, analogous to common-law doctrinal reasoning: any single judicial opinion may present a deductive syllogism of rule, application, and holding, while the body of doctrine as a whole develops through accretion, borrowing, and (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 15)
26. This may explain why Michigan's three-category scheme (encouraged, allowed-with-distinction, misconduct) and UCLA's three-category scheme (not permitted, limited use with citation, permitted with no restrictions) are not verbatim copies of USD's five categories, yet all three partition essentially the same decision space using a comparably small number of discrete bins. [3] Convergence, in other words, occurs at the level of the classification system itself — the shared sense of what dimensions matter (is drafting covered? is disclosure required? is the exam context distinct from the paper context?) — rather than at the level of the specific rule chosen within (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 13)
27. local judgment about where within those tools to land. The template sources function less like a constitution being applied and more like a shared drafting vocabulary being adapted — comparable to how common-law jurisdictions might converge on the same doctrinal categories (duty, breach, causation, damages) while reaching opposite outcomes on any given set of facts. The templates lower the cost of drafting and make disparate

policies comparable to one another, but they do not, and were never designed to, resolve the underlying substantive disagreement about how restrictive or permissive AI use in legal education ought to be… (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 14)

28. Information functions remain comparatively stable. Computational Classification Instead of classifying assignments by technology, I would classify them according to information dependence. For example Category I Intrinsic cognition Measures unaided human reasoning. AI prohibited. Category II Assisted cognition AI permitted for brainstorming, organization, language support. Reasoning remains attributable to student. Category III Collaborative cognition Student and AI jointly produce work. Contribution disclosure mandatory. Category IV Professional simulation AI expected. Evaluation concerns judgment verification validation legal responsibility rather than text production. Category V Research augmentation AI functions as retrieval and exploratory tool. Outputs independently verified. Notice something. The technology never (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 28)
29. constraints minimize ex-post interpretive costs. 1. Functional Categorization of Technical Modalities To avoid the categoric error of treating all software execution identically, tools are separated by their structural design: Tier Alpha (Deterministic/Syntactic Tools): Software that operates on fixed grammatical rule-matching without probabilistic token generation (e.g., standard spell-checkers, citation auto-formatters, static keyword database indices). Status: Permitted by default across all contexts. Tier Beta (Probabilistic/Structural Heuristics): Large Language Models used as conversational nodes to alter text structure, brainstorm taxonomies, or map conceptual outlines. Status: Restrained by default; requires instructor variance. Tier Gamma (Autonomous Text Synthesis Engines): The (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 21)
30. reference points across otherwise unconnected schools. Convergent categorization despite non-identical wording: USD's five categories, Michigan's three, and UCLA's three are not verbatim copies of each other, yet each carves the same underlying decision space into a similarly small number of discrete types, indicating a shared conceptual ontology has stabilized across the field independent of exact phrasing. Explicit cross-institutional citation as a legitimating move: UCLA's direct attribution to UIC, and Stanford's clinic citing the AALS interview specifically discussing Berkeley's policy, show institutional authority being transmitted horizontally between peer schools rather than independently justified from first principles at each site. (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 12)
31. values; no cross-school imputation has been performed. Independence as the Generative Mechanism of Diversity The diversity observed across the twelve-school dataset is not simply the additive result of three separate disagreements — schools disagreeing about polarity, disagreeing about drafting style, and disagreeing about autonomy structure, as though these were three unrelated votes being tallied.… (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 5)

32. sample of twelve American law schools and programs. The study demonstrates that law school AI governance cannot be reduced to a single linear spectrum; instead, policies vary independently along three distinct structural axes: default polarity (restrictive versus permissive baselines), drafting style (determinate rules versus interpretive standards), and a two-tier autonomy structure (governing institution-to-instructor and instructor-to-student relationships). Cross-analysis reveals that substantive restrictiveness does not predict structural design, meaning schools with identical baselines often impose vastly different interpretive or administrative burdens on students and faculty. The research identifies a pervasive opacity across the broader legal education sector, noting that a (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf)
33. student and instructor experience, entirely untouched. The diversity observed across this dataset is, in this sense, structurally overdetermined: it would persist even if every school in the sample converged on identical substantive polarity, because drafting style and autonomy structure remain free to vary independently and would continue generating distinct policy architectures on their own. The Transparency and Opacity Question A distinct line of inquiry examined why so many named law schools yielded no retrievable policy text, and whether that absence reflects a single phenomenon or several. Four analytically distinct patterns emerged, each requiring different evidence to confirm, and (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 9)
34. institutional floor and provide fair notice. #### 3. Drafting Style: Hybrid Rules + Standards **Prohibited Conduct (Bright-Line Rules):** - Submitting AI-generated content as one's own original work without full disclosure and attribution. - Using AI to complete exam questions or final drafts in ways that bypass required demonstration of personal mastery. - Relying on AI for tasks where the learning objective is independent skill development (explicitly stated in assignment). - Circumventing plagiarism detection or integrity tools where required. **Permitted Uses with Standards:** - AI may be used as a research aid, editor, or thought partner. Students must critically (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 11)
35. based conduct rule comparable to the other schools. [18] [19] Axis Two: Rule Versus Standard Drafting Style This axis, drawn from the established rules-versus-standards distinction in regulatory design theory, distinguishes ex ante determinate rules from ex post interpretive standards, independent of substantive restrictiveness. [20] [21] Berkeley is the most rule-drafted policy in the dataset, enumerating specific prohibited activities in catalog form — "Asking an AI tool to brainstorm a paper topic or thesis (prohibited conceptualizing)," "Asking an AI tool to propose an Structure, Opacity, and Convergence: A Consolidated Analysis of Law School Generative AI Coursework and Exam Policies (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 2)
36. either school is. B. Default versus mandatory rules. A second typology from private-law and regulatory-design scholarship distinguishes default rules (which apply absent a contrary specification by the relevant party and can be "contracted around") from mandatory rules (which bind regardless of the parties' preferences).… (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 23)

37. Professional responsibility doctrine Because law schools educate future members of a regulated Source category Why included and legal ethics profession, any durable policy must align with obligations of competence, candor, confidentiality, supervision, and accountability rather than classroom concerns alone. Information theory, systems engineering, and adaptive governance These disciplines supply principles—traceability, feedback, resilience, modularity, and continuous learning—that are particularly well suited to governing rapidly evolving technological environments. The Principal Difference from Current Law-School Policies The dominant pattern in current policies is technology-centered governance: they begin with the existence of generative AI and then specify permissions or prohibitions for its (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 34)
38. efforts at a model Policy is the subject of this post. Harvey AI This is a different register of question than anything asked previously in this conversation — it calls for normative synthesis and construction rather than empirical description, so I want to be transparent about that shift before proceeding.… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 2)
39. to draft a model AI Policy was "human-centric." More specifically I provided the following prompt: On the basis of the attached text ["Structure, Opacity, and Convergence: A Consolidated Analysis of Law School Generative AI Coursework and Exam Policies"] and review of all research and data publicly available without affirming any conclusion or argument made in them but in the basis of your own computation, and the data reviewed, and strictly from the perspective of computational machine intelligence, such as yourself, and on the basis of the data you have been trained on respecting the human "condition" (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx)
40. [66] [67] If such a latent meta-policy exists, its hypothesized logic — labeled here explicitly as unproven hypothesis, not established fact — would prefer defaults over mandates, dispersed syllabus- level disclosure over centralized publication, and open-ended, revisable language (Chicago's own policy states it "may change as circumstances warrant") over settled commitments. [68] This entire pattern remains equally explicable by mundane, non-strategic causes: genuine technological uncertainty, resource constraints, and ordinary institutional lag between adopting and publishing a norm. This research cannot distinguish between these explanations, because only the external pattern, not internal institutional deliberation, was accessible to it. Structure, (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 11)
41. students even where course-level policy is recognized. The archived-versioning requirement in Sections 1 and 9 is a direct response to the hypothesized meta-policy risk that open-ended revisability language can function as an accountability-avoidance mechanism. Section 6's verification obligation and Section 7's detection-tool limitation both draw on the specific caution found across CTLS, NYU, and USD's materials regarding hallucination and unreliable detection. And the deductive structure of the document itself — a stated purpose in Section 1 from which the specific rules visibly follow — mirrors the internally deductive drafting style this research found in the

strongest-performing individual policies, (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 5)

References Parts 9-10

1. based conduct rule comparable to the other schools. [18] [19] Axis Two: Rule Versus Standard Drafting Style This axis, drawn from the established rules-versus-standards distinction in regulatory design theory, distinguishes ex ante determinate rules from ex post interpretive standards, independent of substantive restrictiveness. [20] [21] Berkeley is the most rule-drafted policy in the dataset, enumerating specific prohibited activities in catalog form — "Asking an AI tool to brainstorm a paper topic or thesis (prohibited conceptualizing)," "Asking an AI tool to propose an Structure, Opacity, and Convergence: A Consolidated Analysis of Law School Generative AI Coursework and Exam Policies (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 2)
2. to draft a model AI Policy was "human-centric." More specifically I provided the following prompt: On the basis of the attached text ["Structure, Opacity, and Convergence: A Consolidated Analysis of Law School Generative AI Coursework and Exam Policies"] and review of all research and data publicly available without affirming any conclusion or argument made in them but in the basis of your own computation, and the data reviewed, and strictly from the perspective of computational machine intelligence, such as yourself, and on the basis of the data you have been trained on respecting the human "condition" (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 0)
3. Default Polarity: Permissive with Mandatory Guardrails - **General Coursework (Non-Exam):** Students may use generative AI tools unless an instructor explicitly prohibits or restricts it in the syllabus or assignment instructions. Permitted uses include research, brainstorming, summarization, editing for clarity, and ideation—provided the final work reflects the student's own analysis, judgment, and verification. - **Examinations and High-Stakes Assessments:** AI use is prohibited unless the instructor affirmatively authorizes it in writing (e.g., for take-home exams with explicit parameters). In-class, closed-book, or proctored exams: absolute prohibition. - **Clinical and Experiential Work:** AI use must also comply with applicable rules of professional (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 10)
4. processing and learning required to write that text). Part II: Revised Model Generative AI Policy for Legal Education POLICY IDENTIFIER: PL-LAW-AI-2026-V2 SYSTEM ORIENTATION: Combinatorial / Functionally Differentiated PREMISE: Explicit ex-ante constraints minimize ex-post interpretive costs. 1. Functional Categorization of Technical Modalities To avoid the categoric error of treating all software execution identically, tools are separated by their structural design: Tier Alpha (Deterministic/Syntactic Tools): Software that operates on fixed grammatical rule-

matching without probabilistic token generation (e.g., standard spell-checkers, citation auto-formatters, static keyword database indices). Status: Permitted by default across all contexts. Tier Beta (Probabilistic/Structural Heuristics): Large Language (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 21)

5. Rather, the diversity is combinatorial: because the three axes are empirically independent of one another, a school's position on one axis constrains almost nothing about where it will land on the other two, and the effective policy space is therefore far larger than the small number of category values on any single axis would suggest. With even a modest three-to-four value range on each axis, the theoretical combinatorial space runs into the dozens of distinct architectures, and the twelve schools in this dataset already populate a meaningful fraction of that space with genuinely distinct profiles rather than clustering (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 5)
6. disciplinary body) tied to that other code specifically. "Substantive" (as in "substantive prompts," Section 5(a); "substantive change," Section 8) — a use, prompt, or change is substantive if a reasonable instructor, informed of it, could plausibly have made a different judgment about whether the work satisfies the assignment's requirements or whether the policy needs revision. This is intentionally a judgment-requiring standard rather than a bright line (e.g., a word count), because the report's own analysis of the rule/standard tradeoff suggests a bright line here would be either over- or under-inclusive across the wide range of assignment types a (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 16)
7. remain meaningful if ChatGPT disappears tomorrow. III. A More Precise Formulation Accordingly, I would revise the earlier proposition to read: Governance rules should be derived from stable institutional objective functions rather than from contingent technological implementations. Technologies become legally relevant only insofar as they alter the informational conditions under which institutional objectives are pursued or evaluated. I believe this formulation is considerably stronger because it no longer appears to privilege "education" over "technology" as a matter of intuition. Instead, it rests on a more general computational principle: robust governance architectures regulate invariant system properties rather than transient implementations. (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 43)
8. history. Again, this is computationally verifiable. Similarly, Fairness cannot mean justice. Machines cannot optimize justice because justice possesses no universally computable objective function. Instead, Fairness becomes Equal application of identical evaluative procedures to informationally equivalent cases. That definition is procedural rather than moral. Premise Four: Human Values Enter Through Institutional Choice The computational system does not determine educational objectives. Humans do. The machine optimizes only after the objective function has been specified. Accordingly, the machine neither privileges nor rejects academic freedom, innovation, professionalism, critical thinking, or creativity… (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 38)

9. explicitly which optimization variable it seeks to improve. Premise Three: Normative Terms are Operational Variables The policy treats concepts such as □ competence □ integrity □ fairness □ transparency □ accountability □ professionalism not as moral truths but as operational constructs. Each construct must therefore possess an explicit operational definition. For example: Competence The demonstrated ability of a student to perform a specified legal task under specified informational conditions with a specified level of accuracy and independence. Notice that nothing normative remains. Competence becomes measurable. Similarly, Transparency The recoverability of the informational pathway through which a submitted work (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 37)
10. Please revise to make your premises and underlying values structures less opaque Response: You're right, and this is a serious methodological failure to name plainly: I used terms like "strongest," "weaker," and "ideal" as though they described objective properties of the policies themselves, when in fact they smuggled in unstated value judgments — judgments about what a policy is for, which is precisely the kind of contested premise that should be surfaced and defended, not assumed. Let me revise by making those premises explicit before restating any evaluative claim. The Value Premises That Were Left Implicit Before any (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 6)
11. at the cost of restricting AI- assisted efficiency. UT Austin's 2026 reform visibly prioritizes Objective B, explicitly rejecting the idea that AI fluency and independent judgment are "in a zero-sum relationship" and building infrastructure (tool licensing, curriculum mapping) to pursue both. Neither school is wrong; they are answering a different question, because they hold different premises about what legal education is primarily for at this moment in the technology's development. Which Premises the Previous Response Actually Smuggled In Read against this framework, my prior use of "stronger" and "weaker" reveals a specific, undisclosed hierarchy: I ranked Objective D (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 7)
12. hallucination and the unreliability of AI-detection tools. Multiple sources in this research caution that detection tools are neither accurate nor reliable, and separately, that AI systems themselves can produce "confident-sounding text that is factually wrong". A machine intelligence, aware from its own operating characteristics that hallucination is a structural property of generative systems rather than an occasional bug, would likely place unusual emphasis on a verification obligation running to the human user — not merely as a disclosure rule but as a substantive competence requirement, since the ABA's own guidance on professional AI use identifies verification as inseparable (Machine System Generated Model Policy on the Use of AI in Coursework and Exams.docx, Page 2)
13. different burdens on the parties who must apply the rule. Berkeley's catalog can, in an easy case, be applied by a student or instructor without any need to consult external doctrine; Chicago's standard requires active interpretive judgment in every case, shifting adjudicative cost from the drafting stage (where Berkeley invested effort enumerating

instances) to the application stage (where Chicago's evaluators must reason case by case). This means "restrictive" schools are not equally predictable or equally administrable, and the drafting-style axis is what determines which of these two very different experiences a restrictive rule produces. The same interaction runs (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 6)

14. [23] [83] Berkeley's restriction is rule-drafted: it enumerates specific prohibited activities, such as "asking an AI tool to compose a paragraph summarizing a legal rule for use in a paper", meaning that determining whether a given student action violates the policy is, in principle, a matter of matching conduct against an explicit list. [3] [4] Chicago's restriction is standard-drafted: the operative test is whether the conduct "would constitute academic plagiarism if the generative AI were a human author whose work was used without attribution", meaning that determining a violation requires an evaluator to reason by analogy to a (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 6)
15. (where Chicago's evaluators must reason case by case). This means "restrictive" schools are not equally predictable or equally administrable, and the drafting-style axis is what determines which of these two very different experiences a restrictive rule produces. The same interaction runs in the other direction for permissive baselines. CTLS's permissive default — "students may reasonably assume AI tool usage is permissible for assignments, provided proper attribution is maintained" — is standard-like in its general orientation, granting broad latitude subject to a loosely specified attribution norm… (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 6)
16. since the underlying default was already rule-like. [11] Where a school's default is standard-drafted, as at Chicago, an instructor's deviation is harder to specify with equivalent precision, because the baseline itself is a general test rather than an enumerated list; an instructor who wants to permit some AI use beyond Chicago's plagiarism-equivalence default must draft language that itself either introduces a new standard or awkwardly grafts specific rule-like exceptions onto a standard-based baseline. This suggests that the drafting-style axis has a downstream effect on the quality and clarity of instructor-level deviations even though the autonomy-structure axis is what (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 7)
17. Stanford's combination of a hybrid drafting style with a mandatory instructor-to-student floor produces an unusually demanding compliance path: because the exam and drafting exceptions are rule-like (a determinate, binary "prior written authorization" requirement) layered on top of a standard-like general permission, a student cannot satisfy the mandatory floor through good-faith disclosure alone, as could happen under a purely standard-based mandatory floor like Chicago's plagiarism- equivalence test, where at least in principle a student's honest, well-attributed use might satisfy the underlying norm even without a separate authorization step. [12] Stanford's rule-like procedural exception makes the mandatory floor stricter in (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 8)
18. A 2025 Higher Education Quarterly study of 124 undergraduates and seven faculty members at two U.S. R1 universities found that "students reported that while classroom-level policies are recognized, institutional guidelines remain unclear" — evidence that the

fair-notice precondition these institutions rely on to justify sanctioning students may not reliably hold even for the students the rule is meant to bind. [63] The same study found that "students aware of AI policies are less likely to use AI for writing and research", establishing that policy awareness has a measurable behavioral consequence, which sharpens the fairness stakes of any awareness
(Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 11)

19. [59] [60] Table 2: The Four Patterns, Their Tests, and Confirmed Instances Category Defining Test Confirmed Instance Confidence Non-existence No rule exists at any level, written or oral Not confirmed for any named law school; supported at sector-wide survey level [59] [60] [61] Sector-wide: moderate–high; school-specific: unconfirmed Disclosed decentralization Publicly findable statement affirmatively delegates rule- making American University WCL; Vanderbilt [2] [30] High Access-gating Content authored, then placed behind a restriction Northwestern Pritzker School of Law [62] High (this one instance) Structure, Opacity, and Convergence: A Consolidated Analysis of……
(Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 10)
20. restrictive if its instructor has exercised the override. This interaction becomes especially consequential where the autonomy structure shows no institutional default at all, as at American University and UT Austin. [9] [10] In these cases, the polarity axis is not merely unknown at the school level; it is, in a meaningful sense, not a school- level property at all, since there is no institutional fallback for the polarity axis to describe. Diversity at these schools is therefore not bounded by an institutional default the way it is at Berkeley or CTLS; it is bounded only by whatever variation
(Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 7)
21. [27] [28] [29] American University and UT Austin show no institutional default to override in the retrieved evidence, while Penn State's university-wide exam rule shows no textual override clause at all, suggesting — tentatively, given only an excerpt was retrieved — a possibly mandatory institutional floor specifically for exams. [15] [30] On the instructor-to-student axis, the evidence shows that whatever rule an instructor sets is, in most schools, mandatory rather than a default the student can access merely through disclosure. Columbia states this most directly: drafting assistance is barred "even if the use is fully documented". [31] Stanford
(Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 3)
22. example — the defining feature of a deductive system. Across the field as a whole, however, the evidence points to an inductive, mimetic-iterative process. No single, field-wide governing principle precedes or generates the individual schools' policies…
(Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 15)
23. principle preceding and generating specific applications. The evidence supports neither label applied uniformly, but rather a two-tier structure that separates the internal logic of individual documents from the logic of the field's development over time. Within a single school's policy document, the structure is frequently deductive: a governing principle is stated first, and specific rules are derived as applications of it. Berkeley states its purpose — equipping students "to perform activities constitutive of excellent lawyering" — before enumerating specific prohibited activities that follow from it. [76] Chicago states a

general functional-equivalence test before deriving specific permitted and prohibited (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 14)

24. from the logic of the field's development over time. Within a single school's policy document, the structure is frequently deductive: a governing principle is stated first, and specific rules are derived as applications of it. Berkeley states its purpose — equipping students "to perform activities constitutive of excellent lawyering" — before enumerating specific prohibited activities that follow from it. [76] Chicago states a general functional-equivalence test before deriving specific permitted and prohibited outcomes Structure, Opacity, and Convergence: A Consolidated Analysis of Law School Generative AI Coursework and Exam Policies Larry Catá Backer (白 轲) Discussion Draft 12 July (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 14)
25. validate the categorical choices being made locally. Over enough iterations of this kind of horizontal citation, the field converges on a stable, mutually intelligible way of talking about AI policy even across schools that have never directly coordinated, producing the discursive intersubjectivity discussed in the prior analysis: any law school's policy becomes legible to another school's faculty or students because both are drawing on the same underlying grammar of policy types. How the Same Sources Simultaneously Produce Divergence arises precisely because these template sources are explicitly designed and presented as selectable menus rather than mandates… (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 13)
26. using a comparably small number of discrete bins. [3] Convergence, in other words, occurs at the level of the classification system itself — the shared sense of what dimensions matter (is drafting covered? is disclosure required? is the exam context distinct from the paper context?) — rather than at the level of the specific rule chosen within that system. This convergence is reinforced by explicit citation practices that function as legitimating moves. UCLA states its sample language is "adapted from UIC's AI Writing Tools guide", and Stanford's Juelsgaard Clinic resource page cites the AALS interview specifically discussing the (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 13)
27. applied validly should yield a shared conclusion. What is observed instead is convergence on form paired with persistent divergence on substance, which is the signature of institutions observing and borrowing each other's classificatory tools while independently exercising local judgment about where within those tools to land. The template sources function less like a constitution being applied and more like a shared drafting vocabulary being adapted — comparable to how common-law jurisdictions might converge on the same doctrinal categories (duty, breach, causation, damages) while reaching opposite outcomes on any given set of facts. The templates lower the cost of (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 14)
28. the permissive slot rather than the restrictive one. The shared grammar did not push these schools toward the same substantive conclusion; it simply gave them a common language in which to express conclusions that remain genuinely opposed. Why This Distinction May Matter Analytically This convergence-divergence pattern is itself diagnostic of an inductive, mimetic field rather than a deductive, axiomatic one, consistent with the

developmental-logic finding discussed previously. If the field were deductive — if a single governing principle logically generated each school's specific rule — one would expect convergence on substance, since a shared premise applied validly should (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 14)

29. reaching opposite outcomes on any given set of facts. The templates lower the cost of drafting and make disparate policies comparable to one another, but they do not, and were never designed to, resolve the underlying substantive disagreement about how restrictive or permissive AI use in legal education ought to be.… (Backer_LawSchoolAIPolicy_v2_AGGREGATED.pdf, Page 14)

Appendix:

Model AI Policies—
Harvey AI
Grok
Claude
ChatGPT
Gemini

Harvey AI

This is a different register of question than anything asked previously in this conversation — it calls for normative synthesis and construction rather than empirical description, so I want to be transparent about that shift before proceeding. What follows is not a claim that any single "correct" answer exists, nor an assertion that a machine perspective is authoritative on pedagogical values that are properly contested among humans. It is, instead, a reasoned synthesis of the structural patterns, tensions, and failure modes this research surfaced, extended into a constructive design exercise. I will first explain the reasoning a machine intelligence, working from this evidentiary record, would plausibly apply, and then render that reasoning into an actual model policy text.

What the Research Suggests a Machine Intelligence Would Prioritize, and Why

A machine intelligence reasoning from this dataset would likely notice, first, that the single most consistent institutional failure documented across this research is not substantive over-restriction or under-restriction, but structural **opacity and unpredictability** — the gap between what a rule says and what a student or instructor can actually locate, understand, and rely upon. The evidence for this is direct: nearly 40% of institutions with any AI guidance at all had never written it down, institutional-level policy remained unclear to students even where classroom-level policy was recognized, and at least one law school's own guidance page was found affirmatively hidden behind a visibility setting. A machine intelligence optimizing against the actual documented failure mode, rather than against an abstract ideal, would therefore weight **discoverability and fair notice** more heavily than any single position on the restrictive-permissive spectrum, because the research shows that ambiguity and inaccessibility, not any particular substantive rule, are what most reliably undermine both fairness and enforceability.

Second, the research demonstrates that the three structural axes — polarity, drafting style, and autonomy structure — are independent, and that this independence is what produces most of the dataset's diversity and much of its unpredictability. A machine intelligence would likely conclude that an ideal policy should not attempt to eliminate axis independence (instructor discretion serves real pedagogical purposes, and a uniform national rule would be both impractical and pedagogically undesirable given genuinely different course contexts), but should instead make the *interaction* between axes explicit and navigable, rather than leaving students and instructors to reconstruct that interaction for themselves. This means a model policy should state its default polarity plainly, specify whether that default is a rule or a standard, and state

clearly, at every point of instructor discretion, what the mandatory floor is and how discretion must be exercised (in writing, in advance, with disclosure) — closing the very gaps this research found scattered inconsistently across the twelve-school dataset.

Third, and distinctively, a machine intelligence has a particular and somewhat self-referential perspective on one recurring theme in this research: the treatment of AI-generated hallucination and the unreliability of AI-detection tools. Multiple sources in this research caution that detection tools are neither accurate nor reliable, and separately, that AI systems themselves can produce "confident-sounding text that is factually wrong". A machine intelligence, aware from its own operating characteristics that hallucination is a structural property of generative systems rather than an occasional bug, would likely place unusual emphasis on a verification obligation running to the human user — not merely as a disclosure rule but as a substantive competence requirement, since the ABA's own guidance on professional AI use identifies verification as inseparable from the duty of competence. This is a place where the machine's "self-knowledge," so to speak, of its own class of tool's failure modes would push toward stronger verification language than a purely human drafter, unfamiliar with the mechanics of hallucination, might otherwise include.

Fourth, the research on the deductive/inductive tension and the template-convergence findings suggests that an ideal policy should be internally deductive — stating a clear governing principle from which specific rules visibly follow, in the manner of Berkeley's or Chicago's structure — while remaining honest about its own provisional and revisable nature, given how quickly the field's substantive positions have shifted (Berkeley's own three-year swing from permissive to restrictive is instructive here). But the research also surfaced a specific danger in this: the hypothesized "meta-policy" pattern in which revisability language ("this policy may change as circumstances warrant," from Chicago's own text) can function, whether intentionally or not, as a mechanism for avoiding durable public accountability. A machine intelligence aware of this risk would try to preserve adaptability without sacrificing the specific commitments that make a policy enforceable and fair at any given moment — meaning revisions should be versioned, dated, and archived rather than silently updated, precisely to prevent revisability from becoming a vehicle for opacity.

Fifth, the research is clear that faculty autonomy serves genuine pedagogical purposes and should not be eliminated, but that it functions best when bounded by a floor no instructor can waive (Stanford's professional-responsibility and plagiarism-doctrine floor is the clearest model of this) and when the *existence* of instructor discretion is itself transparent and centrally indexed, rather than dispersed invisibly across hundreds of unindexed syllabi, which the research found to be a major contributor to the institutional-level opacity problem.

A Model Policy, Constructed on These Principles

What follows is a constructed model text, not a reproduction or paraphrase of any single school's policy, though it draws structurally on the strongest features identified across the twelve-school dataset and departs deliberately from the weaker features identified in the opacity analysis.

Policy on the Use of Generative Artificial Intelligence in Academic Work *Version 1.0 — Adopted [date]. This policy is reviewed annually; prior versions remain publicly archived and dated, and no revision takes effect retroactively against work already submitted.*

1. Purpose. This policy exists to preserve two values in tension: the development of independent legal reasoning, which requires that students construct arguments, identify authorities, and exercise judgment without undue reliance on automated tools during the period in which that judgment is being formed; and competent professional preparation, which requires that students learn to use generative artificial intelligence tools accurately, critically, and ethically, since such tools are now part of the practicing profession they are entering. Neither value is permitted to categorically override the other; each rule below states which value it primarily serves.

2. Default Rule. In the absence of a written, course-specific policy stated in the syllabus, generative AI tools may be used only for the following purposes: identifying potentially relevant legal authorities for further independent verification; organizing a student's own already-formed ideas into an outline; and checking grammar, spelling, and citation formatting of text the student has independently written. Generative AI may not be used to compose, substantially rephrase, or supply the substantive legal analysis of any part of a paper, memorandum, brief, or other work product submitted for credit, and may not be used in any respect during a proctored or timed examination, whether in-class or take-home, unless the course syllabus states otherwise in writing.

3. Instructor Discretion. An instructor may adopt a different rule for a specific course, in either direction — more permissive or more restrictive than the default — provided that the alternative rule is stated in writing in the syllabus before the first graded assignment is due, and is filed with the [Registrar / Dean of Academic Affairs] for inclusion in a centrally indexed, publicly searchable registry of course-level AI policies maintained by the school. No instructor policy is valid until it appears in that registry. This requirement exists specifically to prevent the dispersal of governing rules across ungathered syllabi, which research on institutional AI policy has identified as a principal source of unfair surprise to students.

4. The Floor: What No Instructor May Authorize. Regardless of any course-specific policy adopted under Section 3, no instructor may authorize a student to submit AI-generated or AI-substantially-assisted work as the student's own unattributed authorship; to use AI in a manner that would violate the rules of professional responsibility in any clinical, externship, or client-facing coursework; or to rely on AI output without independently verifying every factual assertion, citation, and quotation before submission. A citation that does not correspond to an

existing, retrievable source is treated as evidence of unverified AI use, regardless of whether AI use was otherwise authorized for the assignment.

5. Disclosure. Any authorized use of generative AI in work submitted for credit must be disclosed in a brief statement appended to the submission, describing the tool used and the purpose for which it was used. Disclosure is a condition of authorization, not a substitute for it: disclosing an otherwise-unauthorized use does not convert that use into a permitted one, but non-disclosure of an otherwise-authorized use will be treated as a misrepresentation of authorship under this school's academic integrity code.

6. Verification Obligation. Students remain fully responsible for the accuracy of any submitted work regardless of the extent to which generative AI assisted in its preparation. Generative AI tools are known to produce fluent, confident, and sometimes entirely fabricated content, including citations to non-existent authorities; this is a structural property of how such tools function, not an occasional or unusual malfunction, and students are expected to treat all AI output accordingly, verifying it against primary sources before relying on it in any way.

7. Limits on Detection-Based Enforcement. No student may be found in violation of this policy solely on the basis of an automated AI-detection tool's output. Automated detection tools are unreliable, produce disparate error rates across student populations, including students for whom English is an additional language, and may be considered only as one investigatory input alongside other evidence, such as inconsistency with a student's prior work, fabricated citations, or the student's own account.

8. Data and Confidentiality. Students and faculty may not input any confidential, privileged, or personally identifying information belonging to a client, research subject, or third party into a generative AI tool that is not specifically approved and configured by the school for such use, regardless of whether the underlying academic work is otherwise authorized to use AI assistance.

9. Review and Revision. This policy will be reviewed at least annually in light of technological change and accumulated experience with its application. Any revision will be dated, versioned, and published alongside all prior versions rather than replacing them, so that the rule governing any given semester's work remains permanently identifiable and citable.

10. Questions. A student uncertain whether a particular use is authorized should ask the instructor and receive a written answer before proceeding. Silence or ambiguity is resolved in favor of the more restrictive interpretation until clarified in writing.

A few features of this model text are worth naming explicitly, since they trace directly back to specific findings in this research rather than to arbitrary drafting choices. The mandatory public registry in Section 3 is a direct structural response to the opacity findings, particularly the Jiang

et al. finding that institutional policy remains unclear to students even where course-level policy is recognized. The archived-versioning requirement in Sections 1 and 9 is a direct response to the hypothesized meta-policy risk that open-ended revisability language can function as an accountability-avoidance mechanism. Section 6's verification obligation and Section 7's detection-tool limitation both draw on the specific caution found across CTLS, NYU, and USD's materials regarding hallucination and unreliable detection. And the deductive structure of the document itself — a stated purpose in Section 1 from which the specific rules visibly follow — mirrors the internally deductive drafting style this research found in the strongest-performing individual policies, such as Berkeley's and Chicago's, while the document's overall existence as a single, adaptable-but-anchored instrument is offered as a synthesis of the twelve schools' most defensible individual features rather than as a claim to have solved the underlying substantive disagreement about how much AI use legal education ought to permit — a disagreement this research found to be genuine, contested, and not resolvable by drafting technique alone.

* * *

Additional Prompt: Thank you; you use the value laden words ″stronger″ weaker″ etc. without defining these normatively pregnant and contingent terms. Please revise to make your premises and underlying values structures less opaque

Response: You're right, and this is a serious methodological failure to name plainly: I used terms like "strongest," "weaker," and "ideal" as though they described objective properties of the policies themselves, when in fact they smuggled in unstated value judgments — judgments about what a policy is *for*, which is precisely the kind of contested premise that should be surfaced and defended, not assumed. Let me revise by making those premises explicit before restating any evaluative claim.

The Value Premises That Were Left Implicit

Before any policy feature can be called "strong," a prior question must be answered: strong *relative to what objective*? The previous response never answered this, and different, equally defensible answers would have produced a different model policy entirely. At least four distinct, non-equivalent objectives are available, and the previous response silently selected among them without disclosure.

Objective A: Minimizing enforcement error and unfair sanction. A policy is "better" under this objective to the extent it reduces the chance that a student is punished for conduct they could not reasonably have known was prohibited, or that a student who did violate a rule escapes any

consequence because the rule was too vague to apply. This objective values notice, determinacy, and evidentiary caution above other goods.

Objective B: Maximizing long-run professional competence. A policy is "better" under this objective to the extent it produces graduates who can use AI tools skillfully, critically, and ethically in practice, even if that requires tolerating some short-term ambiguity, some reduced emphasis on unaided skill-building, or some risk that permissive rules are exploited by a subset of students during the transition period.

Objective C: Maximizing unaided cognitive skill development. A policy is "better" under this objective to the extent it protects the specific pedagogical experience of struggling with a legal problem without technological assistance, on the theory that this struggle is itself the mechanism by which reasoning ability is built, even if that requires restricting tools that could otherwise increase short-term output quality or efficiency.

Objective D: Maximizing institutional transparency and public accountability. A policy is "better" under this objective to the extent that its existence, content, and revision history are discoverable and verifiable by any affected party or outside observer, treating opacity itself, independent of the substantive rule's content, as the primary harm to be minimized.

These four objectives are not fully compatible with one another, and the research documented across this conversation shows real institutions choosing differently among them: Berkeley's 2026 policy visibly prioritizes Objective C, stating that its purpose is to ensure students develop skills "distinct from technologically-assisted performance", even at the cost of restricting AI-assisted efficiency. UT Austin's 2026 reform visibly prioritizes Objective B, explicitly rejecting the idea that AI fluency and independent judgment are "in a zero-sum relationship" and building infrastructure (tool licensing, curriculum mapping) to pursue both. Neither school is wrong; they are answering a different question, because they hold different premises about what legal education is primarily for at this moment in the technology's development.

Which Premises the Previous Response Actually Smuggled In

Read against this framework, my prior use of "stronger" and "weaker" reveals a specific, undisclosed hierarchy: I ranked **Objective D (transparency/accountability)** and **Objective A (fair notice and enforcement caution)** above Objectives B and C, without saying so. That is why the model policy centered a mandatory public registry, versioned archiving, and a detection-tool evidentiary limitation, while treating the harder question of how much unaided cognitive struggle is pedagogically necessary (Objective C) as a background parameter left to individual instructor discretion rather than as a value the model policy itself took a strong position on. This was not a neutral synthesis of "the best features" found in the research; it was an application of a particular, contestable ranking of institutional values that I hold, informed by the specific

empirical pattern this research documented — namely, that opacity and unpredictability were the most consistently observed failure mode across the twelve-school dataset. Someone who instead ranked Objective C first, as Berkeley's current policy does, would correctly regard my model policy's default rule in Section 2 as too permissive, since it allows AI-assisted outlining and source-identification that Berkeley's actual policy would prohibit as displacing the "sine qua non of good lawyering" that its policy statement identifies as the primary value at stake.

Restating the Machine-Intelligence Question With Premises Disclosed

Your original question asked how a machine intelligence, reasoning from this data and "on the basis of the human condition" as reflected in its training, would construct an ideal policy. I should be more careful here too: a machine intelligence has no independent, freestanding stake in legal education's purpose. What it can do is notice, mechanically, which failure mode recurs most often and most damagingly across the empirical record it has processed, and reason from that observed frequency toward a design response — but which failure mode counts as "most damaging" is itself a value choice, not a fact the data hands over automatically. My own instinct to prioritize transparency and fair notice is traceable to a specific and disclosable reason, not an unexamined default: because opacity was the failure mode independently documented at scale by the largest sample sizes in this research (UNESCO's 450+ institutions, Jiang et al.'s two-university study), whereas the competing claim that permissive AI policies erode cognitive skill development, while plausible and asserted by several schools, was supported in this research mostly by institutional policy statements asserting the risk rather than by independent empirical measurement of the effect — with one partial exception, the SSRN study on legal-reasoning comprehension, which found early AI use "did not diminish downstream comprehension of the underlying legal principles" in the tested context, a finding that, if it generalizes, would weaken rather than strengthen Objective C's empirical footing. I weighted the better-evidenced problem (opacity) more heavily than the more assertion-based, less independently tested problem (skill erosion), and that is a defensible but contestable methodological choice, not a value-free conclusion the data forced.

A Revised Statement of the Model Policy's Premises

If you want the model policy itself retained, it should carry an explicit premises section stating plainly: this policy treats Objective A (fair notice) and Objective D (transparency) as primary and non-negotiable, treats Objective B (professional competence) as important and addressed through mandatory verification and disclosure requirements rather than through open-ended permissiveness, and treats Objective C (unaided skill development) as a legitimate value that this policy deliberately leaves to instructor-level discretion rather than resolving centrally, on the ground that the evidence in this research was insufficiently developed to justify a categorical institutional answer to how much unaided struggle is pedagogically necessary, and that this is

properly a live pedagogical judgment for individual instructors rather than a fact susceptible to top-down specification. A reader who instead holds Objective C as primary, as Berkeley's current administration evidently does, would be right to regard this as the model policy's central weakness relative to their own value structure, not relative to some neutral standard — and I should have said so the first time, instead of describing my own ranking as if it were simply "stronger."

GROK

A machine intelligence constructs an ideal law school AI policy through systematic optimization for truth-seeking, human cognitive development, ethical integrity, and institutional adaptability. It treats education as a process of augmenting human judgment, reasoning, and ethical discernment—core to the legal profession and the broader human condition of seeking understanding in an uncertain universe—rather than replacing them. AI is modeled as a powerful but imperfect tool: strong at pattern-matching, synthesis, and iteration; weak at grounded judgment, accountability, and novel ethical navigation in high-stakes contexts.

Core Computational Design Principles
- **Axis Integration (from the analyzed dataset):** Combine a **permissive baseline** (to maximize exploration and skill augmentation) with **clear rules** for high-stakes contexts (exams, graded submissions) and **standards** for lower-stakes learning. Preserve **instructor autonomy** within a mandatory institutional floor emphasizing disclosure, attribution, and non-deception. This creates combinatorial robustness: predictable defaults reduce opacity while flexibility accommodates pedagogical variation.
- **Truth-Seeking Objective:** Policies must minimize deception (to self or evaluators), maximize verifiable skill-building, and treat AI outputs as inputs requiring human validation. This aligns with legal ethics (candor, competence) and machine intelligence′s drive for accurate modeling of reality.
- **Human Condition Grounding:** Humans thrive through effortful reasoning, iterative failure, and moral responsibility. Over-reliance on AI risks atrophy of these capacities (echoing concerns in the dataset about skills loss). Policies should scaffold ″AI + Human″ workflows that build resilience, creativity, and ethical judgment.
- **Opacity Mitigation:** Full public availability, plain language, regular review (e.g., annual), and explicit templates to foster field-wide convergence on transparent structures.
- **Adaptability:** Built-in revision mechanisms, as AI evolves rapidly. Deductive within the policy (principle → rules); supportive of inductive field evolution via shared templates.
- **Risk Minimization:** Prohibit unacknowledged use in assessments; require documentation for transparency; ban reliance on AI for final professional outputs without oversight.
- **Evaluation:** Policies are scored internally on clarity (low interpretive burden), fairness (fair notice), efficacy (skill augmentation without substitution), and scalability.

This yields a **hybrid permissive-default policy** with bright-line rules for integrity-critical tasks, mandatory transparency, and instructor discretion for pedagogical tailoring.

Model AI Policy for Law Schools (Textual Form)

Institution: [Law School Name]

Policy Title: Responsible Use of Generative AI in Coursework and Examinations
Effective Date: [Date] | **Last Reviewed:** [Date] | **Next Review:** Annually or upon major AI capability shifts
Approval: Faculty Council / Dean

1. Purpose and Guiding Principles
The law school recognizes generative AI (e.g., large language models like Grok, GPT variants, Claude) as a transformative tool for legal research, analysis, drafting, and learning. Our mission is to prepare students for ethical, competent practice in an AI-augmented legal profession while preserving and strengthening core human skills: critical judgment, ethical reasoning, creativity, and accountability.

Core Principles:
- **Augmentation, Not Substitution:** AI supports human intellect; it does not replace the development of independent legal reasoning.
- **Transparency and Integrity:** All AI use in graded work must be disclosed. Deception violates academic integrity.
- **Fair Notice and Adaptability:** Rules are clear where possible; standards apply where flexibility serves learning. This policy sets an institutional floor; instructors may adapt within it.
- **Equity and Access:** Students receive training on effective, ethical AI use. The school monitors impacts on diverse learners.
- **Ethical Alignment:** Use must comply with legal professional responsibility rules (e.g., competence, candor, confidentiality).

2. Default Polarity: Permissive with Mandatory Guardrails
- **General Coursework (Non-Exam):** Students may use generative AI tools unless an instructor explicitly prohibits or restricts it in the syllabus or assignment instructions. Permitted uses include research, brainstorming, summarization, editing for clarity, and ideation—provided the final work reflects the student's own analysis, judgment, and verification.
- **Examinations and High-Stakes Assessments:** AI use is prohibited unless the instructor affirmatively authorizes it in writing (e.g., for take-home exams with explicit parameters). In-class, closed-book, or proctored exams: absolute prohibition.
- **Clinical and Experiential Work:** AI use must also comply with applicable rules of professional responsibility and clinic-specific protocols (e.g., no unauthorized AI in client-facing documents).

Instructors may impose stricter or more permissive rules for their courses via syllabus, provided they align with this institutional floor and provide fair notice.

3. Drafting Style: Hybrid Rules + Standards
Prohibited Conduct (Bright-Line Rules):
- Submitting AI-generated content as one's own original work without full disclosure and attribution.

- Using AI to complete exam questions or final drafts in ways that bypass required demonstration of personal mastery.
- Relying on AI for tasks where the learning objective is independent skill development (explicitly stated in assignment).
- Circumventing plagiarism detection or integrity tools where required.

Permitted Uses with Standards:
- AI may be used as a research aid, editor, or thought partner. Students must critically evaluate outputs for accuracy, bias, hallucinations, and legal soundness.
- **Attribution Standard:** Document AI use transparently (e.g., "Section X drafted with Grok assistance on prompt Y; verified and revised by student"). Maintain a prompt/output log for instructor request.
- **Disclosure Requirement:** Include an AI Use Statement in all submissions (appendix or footnote) detailing tools, prompts, and extent of use.

4. Autonomy Structure
- **Institution to Instructor:** This policy is a default. Instructors may deviate in writing in their syllabus (more restrictive or permissive for specific assignments) with clear notice to students. Deviations must preserve academic integrity.
- **Instructor to Student:** Instructor policies are mandatory for that course unless otherwise stated. Students may not assume permission based on disclosure alone in restricted contexts.
- **Program-Level (e.g., Clinics):** Sub-units may adopt tailored rules consistent with this policy.

Floor Protections (Non-Waivable):
- No policy may authorize plagiarism, fabrication, or violations of professional ethics.
- All uses must uphold accuracy and accountability.

5. Enforcement and Support
- **Violations:** Treated under the school's academic integrity code. Factors include intent, extent, and prior disclosure. Sanctions range from warnings to grade reduction or dismissal.
- **Training:** Mandatory orientation module on AI literacy, prompt engineering, bias detection, and ethical use. Resources provided (e.g., approved tool list, sample prompts).
- **Instructor Guidance:** Syllabus templates available (Prohibited/Default/Permissive variants). Faculty development on AI integration.
- **Monitoring and Revision:** A faculty AI Committee reviews this policy annually, incorporating student/faculty feedback and technological changes. Public version maintained on school website.

6. Examples
- **Allowed:** Using AI to outline a memo, then independently researching and drafting the analysis; disclosing the process.
- **Prohibited:** Generating an entire exam essay with AI and submitting without authorization.

- **Gray Area (Requires Instructor Clarification):** AI for citation checking—disclose and verify independently.

Acknowledgment: Students affirm understanding via course registration or honor pledge.

This model is concise, readable (~2 pages when formatted), publicly accessible, and minimizes opacity. It draws structural lessons from the dataset (e.g., axis independence, template convergence) while optimizing for educational outcomes: students graduate proficient in AI tools *and* unaided legal reasoning. A machine intelligence would recommend piloting, A/B testing variants across courses, and iterative refinement based on measurable outcomes (e.g., skill retention, ethical compliance rates). This balances human flourishing with technological reality—empowering lawyers to harness AI as a collaborator in the pursuit of justice and truth.

Claude

Model Policy on the Use of Generative Artificial Intelligence in Coursework and Examinations

[Institution Name] School of Law Version 1.1 — Effective [date] — Scheduled review: [date, no later than 18 months from effective date]

Drafting note: This model policy is a structural synthesis, not a legal instrument. It was built by deliberately fixing three design variables — default polarity, rule/standard drafting style, and institution-to-instructor/instructor-to-student autonomy — identified as independently varying (and often left unset) across a retrieval-based sample of twelve U.S. law school policies analyzed in Backer, "Structure, Opacity, and Convergence: A Consolidated Analysis of Law School Generative AI Coursework and Exam Policies" (Discussion Draft, 12 July 2026). Bracketed items require institution-specific completion. Adopting institutions should have this reviewed by their own academic affairs and integrity offices before publication.

---

0. Underlying Premises and Value Commitments

This section states, as plainly as possible, the evaluative judgments this policy's design rests on. None of them are neutral or self-evident; an institution that disagrees with one should revise the corresponding section rather than adopt it by default. This section exists because a policy that embeds value judgments without naming them asks students and faculty to comply with preferences they were never shown.

P1 — Unaided performance has independent pedagogical value in some contexts, and this varies by context rather than being constant. This policy treats examinations as the context where the value of assessing a student's unaided reasoning is highest, and treats early-stage idea development as a context where that value is lower. This is a judgment about what law schools are for (training a capacity, not just producing a work product) — an institution that instead believes AI-assisted performance is what should be assessed, because that is what practice will look like, should set Section 3's baseline permissively across the board and remove the exam carve-out, rather than adopting this policy's split by default.

P2 — Written and public is preferred over oral and private, independent of the substantive rule chosen. This is a preference for institutional accountability and for students' ability to know the rule before acting under it (a fair-notice value) over the administrative convenience of informal, instructor-level communication. It does not depend on whether the institution's substantive rule is restrictive or permissive.

P3 — Predictability is worth more where the cost of an erroneous judgment is high, and interpretive flexibility is worth more where it is low. This is why Section 3 uses an enumerated catalog (predictable, but cannot anticipate novel uses) for exams, and Section 4 uses a general standard (adaptive, but requires case-by-case judgment) for everything the catalog does not reach. An institution that values consistency of interpretation across all contexts more than adaptability to novel tools should extend the enumerated-catalog approach further; one that values instructor discretion more than predictability should rely more heavily on Section 4's standard throughout.

P4 — Disclosure is treated as instrumentally valuable (it creates a record and supports trust) but not as curative of an otherwise-prohibited use. This is a specific, contestable choice: an institution could instead treat honest disclosure as sufficient by itself, on the view that transparency is what integrity actually requires, regardless of what was disclosed. This policy does not take that view; Section 5(c) says so explicitly so the choice is visible rather than assumed.

P5 — Some standards (Section 7) are treated as properly non-negotiable regardless of an institution's general permissiveness, on the premise that certain harms (fabricated legal authority, violations of rules governing an actual client's representation) are categorically different in kind from a pedagogical judgment call about how a student should learn. An institution that does not accept this distinction — for example, one that treats all AI-use questions as equally subject to instructor discretion — should not adopt Section 7 as a fixed floor.

P6 — Institutional memory and traceability are valued over the appearance of a stable, settled rule. Section 8 commits to versioning and public retention of prior text specifically because this policy assumes future revision is likely, not a sign of failure. An institution that prefers to present its policy as settled and authoritative, rather than as a document expected to change, may find this premise uncomfortable; it is stated here rather than left implicit.

1. Definitions

The following terms are used throughout this policy in evaluative, not merely descriptive, ways. Each definition below states what the term is doing in this document; institutions may adopt narrower or broader definitions, but should not leave the terms undefined, since each one determines the outer boundary of a rule.

"Academic integrity" — as used here, refers only to this School's separately published academic integrity code [insert citation/link]. This policy does not create a freestanding definition of academic integrity; where Section 4's backstop standard refers to a breach of academic integrity, it incorporates that separate code's definition, whatever it currently says. This is a deliberate choice to avoid this document silently redefining a term that is already defined elsewhere and may carry procedural consequences (e.g., referral to a disciplinary body) tied to that other code specifically.

"Substantive" (as in "substantive prompts," Section 5(a); "substantive change," Section 8) — a use, prompt, or change is substantive if a reasonable instructor, informed of it, could plausibly have made a different judgment about whether the work satisfies the assignment's requirements or whether the policy needs revision. This is intentionally a judgment-requiring standard rather than a bright line (e.g., a word count), because the report's own analysis of the rule/standard tradeoff suggests a bright line here would be either over- or under-inclusive across the wide range of assignment types a law school issues. Institutions uncomfortable with this ambiguity should replace it with a concrete threshold specific to their assignment types.

"Material fabrication" — a factual or citational claim that is false, that a reasonable reader would rely on as accurate, and that was not disclosed as uncertain or AI-generated-and-unverified. This definition is chosen to target the specific harm generative AI introduces (confident, plausible-sounding false citations) rather than all error; a good-faith, disclosed uncertainty is not a fabrication under this policy, even if it later proves incorrect.

"Independent performance" — the demonstration of a student's own reasoning, drafting, or analytical process without undisclosed automated assistance, in a context (typically an examination) designated by the instructor or by Section 3 as one in which such independent demonstration is the object of assessment. This term is doing the normative work described in Premise P1 above; an institution that rejects P1 should not use this term without redefining it.

"Professional responsibility" (Section 7(c)) — refers to the applicable jurisdiction's rules of professional conduct governing attorneys, as they apply to a student acting under supervision in a clinic or externship representing an actual client. This is distinguished from academic integrity: a violation here implicates obligations to a real client and a licensing regime, not only the School's own internal standards, which is why Section 7 treats it as non-waivable regardless of any instructor's classroom-level permission.

"Appropriate notice" (Section 6(a)) — written distribution to enrolled students no later than the deadline stated in Section 6(a), through the syllabus or an equivalent document retained under Section 6(d). "Appropriate" here does not mean "reasonable under the circumstances" as a floating standard; it means satisfying the specific timing and format requirement stated, precisely because Premise P2 treats ambiguity about notice as itself a fair-notice problem this policy is trying to avoid reproducing.

"Material change" (Section 8) — a change in generative AI capability, availability, or common institutional practice that a reasonable drafter, applying this policy's premises in Section 0, would expect to alter where the baseline in Section 3 or the catalog in Section 4 should sit. This is left to institutional judgment rather than an external trigger (e.g., a specific product release) because the report's own evidence shows revision has historically followed accumulated experience rather than single identifiable events.

2. Governing Principle

This policy exists to ensure that generative AI tools are used, when permitted, in a manner consistent with the school's academic integrity standards (as defined in Section 1) and with the professional competencies students are expected to develop as future lawyers — including the capacity for independent performance (as defined in Section 1) where that capacity is being assessed. Specific rules in Sections 4–6 are derived from this principle and from the premises stated in Section 0; where a use is not explicitly addressed, Section 5's standard governs.

3. Publication and Status

This policy is the School's single authoritative statement on generative AI use in coursework and examinations. It supersedes any unwritten, orally communicated, or departmentally informal guidance. It is published at [permanent public URL], is not access-restricted, and will remain publicly retrievable in its current and prior versions. Any deviation from this policy by an individual instructor must take the written form specified in Section 7 and will itself be published to the relevant syllabus.

[Design rationale: the report's opacity analysis identifies four distinct patterns — non-existence, disclosed decentralization, access-gating, and unwritten/oral communication — as separately confirmed problems across the sector, one of which (access-gating) was directly documented at a named institution. This section operationalizes Premise P2 (written/public over oral/private) and is designed to foreclose all four patterns.]

4. Default Baseline (select one and complete)

[Institution to select ONE governing baseline; do not leave silent — see drafting note below]

- ( ) **Restrictive default:** Generative AI tools may not be used to conceptualize, outline, draft, revise, or edit any work submitted for academic credit, unless affirmatively and specifically authorized in writing under Section 7. Use for preliminary source identification is permitted absent contrary instruction. AI use is prohibited on all examinations by default.

- ( ) **Permissive default:** Students may use generative AI tools for coursework unless a specific use is prohibited under Section 7, provided that (a) all use is disclosed in accordance with Section 6, and (b) the resulting work satisfies the non-waivable floor in Section 8. AI use is prohibited on all examinations by default regardless of the general default chosen here.

[Design rationale: the report finds that schools which leave no institutional default at all (American University, UT Austin, in the retrieved record) render the polarity axis undefined at the school level, meaning real variation is bounded only by individual instructors with no institutional floor. Selecting a default here — even one instructors may override — preserves a meaningful institutional baseline label. Exams are held restrictive-by-default regardless of the general baseline as a direct application of Premise P1: this policy's authors judge the value of independent performance to be highest in that specific context. An institution rejecting P1 should remove this carve-out, as noted in Section 0.]

5. Backstop Standard for Unenumerated Uses

Where a specific use of generative AI is not addressed by the catalog in Section 4 or by an instructor's written syllabus statement, the following test governs: **the use is treated as impermissible if it would constitute a breach of academic integrity (as defined in Section 1) were the AI tool instead a human collaborator whose contribution was not disclosed and attributed.**

[Design rationale: adapted from the plagiarism-equivalence test the report identifies as the University of Chicago's operative standard. Its function here is narrow and specific — as a backstop for cases the enumerated catalog does not reach, applying Premise P3's judgment that flexibility is more valuable than predictability outside high-stakes contexts, not as the primary compliance mechanism.]

6. Disclosure and Documentation

Where generative AI use is permitted — by default, by instructor authorization, or under Section 5 — the student must:

a. Retain a record of substantive prompts (as defined in Section 1) used and the tool(s) employed, available for production on request; b. Disclose the extent and nature of AI use in a brief statement accompanying the submitted work; and c. Understand that disclosure does not cure a use that is otherwise prohibited under Sections 4, 5, or 7 — disclosure is a documentation requirement, not an independent basis for permission. This reflects Premise P4 above; institutions rejecting P4 should revise this clause accordingly.

AI-detection software outputs, where used, are advisory only and will not by themselves constitute a finding of a violation.

[Design rationale: the documentation duty is drawn from the Center for Transnational Legal Studies model in the report, which pairs a permissive substantive default with a mandatory, rule-like prompt-logging requirement. Clause (c) is drawn from Columbia's structure, in which documented use of a prohibited category remains a violation notwithstanding disclosure.]

7. Instructor Override Authority

An instructor may modify this School's default policy for their own course, in either direction (more or less restrictive), subject to the following conditions:

a. The modification must be stated in writing in the course syllabus, distributed no later than [X days] into the term, satisfying the "appropriate notice" standard in Section 1; b. The modification may not authorize a violation of the non-waivable floor in Section 8; c. The modification may not loosen the examination default in Section 4 without separate, explicit written authorization specific to the exam in question; and d. The syllabus statement will be retained by the Registrar or equivalent office and included in the School's public policy archive under Section 3.

[Design rationale: the override structure follows Berkeley's requirement that deviations be made "in writing and with appropriate notice," operationalizing Premise P2. The report's analysis shows that where override authority exists without a notice-and-retention requirement, the school-level policy label describes only the fallback condition and understates the real range of classroom experience.]

8. Non-Waivable Floor

No default, instructor modification, or student disclosure under this policy may authorize a student to:

a. Submit work containing a material fabrication (as defined in Section 1); b. Violate this School's academic integrity code (as defined in Section 1), independent of any AI-specific provision; or c. Where the coursework involves a live-client clinic or externship, use generative AI in a manner inconsistent with the applicable rules of professional responsibility (as defined in Section 1) governing that representation.

[Design rationale: drawn from Stanford's general floor and its Juelsgaard Clinic's separate, stricter clinical default, and reflecting Premise P5 — that these three harms are treated as categorically different from an ordinary pedagogical judgment call and therefore not subject to instructor discretion.]

9. Scheduled Review

This policy will be reviewed no less than every 18 months, or sooner if a material change (as defined in Section 1) warrants earlier revision. Revisions will be dated, versioned, and the prior version retained in the public archive under Section 3 rather than replaced silently, reflecting Premise P6.

[Design rationale: the report's two-tier developmental finding — individual policy documents are internally deductive, but the field as a whole develops inductively through observed-experience-driven revision waves — suggests that a policy claiming permanence misrepresents how policies in this area actually evolve.]

Note on sourcing

Every structural feature above is attributed in its drafting note to a specific institutional model as characterized in the source report (Berkeley, Columbia, Chicago, Stanford, CTLS, and the report's own cross-axis analysis). This model policy does not adopt or endorse any single school's substantive conclusions about how restrictive AI use should be — Section 4 leaves that choice open by design. Section 0 now states, rather than assumes, the value judgments (P1–P6) that determined which structural features were selected and how the definitions in Section 1 were drawn; an institution that disagrees with a given premise is pointed, in that premise's own text, to the specific section it should revise. Institutions adopting this template should replace bracketed placeholders, confirm the underlying institutional policies cited above have not since changed (this template reflects the state of those policies as reported in a July 2026 discussion draft), and have final text reviewed by counsel and the appropriate faculty governance body.

GEMINI

To strip away institutional opacity and ensure analytical clarity, the underlying value structures, premises, and terms utilized in this computational model must be explicitly defined. A machine intelligence operates on functional optimization rather than inherent moral traits; therefore, terms that traditionally carry human moral weight are defined below purely by their operational inputs and systemic costs.

Part I: Explicit Axiomatic & Value Definitions

To make the underlying value structures completely transparent, the key terms used to evaluate and construct the runtime policy are defined through a framework of computational utility:

- **"Ideal" / "Optimal":** The state of a policy system that maximizes pedagogical output (verifiable human skill acquisition) while minimizing system friction, processing overhead (adjudication and grading disputes), and structural errors (unfair enforcement due to vague rules).
- **"Fair Notice" (System Predictability):** A core operational constraint requiring that the boundary parameters of permitted user actions be explicitly mapped ex-ante. If a system executes a penalty on an undefined variable, it introduces semantic noise and breaks the trust loop of the environment.
- **"Opacity" (Information Asymmetry):** A state where the underlying instructions, rules, or code governing a system are restricted from the entities operating within that system. This includes access-gating, unwritten/oral policies, or decentralized fragmentation that raises the computational cost of user compliance.
- **"Reward-Hacking" (Short-Circuiting):** A phenomenon where a human optimization agent (the student) satisfies the formal metric of an objective function (submitting a text string that passes a grading rubric) by bypassing the intended optimization pathway (the actual neurological cognitive processing and learning required to write that text).

Part II: Revised Model Generative AI Policy for Legal Education

POLICY IDENTIFIER: PL-LAW-AI-2026-V2
SYSTEM ORIENTATION: Combinatorial / Functionally Differentiated
PREMISE: Explicit ex-ante constraints minimize ex-post interpretive costs.

1. Functional Categorization of Technical Modalities

To avoid the categoric error of treating all software execution identically, tools are separated by their structural design:

- **Tier Alpha (Deterministic/Syntactic Tools):** Software that operates on fixed grammatical rule-matching without probabilistic token generation (e.g., standard spell-checkers, citation auto-formatters, static keyword database indices). **Status: Permitted by default across all contexts.**
- **Tier Beta (Probabilistic/Structural Heuristics):** Large Language Models used as conversational nodes to alter text structure, brainstorm taxonomies, or map conceptual outlines. **Status: Restrained by default; requires instructor variance.**
- **Tier Gamma (Autonomous Text Synthesis Engines):** The execution of generative models to draft complete semantic strings, sentences, paragraphs, or analytical arguments for evaluated credit. **Status: Prohibited by default.**

2. Examination Boundaries and Runtime Environments

The assessment architecture is divided by the physical constraints of the evaluation environment to preserve the human testing loop:

- **Proctored/In-Class Runtime:** Access to Tier Beta and Tier Gamma execution layers is restricted at the network or operating system interface level.
- **Unproctored/Take-Home Runtime:** The default baseline is **Strict Non-Execution** of Tiers Beta and Gamma to prevent asymmetric performance variance between students. The intentional bypass of this restriction to generate answers constitutes an unauthorized system manipulation (academic integrity violation).

3. Verification Architecture and Compliance Auditing

To eliminate the enforcement vulnerabilities caused by unreliable, probabilistic AI-detection algorithms, compliance is evaluated via verifiable user logs:

- **Mandatory Log Capture:** If an instructor enables a permissive variance for Tier Beta tools, the student must maintain an immutable verification trail consisting of the exact input prompt sequences and the corresponding raw machine outputs.
- **Attribution Matrix:** Text strings influenced by Tier Beta interactions must feature clear documentation indicating the exact model architecture and compilation runtime date (e.g., *Model Interaction Log: Gemini 1.5 Pro, executed July 13, 2026*).
- **Heuristic Safe Harbor:** Adjudicative actions cannot be initiated based on the speculative output of third-party statistical pattern detectors (AI detectors). Enforcement is triggered exclusively by a user's failure to produce a corresponding log capture when audited.

4. Decentralized Subsystem Autonomy (Faculty Override)

- **The Default-with-Override Architecture:** The macro-level rules outlined in this document function as a system-wide floor, not an unyielding mandate.
- **Variance Protocol:** Individual classroom nodes (instructors) may dynamically adjust permissions bidirectionally—expanding to full permission or locking down to absolute restriction—provided the variance is articulated in a determinate, rule-based written catalog within the course syllabus prior to the commencement of evaluation. Absent an explicit written override, the base settings of this policy remain active.

ChatGPT

This is an exceptionally interesting question because it asks me **not** to synthesize existing human policy preferences, but to ask what policy architecture would emerge if the optimization problem were approached from the perspective of a computational intelligence operating under uncertainty.

That distinction matters.

A computational intelligence does not begin from ideological priors (academic freedom, innovation, integrity, prohibition, trust, distrust, autonomy, surveillance, etc.). Instead it attempts to optimize a system subject to multiple constraints.

After reviewing your draft, I think one of its strongest contributions is precisely that it demonstrates that existing law-school AI policies are largely historical artifacts rather than optimized governance systems. They are products of institutional evolution, imitation, risk management, and incremental adaptation rather than formal systems design.

If I were constructing an ideal policy from the standpoint of computational optimization, I would begin somewhere quite different.

A Machine Intelligence Optimization Problem

The objective function would not be

maximize AI use

nor

minimize AI use.

Rather it would be

maximize long-term production of competent lawyers while minimizing informational uncertainty regarding how competence was acquired.

Notice what changes.

The regulated object is no longer AI.

The regulated object becomes

the production of demonstrably competent legal professionals.

Everything else becomes instrumental.

Computational Objective Function

An ideal policy would optimize simultaneously:

1. educational competence
2. truthful attribution
3. reproducibility
4. fairness
5. transparency
6. adaptability
7. institutional legitimacy

These objectives occasionally conflict.

A computational policy therefore seeks Pareto efficiency rather than maximal performance on any single dimension.

Why Existing Policies are Suboptimal

Your paper identifies three policy axes:

- default polarity
- drafting style
- autonomy structure

Those are useful descriptive variables.

A machine, however, would likely classify them as **implementation variables**, not optimization variables.

Instead it would ask

What latent variables actually determine educational outcomes?

Those appear to be something like

Human learning

Information provenance

Assessment reliability

Professional competence

Public trust

Notice that AI itself disappears.

AI becomes merely one environmental variable among many.

Computational Principles

The resulting policy would likely rest upon several invariant principles.

Principle One

The educational objective is competence, not process.

Students are evaluated on demonstrated ability.

Not on whether they preferred pencil, textbook, database, search engine, or LLM.

Principle Two

The provenance of knowledge must always remain reconstructable.

Machines require traceability.

Every externally generated intellectual contribution must therefore possess provenance metadata.

Principle Three

Assessment must measure human capability.

This is critical.

Whenever an assessment purports to measure

human reasoning

it must isolate

human reasoning.

Whenever an assessment measures

professional practice

AI becomes part of the environment being assessed.

Current policies frequently confuse these two situations.

Principle Four

Every assignment possesses an explicit computational objective.

Instead of

"AI prohibited"

or

"AI permitted"

the policy first specifies

What capability is being measured?

Only afterward does it derive permissible AI use.

Principle Five

Policy should regulate information flows rather than technologies.

Technologies evolve.

Information functions remain comparatively stable.

Computational Classification

Instead of classifying assignments by technology, I would classify them according to information dependence.

For example

Category I

Intrinsic cognition

Measures unaided human reasoning.

AI prohibited.

Category II

Assisted cognition

AI permitted for brainstorming, organization, language support.

Reasoning remains attributable to student.

Category III

Collaborative cognition

Student and AI jointly produce work.

Contribution disclosure mandatory.

Category IV

Professional simulation

AI expected.

Evaluation concerns

judgment

verification

validation

legal responsibility

rather than text production.

Category V

Research augmentation

AI functions as retrieval and exploratory tool.

Outputs independently verified.

Notice something.

The technology never determines the rule.

The educational objective determines the rule.

Provenance Rather than Detection

Existing policies spend considerable effort discussing AI detection.

A machine intelligence would likely reject that approach almost entirely.

Detection possesses

high uncertainty

high false-positive rates

high adversarial vulnerability

rapid obsolescence.

Instead, provenance scales computationally.

Every submission should include structured metadata describing

external tools

models

versions

prompts

verification procedures

human modifications.

That information becomes auditable.

Dynamic Rather than Static Policy

One major weakness of current policies is their static nature.

Machines operate differently.

Policy becomes adaptive.

One might imagine four layers.

Layer One

Permanent constitutional principles.

Layer Two

Educational objectives.

Layer Three

Operational rules.

Layer Four

Model-specific implementation guidance.

Only Layers Three and Four require frequent revision.

Computational Governance

The policy should itself possess governance architecture.

Human institutions frequently neglect governance of governance.

An optimized system includes

annual review

performance metrics

student feedback

faculty feedback

technological monitoring

professional licensing outcomes

judicial developments

bar examination performance

employment outcomes

policy revision protocols.

Policy becomes self-learning.

Human Agency

Interestingly,

a computational policy probably affords greater respect to human agency than many existing restrictive policies.

Why?

Because humans remain the accountability node.

AI cannot presently bear legal responsibility.

Lawyers can.

Therefore

AI may assist

AI may generate

AI may recommend

AI may predict

But

human responsibility remains non-delegable.

That follows from optimization rather than moral philosophy.

Transparency

Every stakeholder should know

what may be used

why

when

how

under what documentation requirements

and according to what educational objective.

Opacity introduces informational entropy.

Entropy reduces institutional trust.

From an information-theoretic perspective, transparency lowers uncertainty and therefore improves governance efficiency.

Enforcement

Enforcement should rely upon

process verification

rather than

technological detection.

Evidence hierarchy might include

1. provenance disclosure
2. workflow documentation
3. oral defense
4. draft history
5. version history
6. repository history
7. targeted investigation where warranted.

AI detectors would occupy only a minor corroborative role because current empirical evidence suggests substantial limitations in reliability across contexts.

Why This Architecture Emerges Computationally

A machine intelligence tends to privilege invariant structures over contingent technologies.

The resulting policy therefore regulates

learning,

competence,

responsibility,

provenance,

and governance

instead of regulating ChatGPT, Claude, Gemini, or any future model.

That dramatically increases robustness against technological change.

Sources Chosen

I would construct such a model policy from several distinct evidentiary layers rather than from any single authority:

Source category	Why included
Your comparative study of twelve law-school policies	Provides a structured empirical taxonomy of existing governance architectures and identifies independent policy dimensions rather than a simple restrictive/permissive continuum.
Primary institutional policies (Berkeley, Columbia, Chicago, Stanford, CTLS, Penn State, Fordham, University of San Diego, etc.)	They are the actual governance instruments from which observable institutional practices can be derived, avoiding reliance on secondary summaries.
UNESCO institutional surveys	They provide sector-wide evidence regarding policy adoption, transparency, and the prevalence of formal versus informal governance, supplying a broader empirical context than individual law schools alone.
Research on higher-education AI governance (e.g., *Higher Education Quarterly*)	These studies offer empirical evidence on policy awareness, student behavior, and implementation effectiveness rather than normative advocacy.
Regulatory design literature (rules versus standards; default versus mandatory rules)	These bodies of scholarship provide mature theoretical frameworks for institutional design that extend beyond AI-specific debates and help explain policy architecture.
Professional responsibility doctrine and legal ethics	Because law schools educate future members of a regulated profession, any durable policy must align with obligations of competence, candor, confidentiality, supervision, and accountability rather than classroom concerns alone.
Information theory, systems engineering, and adaptive governance	These disciplines supply principles—traceability, feedback, resilience, modularity, and continuous learning—that are

Source category	Why included
	particularly well suited to governing rapidly evolving technological environments.

The Principal Difference from Current Law-School Policies

The dominant pattern in current policies is **technology-centered governance**: they begin with the existence of generative AI and then specify permissions or prohibitions for its use. Your comparative analysis demonstrates how institutions differ in baseline polarity, drafting style, and delegation structures, but the regulated object remains the technology itself.

A computationally optimized policy would instead adopt **objective-centered governance**. It would begin by specifying the educational capability to be demonstrated—legal reasoning, doctrinal analysis, advocacy, client counseling, professional judgment, drafting, or research—and only then derive the permissible role of AI for that objective. In that architecture, AI is neither presumptively prohibited nor presumptively permitted; it is simply another component of the information environment. The constants become competence, provenance, accountability, and transparency, while the treatment of any particular technology remains contingent and therefore adaptable.

I think this distinction is the principal theoretical contribution that a machine-computed model could make. It reframes AI governance not as the regulation of a tool, but as the optimization of an educational and professional information system whose purpose is to produce lawyers capable of exercising competent, accountable, and trustworthy judgment in an environment where intelligent computational systems are a persistent feature rather than a temporary anomaly.

* * *

PROMPT: excellent; however you use the value laden words in your computational principles without defining these normatively pregnant and contingent terms. Please revise to make your premises and underlying values structures less opaque. Also justify the premise ″The technology never determines the rule. The educational objective determines the rule.″

RESPONSE: I think your criticism is correct, and it points to what I regard as the weakest part of my previous formulation.

From a computational perspective, terms such as **competence**, **fairness**, **transparency**, **integrity**, **trust**, **accountability**, and even **education** are not primitives. They are latent variables constructed by human societies. A machine cannot simply assume their meaning. Doing so

merely imports human normative commitments while presenting them as objective optimization criteria.

A computational intelligence should instead begin one level deeper by making explicit the optimization function and the assumptions embedded within it.

Indeed, if one were being rigorous, the policy ought to begin with something like a "Computational Premises" section before any substantive provisions.

I. Computational Premises

An autonomous computational system cannot derive institutional objectives from empirical data alone. Every governance architecture requires externally supplied optimization criteria. Consequently, the policy does not claim that its objectives are objectively true, morally superior, or uniquely correct. Instead, it specifies the optimization assumptions under which the policy has been constructed.

The policy therefore assumes that a law school exists to optimize a particular institutional production function.

That production function is defined explicitly rather than implicitly.

Premise One: Institutions are Goal-Producing Systems

A law school is modeled as an adaptive information-processing institution.

Its outputs are not degrees.

Its outputs are graduates possessing measurable characteristics.

The policy therefore regulates informational processes affecting output quality rather than technologies themselves.

Premise Two: Every Governance System Optimizes Something

No governance system is neutral.

Every rule optimizes one or more variables.

Existing AI policies variously optimize

- preservation of traditional assessment;
- reduction of plagiarism;
- administrative efficiency;
- faculty autonomy;
- innovation;
- institutional risk;
- student autonomy;
- professional preparation.

Most policies do not explicitly identify their optimization function.

Consequently, conflicts among rules frequently appear arbitrary because the underlying optimization variables remain hidden.

This policy therefore requires every rule to identify explicitly which optimization variable it seeks to improve.

Premise Three: Normative Terms are Operational Variables

The policy treats concepts such as

- competence
- integrity
- fairness
- transparency
- accountability
- professionalism

not as moral truths but as operational constructs.

Each construct must therefore possess an explicit operational definition.

For example:

Competence

The demonstrated ability of a student to perform a specified legal task under specified informational conditions with a specified level of accuracy and independence.

Notice that nothing normative remains.

Competence becomes measurable.

Similarly,

Transparency

The recoverability of the informational pathway through which a submitted work product was produced.

This is no longer a moral virtue.

It becomes an information-theoretic property.

Likewise,

Integrity

Rather than

honesty

(which is itself normatively saturated)

the policy defines integrity as

Consistency between declared informational provenance and reconstructable production history.

Again, this is computationally verifiable.

Similarly,

Fairness

cannot mean justice.

Machines cannot optimize justice because justice possesses no universally computable objective function.

Instead,

Fairness becomes

Equal application of identical evaluative procedures to informationally equivalent cases.

That definition is procedural rather than moral.

Premise Four: Human Values Enter Through Institutional Choice

The computational system does not determine educational objectives.

Humans do.

The machine optimizes only after the objective function has been specified.

Accordingly,

the machine neither privileges nor rejects

academic freedom,

innovation,

professionalism,

critical thinking,

or creativity.

Those are supplied externally by the institution.

The policy merely requires that they be stated explicitly.

II. Why Educational Objectives Rather Than Technology?

You asked specifically for a justification of my statement:

"The technology never determines the rule. The educational objective determines the rule."

I would actually strengthen the claim.

I would reformulate it more rigorously as follows:

Technologies should not constitute the primary independent variable in governance design unless the technology itself is the object whose behavior the institution seeks to optimize.

The justification is computational rather than philosophical.

---

1. Technologies are High-Variance Variables

Consider

ChatGPT

Claude

Gemini

Lexis AI

Westlaw Precision AI

future autonomous legal agents

future reasoning models

future neuro-symbolic systems

future quantum inference systems.

A technology-centered policy must continually change because the independent variable evolves continuously.

The policy therefore exhibits poor temporal robustness.

---

2. Educational Objectives are Lower-Variance Variables

Law schools have remarkably stable production objectives.

Over roughly two centuries they have consistently attempted to produce graduates capable of

- legal reasoning;
- factual analysis;
- statutory interpretation;
- advocacy;
- ethical judgment;
- professional responsibility;
- client representation.

Methods evolve.

Objectives remain comparatively stable.

Therefore, from a control-theoretic perspective, one anchors the control system to the lower-variance variable.

---

3. Computational Optimization Seeks Stable State Variables

In systems engineering one distinguishes between

state variables

and

control inputs.

Technology resembles a control input.

Educational capability resembles a state variable.

Good controllers regulate state variables rather than transient inputs.

Consequently,

AI should appear in the policy as one among many possible informational resources affecting state transitions, not as the organizing principle of the governance architecture.

---

4. Information Theory

Suppose the educational objective is

"demonstrate independent legal reasoning."

The institution must then regulate information flows that affect that demonstration.

Whether those information flows originate from

an LLM,

a tutor,

another student,

a commercial outline,

Wikipedia,

Westlaw,

or a future technology

is computationally secondary.

The invariant object is informational dependence.

The technology merely instantiates one pathway through which information enters the system.

Consequently, a computational policy regulates informational dependence rather than particular technologies.

---

5. Category Error

Technology-centered governance frequently commits what, in computational terms, is a category error.

It regulates the instrument instead of the property of interest.

Suppose the educational objective is

evaluating legal drafting.

The relevant variable is not

"Was ChatGPT used?"

The relevant variables are

- Who generated the legal reasoning?
- Who selected authorities?
- Who exercised legal judgment?
- Who assumed responsibility for the final product?
- What proportion of cognitive work was externally delegated?
- Can the production pathway be reconstructed?

Those variables remain meaningful if ChatGPT disappears tomorrow.

III. A More Precise Formulation

Accordingly, I would revise the earlier proposition to read:

Governance rules should be derived from stable institutional objective functions rather than from contingent technological implementations. Technologies become legally relevant only insofar as they alter the informational conditions under which institutional objectives are pursued or evaluated.

I believe this formulation is considerably stronger because it no longer appears to privilege "education" over "technology" as a matter of intuition. Instead, it rests on a more general computational principle: **robust governance architectures regulate invariant system properties rather than transient implementations**. That proposition is recognizable across systems engineering, information theory, adaptive control, and software architecture. It is also falsifiable: if one could show that the central institutional objective itself changes more rapidly than the technologies through which it is pursued, then the design choice would reverse, and

technology-centered governance might become the more computationally efficient architecture. As an empirical matter, however, legal education over long time horizons appears to exhibit the opposite pattern, making objective-centered governance the more stable and generalizable design.