The Missing Variable in AI-Assisted Litigation

Architecture and the Quality of Pro Se Access to Justice

Vadym Chernets, PhD[*]

First posted July 2026; this version September 2026

AUTHOR'S DISCLOSURE STATEMENT

The author's interest in this Article is threefold, and each strand is material. First, the author is the plaintiff in Chernets v. Google LLC et al., No. 1:25-cv-05691-RER-RML (E.D.N.Y.), the litigation that serves as the case study. Second, the author is the developer of the multi-model orchestration system whose use is examined here, and is a named applicant on international applications within the patent family described in Part III; the author therefore has an interest in the methods this Article analyzes. Third, the author is the analyst. The methodological constraints set out in Part IV govern the analysis: reliance solely on the public docket, no inference of settlement motives, and non-evidentiary treatment of the system's self-reported performance metrics. Part VII states the limits.

Two paragraphs, one in Part II and one in the Appendix, are the author's own words about why he built the system and why he sued; they are marked as his, and no finding rests on them. Otherwise this Article draws exclusively on the public federal docket and publicly available sources; it does not disclose privileged litigation strategy concerning claims that remain pending against the non-settling defendants. The author is not an attorney. AI assistants were used as instruments under the author's direction; the ideas, research and conclusions are the author's.

ABSTRACT

Nearly every account of AI in litigation rests on an unstated assumption: that AI assistance is one thing. The debate that assumption produces is about volume and fabrication: a federal pro se plaintiff rate up from 11% to nearly 17%, courts describing an "existential threat," a growing docket of sanctions decisions over fabricated citations. What the aggregate studies cannot yet show is whether any configuration of AI assistance changes

[*] AI systems architect. The author's roles as plaintiff, system developer, and analyst are set out in the Author's Disclosure Statement below. Comments welcome: vadimchernets9@gmail.com.

the quality of what self-represented litigants do. The record used here is the public docket of one federal case, and it holds a pattern the volume-and-hallucination account does not predict. A self-represented litigant with no attorney of any kind, using, by the author's account, a multi-model orchestration system (independent models queried in parallel, disagreement surfaced rather than averaged away, citations verified against external sources), answered six coordinated pre-motion letters within four days, filed a substantive opposition six weeks ahead of deadline, and identified a verifiable statutory-citation error in a brief filed for seven defendants represented by counsel from leading firms, in a case naming thirteen technology defendants. The last of those three is the inverse of the failure mode that dominates the literature. The Article proposes the architecture-dependence hypothesis: the architecture of AI assistance, single-model prompting versus multi-model orchestration, is an empirical variable in the quality of self-represented litigation in its own right, not an implementation detail and not the same thing as the availability of AI. The doctrinal consequences taken up here are for the work-product doctrine, where federal courts split in early 2026, and for the unauthorized-practice framework. The case is offered as a falsifiable existence proof and a research agenda; the author is the plaintiff, the system developer, and the analyst.

Keywords: access to justice; pro se litigation; AI-assisted litigation; generative AI; AI hallucination; multi-model orchestration; architecture-dependence hypothesis; blind-consolidation protocol; work-product doctrine; unauthorized practice of law; federal courts. **JEL Classification:** K41 (Litigation Process); K10 (Basic Areas of Law, General); O33 (Technological Change: Choices and Consequences).

Contents

I. Introduction

One litigant, thirteen AI companies: the case examined here compresses into a single docket the question this Article puts to the literature. The gap between the civil legal needs of Americans and the resources available to meet them is among the most durable problems in the U.S. justice system. The Legal Services Corporation's most recent national study found that low-income Americans received no or inadequate help for ninety-two percent of the civil legal problems that substantially affected them, with nearly three in four low-income households experiencing at least one such problem in a single year.[1] For most self-represented litigants the barrier is not the filing fee but everything after it: locating the correct court, identifying the operative cause of action, effecting service on corporate defendants, and answering motions drafted by experienced counsel under compressed deadlines.

That population has grown quickly since generative AI became widely available. A recent large-scale empirical study reports that the federal civil pro se plaintiff rate, stable near eleven percent for two decades, rose to 16.8 percent in fiscal year 2025 (Figure 1), and that the share of sampled federal civil complaints bearing markers of AI-generated text climbed from virtually zero in 2019 to more than eighteen percent by early 2026, as measured in that study. That sample draws on complaints filed both pro se and with counsel, taken from the RECAP archive, which the authors caution underrepresents filings made without a lawyer.[2]

Courts and commentators have reacted mostly with apprehension. The then-chief judge of a federal district court described the overall problem as "an existential threat to the federal courts."[3]Bar associations warn that consumer chatbots may cross into the unauthorized practice of law, and a growing body of sanctions decisions addresses litigants, represented and unrepresented alike, who file briefs citing cases that do not exist.

---

[1]Legal Services Corporation, The Justice Gap: The Unmet Civil Legal Needs of Low-Income Americans (2022), https://justicegap.lsc.gov. On the scale of self-representation generally, at least one party was self-represented in more than three-quarters of civil cases in a leading multi-jurisdiction study. See Paula Hannaford-Agor, Scott Graves & Shelley Spacek Miller, The Landscape of Civil Litigation in State Courts iv (National Center for State Courts 2015).

[2]Anand V. Shah & Joshua Y. Levy, Access to Justice in the Age of AI: Evidence from U.S. Federal Courts (Mar. 20, 2026) (unpublished manuscript), https://ssrn.com/abstract=6766859 (analyzing over 4.5 million non-prisoner federal civil cases, FY2005–FY2026, and 46 million matched PACER docket entries); see also Or Cohen-Sasson, Stochastic Justice: Legal Inconsistency by Humans and AI, 105 Neb. L. Rev. (forthcoming 2026).

[3]Chief Judge Patrick J. Schiltz (D. Minn.), quoted in Mattathias Schwartz & Zach Montague, Artificial Intelligence Floods Court Dockets with Home-Brewed Lawsuits, N.Y. Times (May 25, 2026).

The apprehension is well founded, but it describes a particular mode of AI use. By focusing on that mode, existing empirical work implicitly treats AI assistance as homogeneous, and this Article argues that the assumption is no longer defensible. The documented failures (hallucinated authority, sycophantic agreement, undifferentiated volume) are characteristic of single-model prompting: a litigant poses a question to one general-purpose model and files the result. Almost the entire empirical picture of pro se AI use rests on that mode.

The architecture examined here is a different one, and the question is whether that difference matters. In multi-model orchestration, a query is dispatched simultaneously to several independent models, and their outputs are compared. Agreement and disagreement are made explicit, and divergence, including a minority signal that flags a possible error, is preserved rather than averaged away. The claim advanced here is narrow but, if correct, consequential: the single-model failure modes that animate current concern may not be intrinsic to AI-assisted litigation, and an orchestration architecture may convert AI use from a source of volume into an instrument of quality. The contribution is methodological: the Article proposes that the architecture of AI assistance be treated as a distinct empirical variable rather than an implementation detail, and the case that follows is offered as one documented instance motivating that proposal. The claim, in four words: architecture, not availability.

The claim is developed through a single case study, *Chernets v. Google LLC et al.*, a federal action in which one self-represented plaintiff, using (by his own account) a multi-model orchestration system, litigated against a group of technology defendants represented by counsel from numerous law firms. The analysis rests entirely on the public docket: filings, dates, and document numbers that any reader can verify. That record shows conduct the prevailing volume-and-hallucination account does not describe and that aggregate studies of AI-assisted filing are not designed to detect: a full briefing cycle sustained against a coordinated defense, with six coordinated pre-motion letters answered within four days and a substantive opposition filed weeks ahead of deadline, and the identification of a verifiable statutory-citation error in the defense coalition's brief. Four defendants settled early (on confidential terms; Part V.E). The case remains pending, and the purpose of examining it is to use a verifiable record to isolate a phenomenon that, on the prevailing account, would not be expected to exist and that the literature has not yet examined closely.

A single case authored by a party can establish that a phenomenon occurred, not how often it occurs or whether it generalizes; Part VII states the limits. The argument is that the architecture of AI assistance is an independent variable in the study of AI-assisted litigation, a variable that current framing, focused on availability and volume, does not isolate. That is the architecture-dependence hypothesis; Part IV states it formally.

The Article proceeds as follows. Part II situates the study within the access-to-justice literature and the emerging scholarship on pro se AI use. Part III distinguishes single-model prompting from orchestration and specifies the mechanism. Part IV sets out the case-study method and its evidentiary basis in the public docket. Part V presents the findings. Part VI examines implications for the work-product doctrine and the unauthorized-practice-of-law framework. Part VII addresses limitations and future work. Part VIII concludes.

II. The Access-to-Justice Gap and the Pro Se AI Turn

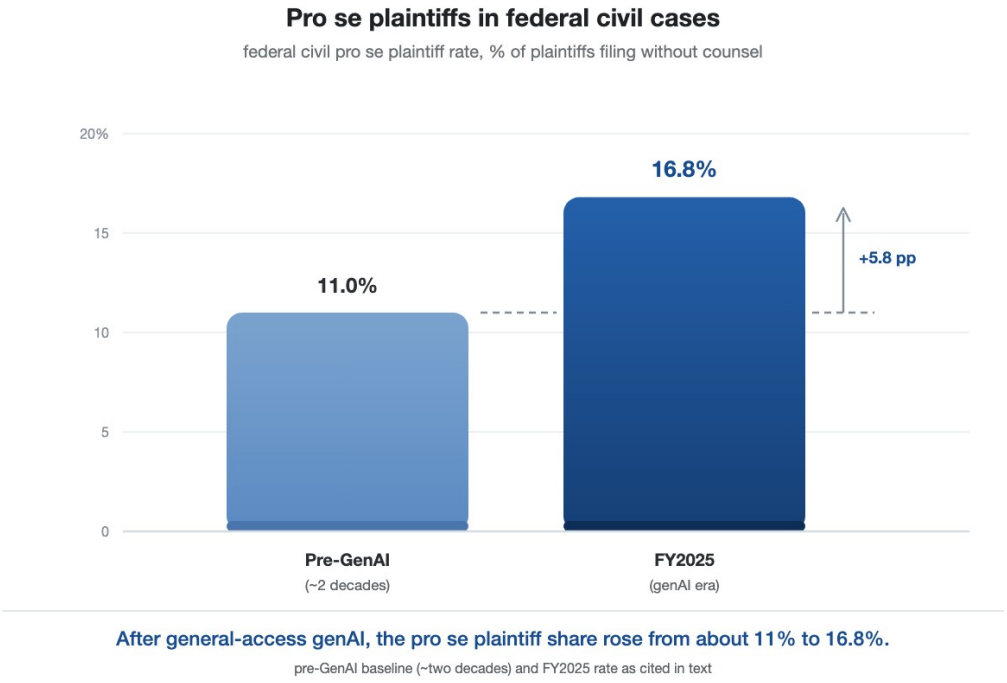
The justice gap is a structural feature of the U.S. civil system, not a marginal one. A large share of low- and middle-income Americans receive no meaningful legal help for their civil problems, and in many categories of case self-representation is the norm rather than the exception (Carpenter, Shanahan, Steinberg & Mark 2022). Since the comparative access-to-justice movement of the 1970s (Cappelletti & Garth 1978), scholars have framed the question as one of distribution: legal expertise is a scarce resource, unevenly allocated, and the litigants least able to afford counsel are often those with the most at stake (Sandefur 2019). That literature asks what access to justice is access to; the question pursued here is the quality of that something.

Against that background, generative AI arrived as an ambivalent development. On one view, it promised to democratize legal capability, giving unrepresented litigants a facsimile of the research, drafting, and analysis that counsel provides. That promise took concrete form when a frontier model passed the Uniform Bar Examination (Katz, Bommarito, Gao & Arredondo 2024), though performance benchmarks of that kind changed expectations about capability without isolating architectural differences. Simshaw cautioned early that AI could instead entrench a two-tiered system, in which the well-resourced obtain expert human judgment while everyone else receives automated approximations.[4] The empirical picture that has since emerged is more

[4]Drew Simshaw, Access to A.I. Justice: Avoiding an Inequitable Two-Tiered System of Legal Services, 24 Yale J.L. & Tech. 150 (2022).

concrete: analyzing millions of federal filings, recent work documents that the pro se plaintiff rate rose sharply following the public release of generative AI and develops indicators of AI-consistent drafting in the complaints themselves. That aggregate record cannot yet show whether any of this leaves self-represented litigants better off. As that study's most careful readers observe, getting new plaintiffs past the courthouse door is not the same as helping them once inside (Engstrom & Caspi 2026).

Figure 1. *The rise in federal civil pro se plaintiffs following broad availability of generative AI.*



Source: Adapted from Shah & Levy (2026); figures rounded.

The reception in the courts and the practicing bar has been predominantly cautionary, and the concerns cluster around fabrication, unauthorized practice, and volume. On fabrication, a now-substantial line of decisions sanctions litigants and lawyers who file briefs citing cases that do not exist,[5] a failure mode traced directly to generative models that produce plausible but fictitious authority. That failure mode has produced a tracked docket of decisions in which a court found, or took as established, that a filing relied on AI-fabricated material: more than two thousand worldwide by September 2026 (Charlotin 2026), and that now includes a widely reported episode in which one of Wall Street's most prominent firms apologized to a federal bankruptcy court for

[5]See, e.g., Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023); Park v. Kim, 91 F.4th 610 (2d Cir. 2024).

an emergency filing containing inaccurate citations and other errors, some of them AI "hallucinations."[6] The composition of that docket is itself instructive: what began overwhelmingly as a self-represented phenomenon has not remained one. By mid-2026, licensed counsel accounted for roughly two-fifths of the documented incidents, and their share of newly documented findings was growing.[7] Bar associations and commentators worry as well about unauthorized practice, questioning whether consumer chatbots dispensing legal guidance cross the line reserving the practice of law to licensed attorneys, and whether the entities operating them might bear responsibility.[8] As for volume, judges have described an influx of AI-assisted pro se filings as a strain on judicial resources, in the sharpest framing an existential threat to the federal courts.

The judicial response has taken a further, revealing form. Since mid-2023, individual federal and state judges have issued standing orders addressing generative AI in their courtrooms (public trackers now list hundreds, no two alike), requiring parties to disclose whether AI was used in drafting, to certify that citations were verified, or both.[9] The orders share a design (Table 1): they regulate the fact of AI use and the verification of its output. Architecture, whether the assistance is single-model or orchestrated, self-verifying or not, is the variable isolated here, and it appears in none of the orders the author has located. A court applying such an order to the case studied here would record only that AI was used. Neither the disclosure regime nor the empirical literature currently has a vocabulary for how, and that absence is the same homogeneity assumption described above, in regulatory form.

Table 1. The emerging AI-disclosure regime: what standing orders ask, and what they do not.

---

[6] Letter from Andrew J. Dietderich, Sullivan & Cromwell LLP, to Chief Judge Martin Glenn, U.S. Bankruptcy Court, S.D.N.Y. (Apr. 18, 2026) (apologizing for "inaccurate citations and other errors," some of them AI "hallucinations," in an April 9 emergency motion in the Chapter 15 proceedings of Prince Global Holdings Ltd.); see Sullivan & Cromwell Apologizes to Judge for AI Hallucinations, Bloomberg Law (Apr. 21, 2026).

[7] AI Hallucination Cases Database (Damien Charlotin), https://www.damiencharlotin.com/hallucinations/ (snapshot of July 2, 2026): of 1,668 documented incidents worldwide, 975 involved self-represented litigants and 653 involved represented counsel; in 2023, roughly seven in ten involved self-represented litigants, and counsel's share of newly documented findings rose steadily through 2025–2026.

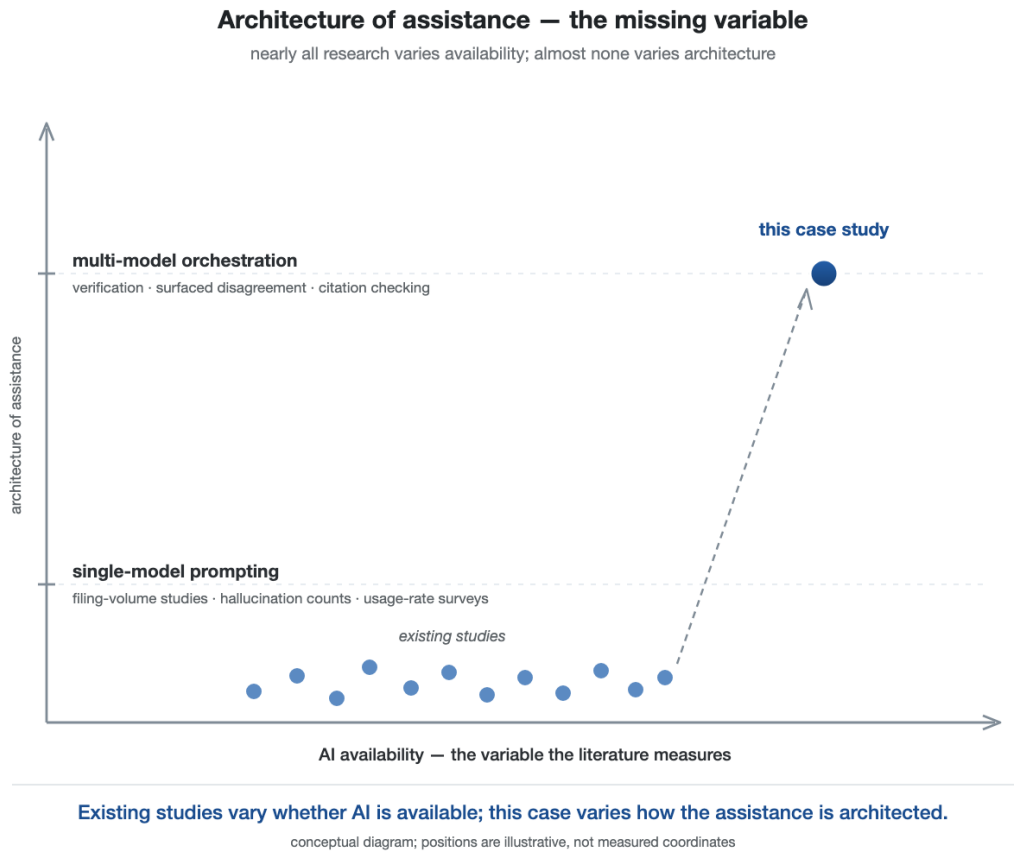
[8] New York State Bar Association, Pro Se Advocacy in the AI Era: Benefits, Challenges, and Ethical Implications (2026); see also commentary on AI and the unauthorized practice of law in the Georgetown Journal of Legal Ethics (2026).

[9] See Law360 Pulse, Tracking Federal Judge Orders on Artificial Intelligence (last visited July 2026); Legal AI Governance, Federal and State Court Orders on AI, https://legalaigovernance.com/tracker/court-orders/ (last visited July 2026).

What the order asks	Presence in the regime
Was generative AI used in drafting?	Routinely required
Were citations and quotations verified?	Routinely required
Which tool or model was used?	Sometimes required
Was assistance single-model or orchestrated (architecture)?	Not found in any located order
Was disagreement among models surfaced and resolved?	Not found in any located order

Source: Public trackers of judicial standing orders on AI (Law360 Pulse; Legal AI Governance), as of July 2026; see note 9. Characterization of coverage is the author's.

Figure 2. The missing variable: existing empirical studies vary the availability of AI assistance; its architecture goes unmeasured.



Source: Author's schematic of the empirical literature discussed in this Part.

The two developments meet in the question pursued here. The justice gap explains why self-representation matters: for most litigants the alternative to self-representation is no representation at all. The recent increase in self-representation, concurrent with the broad availability of

generative AI, explains why the *architecture* of AI assistance has become an important research question rather than an incidental detail. If unrepresented litigants are now acting through AI at scale,[10] then how that AI is configured may shape the quality of what reaches the courts, and may in principle narrow the gap in problem-solving capacity between individuals and the institutions they face.

One further framing changes what orchestration is understood to be for. For a user with a motor, spinal, or visual impairment who cannot operate several browser interfaces at once, the manual workflow that able-bodied users find merely tedious (open each provider, paste the query, read, compare) is not tedious but unavailable. Unified multi-model access is not a productivity preference but a reasonable-modification question (cf. Blanck 2014, on web accessibility for persons with cognitive disabilities; the extension to motor, spinal and visual impairments is the author's own). Blocking the assistive automation of a workflow a person could otherwise perform manually then imposes a burden that falls, by construction, most heavily on the users least able to perform it manually, and for that user architecture turns a debate about unauthorized practice into a question of reasonable modification.[11]

I built it for someone in my family who cannot do that by hand — who cannot crawl between browser tabs. A crutch, a guide dog: something does the part a person cannot do. Say it once, and it is done. That is what this case is about.

What these concerns share, and what this Article seeks to make visible, is an implicit model of how a pro se litigant uses AI: one person, one general-purpose model, minimal verification, maximal output. It is not the only architecture available, though, and the empirical literature (Figure 2) has not yet examined cases built on a different one. The scholarship measures the quantity of AI-assisted pro se litigation and the reliability of its most common form; it has largely not asked whether a different architecture of assistance produces a different quality of litigation

[10] See How Courts Are Coping with a Flood of AI-Generated Lawsuits, MIT Technology Review (June 4, 2026) (reporting, inter alia, that filings by self-represented parties in Vermont rose from roughly forty-five a year before 2022 to more than 1,100 in 2024).

[11] The complaint in the case study characterizes the system in these terms; the framing is noted as the plaintiff's legal theory, not as a merits determination. On the scale of the affected population, more than one in four U.S. adults — over 70 million — reported a disability in 2022, CDC, Disability Impacts All of Us (2022 BRFSS data), and the employment-population ratio for people with a disability was 22.8% in 2025, versus 65.2% for those without, U.S. Bureau of Labor Statistics, People with a Disability: Labor Force Characteristics — 2025, USDL-26-0364 (Mar. 3, 2026).

conduct. The gap is addressed here by supplying a closely documented case of a mode of use the aggregate studies do not isolate.

III. From Single-Model Prompting to Orchestration: The Cognitive Exoskeleton

The term "cognitive exoskeleton" marks a distinction the legal-AI debate tends to collapse. An exoskeleton amplifies the person inside it; it does not act in the person's place, and it is inert without the person. That distinction between augmentation and substitution runs through both doctrinal questions taken up in Part VI. Work-product protection attaches naturally to augmentation, because the strategy protected remains the litigant's own, and awkwardly to substitution. The unauthorized-practice framework is aimed at substitution (an entity practicing law for someone) and grips poorly on a tool through which a person practices for himself. Whether a given system is one or the other is a question of architecture and use, not of labels. The concerns canvassed in Part II share an implicit architecture. In the paradigmatic problem case, a self-represented litigant poses a question to a single general-purpose model and submits its output. The pathologies that follow are, in substantial part, properties of that single-model configuration: fabricated citations, confident misstatements of law, sycophantic agreement with the user's premise (Sharma et al. 2024). A lone model has no internal adversary, nothing in the pipeline positioned to contradict it. Its errors, however confident, pass through unchallenged. The single-model failure modes that dominate the literature are, in this sense, not properties of artificial intelligence; they are properties of a configuration.

For decades, AI-and-law research pursued reliability through explicit knowledge engineering, encoding legal reasoning in rules and case-based models (Ashley 2017), before the machine-learning turn shifted the question from encoding rules to learning patterns (Surden 2014). A distinct body of computer-science research suggests that this configuration is not the only one available, and that its failure modes are partly architectural rather than intrinsic. In the approach commonly termed multi-agent debate, several model instances independently generate answers and then critique one another's reasoning across successive rounds before converging on a response. The foundational study reports that this "society of minds" procedure improves factual validity and reduces the fallacious answers and hallucinations to which single models are prone.[12]

[12]Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum & Igor Mordatch, Improving Factuality and Reasoning in Language Models through Multiagent Debate, Proceedings of the 41st International Conference on

Subsequent work has extended and qualified the finding, reporting that structured disagreement reduces hallucination across mathematics, logic, and question-answering benchmarks, while also identifying limits, such as diminishing returns beyond a handful of agents and rounds, and the need to assess the quality of intermediate argumentation rather than final answers alone.[13] The benchmark domains matter here. That literature validates structured disagreement largely on mathematics, logic, and question-answering benchmarks, domains with verifiable answers. Whether its gains transfer to legal argument, where many questions lack a single checkable answer and models may converge politely on a shared error, is untested, and the transfer is an assumption of the design examined here rather than a demonstrated result.

For the present analysis, the part of the mechanism that matters is where the reliability gain originates. The hypothesis is that the gain comes from the relationships among models rather than from any single model being more capable, from disagreement being surfaced and adjudicated rather than suppressed. The value of orchestration is not more answers but better-supported ones. A fact that one model asserts and another contests is the fact most in need of scrutiny, and a debate procedure routes attention to it, whereas single-model prompting cannot. The mechanism has recognizable relatives in other fields (ensemble methods in machine learning, error-correcting codes in information theory, structured dissent in the science of group deliberation), which suggests that it is a general property of redundant, disagreement-preserving systems rather than an artifact of any one product.

Table 2. *Single-model prompting versus multi-model orchestration, by design property.*

Property	Single-model prompting	Multi-model orchestration
Internal adversary	None; output unchallenged	Models cross-examine one another
Source of citations	Generated from model weights	Validated against external databases
Handling of disagreement	Not surfaced	Preserved; can trigger review
Characteristic failure mode	Fabrication, sycophancy	Reduced fabrication (per debate literature)

Source: Design properties per the multi-agent debate literature (e.g., Du et al. 2024; Liang et al. 2024; Chan et al. 2023) and, for the orchestration column, the system documentation.

Machine Learning (ICML 2024), arXiv:2305.14325.

[13] See, e.g., Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi & Zhaopeng Tu, Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024) 17889; Chi-Min Chan et al., ChatEval: Towards Better LLM-Based Evaluators through Multi-Agent Debate, arXiv:2308.07201 (2023).

The case study has a reflexive structure that, to the author's knowledge, no other litigation yet shares: the technology in dispute is the technology by which the dispute is conducted. Several of the defendants develop models that systems of this kind query, the orchestration system at issue coordinates those same publicly available models, and the litigation conduct examined in Part V occurred while that coordination was in use. On the hypothesis advanced here, no individual model conferred the capability. Each carries the single-model failure modes described above, and each defendant that develops a constituent model had, by construction, access to at most its own constituent model. The configuration that mattered was a relational one: heterogeneous models from adverse providers, queried in parallel, checking one another. The reflexivity does analytical work because it holds the component technology constant: whatever differences the record reveals cannot be attributed solely to the litigant having access to better component models than the defendants, since several of the component models were the defendants' own; configuration, access, and operator skill remain open explanations. This line of research is also not external to the defendants. Major AI laboratories have themselves published on the advantages of multi-agent and ensemble configurations over single models, and during the period studied moved to commercialize cross-model verification in their own products.[14] The architecture at issue is one the industry itself increasingly endorses.

Orchestration generalizes multi-agent debate into an operational system (Figure 3). Where debate typically runs multiple instances of one model, orchestration dispatches a query across heterogeneous models from different providers, normalizes their outputs, computes a consensus, and, critically, preserves divergence as a first-class signal rather than averaging it into a single answer. The orchestration system used in the case study is one implementation of this pattern: per its technical documentation, it incorporates heterogeneous model querying, consensus formation, divergence preservation, and citation verification against external sources.[15]

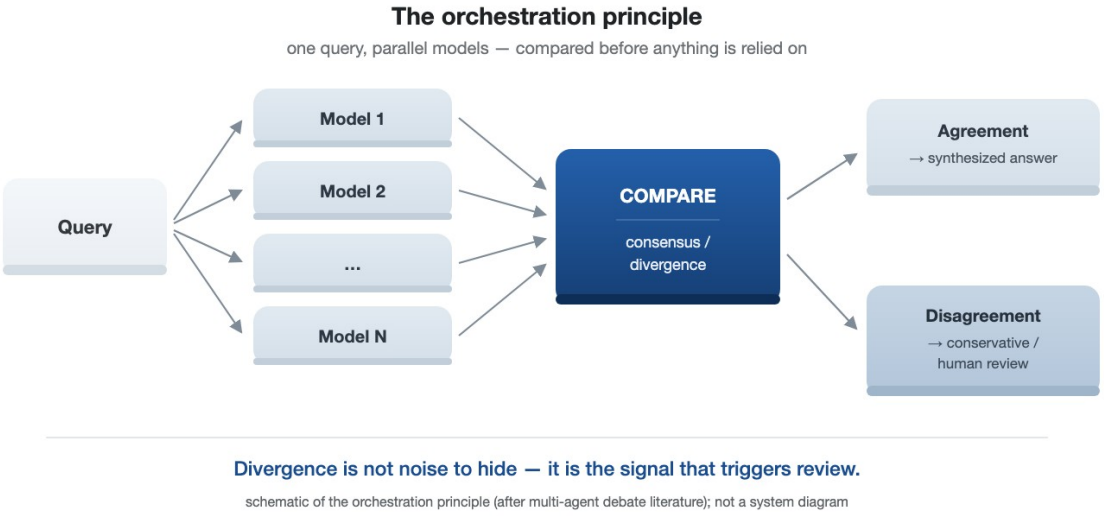
The orchestration architecture at issue is a specified, engineered design. It is the subject of a family of pending patent applications, including several international applications under the Patent

[14] The plaintiff's opposition placed one such development into the record, noting one defendant's public announcement of a cross-provider, multi-model consensus feature. See Plaintiff's Opposition, ECF No. 76.
[15] System technical documentation (cited for design only).

Cooperation Treaty filed in 2025 and 2026.[16]

The two configurations differ along several design axes, summarized in Table 2. Two of those commitments are analytically significant. One is citation validation against external sources. The dominant single-model failure in litigation, the hallucinated case, arises because a model generates citations from learned patterns rather than retrieving them, and an architecture that verifies each citation against a live database, flagging rather than including anything it cannot confirm, targets that failure at its source. The other is preservation of divergence. Where models disagree on a legal interpretation, an orchestration system can be configured to treat disagreement as a trigger for conservatism and human review rather than as noise to be resolved by majority vote. In high-stakes domains, disagreement among independent systems may itself constitute actionable information rather than statistical noise: a minority signal indicating a possible error is retained and elevated. It is not discarded.

Figure 3. *The orchestration principle: parallel query, comparison, and preservation of divergence.*



Source: Schematic of the general principle after the multi-agent debate literature; not a diagram of any specific product architecture.

The structural point is that the specific pathologies that have made courts wary of pro se AI use are characteristic of single-model prompting, and that an orchestration architecture is directed

[16] The architecture is the subject of a family of pending international patent applications on which the author is a named applicant; the family is documented in the public record of the litigation (see Plaintiff's Opposition, Ex. 3), and its identifiers are omitted here as immaterial to the analysis.

at those pathologies. The documented failures are properties of a configuration. They are not properties of AI assistance as such. In this sense the architecture shifts the source of reliability from an isolated model to relations among models: from isolated intelligence to coordinated intelligence. Nor is orchestration exotic: commercial routing and multi-model products increasingly expose these same design properties, which makes the absence of architecture from the empirical legal literature all the more conspicuous. Whether that architectural difference produces a difference in observable litigation conduct is an empirical question, and the remainder of the Article turns to one setting in which it can be examined against a verifiable record.

IV. Method and Evidentiary Basis

The design is a single-case study, a choice that reflects the nature of the phenomenon: multi-model orchestration in live federal litigation is at present rare, and the objective here is not to estimate its frequency but to establish, against a verifiable record, that a particular pattern of conduct occurred and to characterize it. Quality, as used throughout, refers to observable litigation conduct verifiable from the docket: deadline adherence, citation accuracy, procedural correctness, responsiveness to the record, and the coherence of filed argument. It does not refer to outcomes, which no single case can warrant. A case of this kind repays attention: real-world adversarial litigation, with deadlines, opponents, and consequences, may offer a more meaningful setting for observing AI-assisted human reasoning than laboratory evaluation alone. Part VII returns to this limitation. Of these dimensions, the findings in Part V bear chiefly on deadline adherence, procedural correctness, and citation accuracy; the coherence of filed argument is documented in the record rather than independently scored, and no finding depends on the outcome of any pending motion.

The evidentiary basis is independently retrievable: every filing cited in this Article is identified by its ECF number and can be obtained by any reader directly from PACER under the case numbers below, without reliance on the author. It consists of the public federal docket in *Chernets v. Google LLC et al.*, No. 1:25-cv-05691-RER-RML (E.D.N.Y.), together with the docket of the earlier action, No. 1:25-cv-07770 (S.D.N.Y.), which named a different defendant. The structure of the case can be stated in figures, all docket-verifiable: of thirteen named defendants, ten appeared; seven joined a joint motion to dismiss; four settled and were dismissed with prejudice; and six were continuing to litigate as of June 2026 (Table 3). Three of the thirteen

never appeared, and the record shows service and entity-identification difficulties concerning those defendants; it does not establish why they did not appear. One foreign defendant was transmitted for service through China's Central Authority under the Hague Service Convention. Nothing came back: no confirmation, no refusal. One domestic defendant took four attempts across several states before the summons reached its registered agent, confirmed by a process server under oath. No appearance followed. A third invoked its own corporate structure to contest which entity had been named. Corporate structure and cross-border procedure can operate, in practice, as a procedural shield, and working through it is a task a represented party typically hands to specialized counsel or vendors, which here the self-represented plaintiff did himself, for all thirteen. All factual assertions in Part V concerning litigation conduct are drawn from filed documents, identified by docket number and date, that any reader may independently retrieve, except the litigant's mode of AI assistance, which a docket does not record. Where the Article refers to statements made in court, it relies on the official transcript; where it refers to the parties' arguments, it relies on the filings themselves (the defendants' joint motion and supporting memorandum, and the plaintiff's opposition) rather than on any characterization by a party.[17]

Three methodological constraints follow from the author's role, disclosed at the outset. First, the analysis is confined to the public record; it does not draw on the author's privileged knowledge of litigation strategy, and it does not disclose work product concerning claims that remain pending against the non-settling defendants. Second, the Article distinguishes throughout what the docket establishes (dates, filings, rulings, the text of documents) from what it does not: the parties' subjective motives, and in particular the reasons for any settlement, whose terms are confidential. Third, claims made by the orchestration system's own documentation about its performance are treated as self-reported and are not relied upon as evidence of efficacy; the findings rest on litigation conduct visible in the record, not on the system's account of itself.

One asymmetry in the evidentiary basis qualifies the method just described. The dependent variable of this study, litigation conduct, is docket-verifiable in full: every filing, date, and document analyzed in Part V can be retrieved by any reader. The independent variable is not retrievable in that way. The public record establishes that the orchestration system exists and documents its architecture: the operative complaint characterizes the system, the patent

[17]Joint Motion to Dismiss and Memorandum in Support, Chernets v. Google LLC, No. 1:25-cv-05691-RER-RML (E.D.N.Y.), ECF Nos. 69, 69-1; Plaintiff's Memorandum of Law in Opposition, ECF No. 76 (Apr. 2, 2026); Transcript of Pre-Motion Conference, ECF No. 55 (Jan. 27, 2026).

applications specify the design, and the opposition's technical exhibits describe its orchestration features. What the docket cannot show is which mode of AI assistance produced any particular filing, or when the litigant's working method changed. Those facts rest on the author's own account of his process, and the Article marks them as such where they appear. The study therefore pairs a verifiable record of outputs with a self-reported record of inputs, and the controlled designs proposed in Part VII are constructed so that the architecture variable is observed rather than reported; the Appendix specifies a run manifest for that purpose.

The docket establishes that a self-represented litigant using an orchestration system produced particular filings on a particular schedule. It cannot establish that the system, rather than the litigant's own capabilities, produced their observed characteristics. The Article therefore describes the conduct and its architectural context and treats the causal question as open, to be addressed by the controlled study proposed in Part VII.

Stated formally, the hypothesis advanced here is the following:

> *Architecture-dependence hypothesis.* Holding the availability of generative AI constant, the architecture of AI assistance (single-model prompting versus multi-model orchestration) functions as a distinct empirical variable (Figure 4): assistance mode is associated with measurable differences in observable litigation conduct, including deadline adherence, citation accuracy, procedural correctness, and responsiveness to the record. The hypothesis is falsifiable: a systematic comparison of pro se filings coded by mode of assistance that found no such association would refute it.

Figure 4. The architecture-dependence hypothesis as a chain of claims. Each link is a claim to be tested, not a finding.

Source: Author's schematic of the hypothesis stated in the text; dashed arrows mark hypothesized, untested links.

V. Findings

Several features of the litigation bear on the question posed here, and Figure 5 sets them in sequence.

The disparity of legal resources in this case is not incidental to the study; it is the condition that gives the study its point. On one side was a single self-represented litigant with no legal training and no attorney of any kind, not even limited-scope counsel. On the other was a defense in which lawyers from numerous leading firms combined into a common front, filing a joint motion to dismiss on behalf of seven defendants and coordinating their submissions among themselves. A configuration in which one unrepresented individual sustains a full briefing cycle against that

alignment, meeting every deadline, answering on the merits, and identifying an error the aligned firms did not catch, is close to what the access-to-justice literature treats as not happening. The remainder of this Part documents it, on a public docket; what may account for it is the question the Article as a whole pursues.

Figure 5. *Procedural chronology of the case.*

Source: Public docket, Chernets v. Google LLC, No. 1:25-cv-05691-RER-RML (E.D.N.Y.), and the related earlier action.

A. Reconstruction of the legal framework following an initial defective filing.

The litigant's first action, filed in September 2025 and a matter of public record, was procedurally defective in several independent respects: it was brought in the name of a limited-liability company, which cannot proceed without counsel; against the wrong defendant; in the wrong district; and under the wrong title of the ADA. The court closed it at the threshold, declining to issue a summons. A second action, filed personally and in the correct district roughly three

weeks later, reframed the claim under the appropriate provision of the statute, named the proper defendants, and was accompanied by a substantially expanded evidentiary submission. One change in method accompanied the correction. By the author's own account, through the first filing the litigant had worked with a single model on a consumer device, and the reconstruction coincided with the first use of full multi-model orchestration through direct model access. The docket cannot confirm that sequence, and it is not offered as cause. What the docket does show is the correction itself, the first episode this Article examines: the record reflects a rapid transition from a filing dismissed at the threshold to one that proceeded to service and prompted appearances by multiple represented defendants. The reconstruction is the kind of broad, cross-checked legal-framework revision an orchestration workflow, querying multiple models on forum, cause of action, and pleading standard, is designed to support.

B. Response to coordinated pre-motion practice.

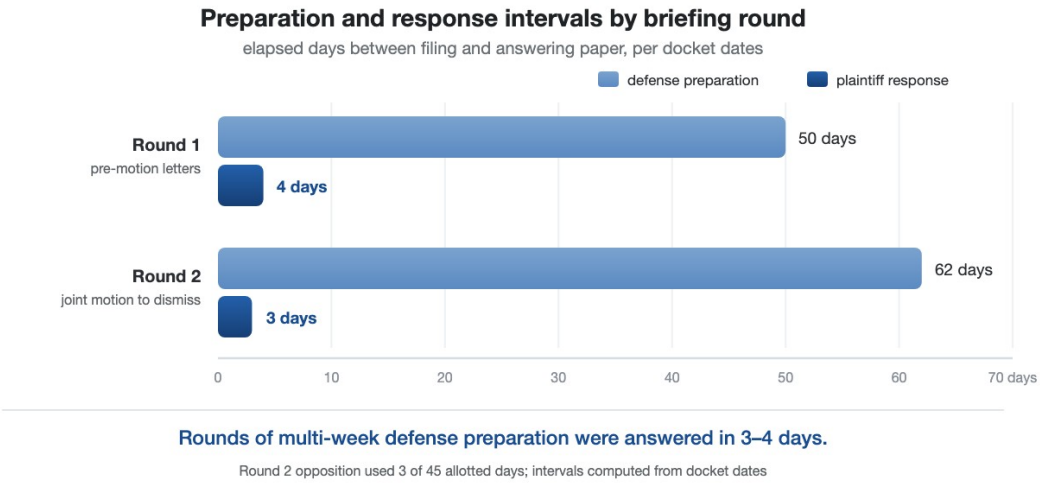
On a single evening, January 12, 2026, six defendants filed six pre-motion letters within a span of roughly four and a half hours, having jointly moved for an extension so that they could file together.[18] Under the court's individual practices, each letter required a response from the plaintiff within five business days. Six responsive memoranda, each addressed to the arguments of a specific defendant, were filed within four days.[19] The episode bears on more than speed: a response to a coordinated, multi-front filing cannot be finalized in advance, because its addressing is fixed by six documents that did not exist until the evening they were filed. But the arguments available to defendants in a case of this type are drawn from a predictable repertoire: standing, the applicability of Title III to digital services, failure to state a claim. A diligent party can prepare modular responses to anticipated arguments and adapt them once the letters arrive. Anticipatory drafting is therefore constrained by this episode rather than foreclosed by it: what the four-day interval establishes is the assembly, adaptation, and filing of six individually addressed memoranda against six distinct submissions. What it cannot establish is how much of their content predated the letters.

[18]ECF Nos. 28 (consent motion), 30, 32–34, 36–37 (Jan. 12, 2026). The letters coordinated among themselves; one defendant expressly adopted the grounds and arguments set out in its co-defendants' pre-motion letters.
[19]ECF Nos. 39–44 (Jan. 15–16, 2026).

C. Timely response to a consolidated defense.

Following a pre-motion conference at which the court invited the defendants to reduce duplicative briefing, seven defendants, including all six that had filed pre-motion letters, filed a joint motion to dismiss with a supporting memorandum. Under the court's schedule, the plaintiff's opposition was due approximately forty-five days later. The opposition was dated April 2, 2026 (three days after the motion's March 30 filing) and entered on the docket the following day, roughly six weeks ahead of the May 14 response deadline. It was a twenty-eight-page memorandum accompanied by a declaration and voluminous exhibits. Figure 6 sets the response intervals against the defense's preparation intervals for both briefing rounds. The observation is the fact of a substantive, timely, and proportionate response, not its merits, filed by a self-represented litigant against a coordinated defense represented by counsel from multiple firms. The relevant point is architectural rather than personal: responding on the merits to several distinct briefs at once is the kind of parallel, multi-thread task a multi-model workflow is structured to distribute, in contrast to the single-thread prompting the literature assumes. The timing matters for something besides compliance: a motion that sits unanswered on a public docket for a full response period presents, to anyone reading the record during that interval, only the movants' account. An opposition entered within days collapses that interval. An asymmetry of that kind ordinarily runs against self-represented parties, who cannot usually compress the interval.

Figure 6. *Preparation and response intervals, by briefing round.*



Source: Public docket; elapsed days between the operative filing and the date of the responsive filing in each round (document dates; docket entry may follow by one day). The defense interval in the first round is measured from completion of service

(November 2025) to the coordinated pre-motion letters of January 12, 2026. Elapsed intervals bound preparation time from above; they do not distinguish drafting time from strategic timing.

D. Identification of a statutory-citation error in the defense memorandum.

The plaintiff's opposition identified an error in the statutory authority relied upon in the joint motion. In addressing the "full and equal enjoyment" element, the defense brief cited 42 U.S.C. § 12184 (ECF No. 69-1, at 15), the provision governing specified public transportation provided by private entities, such as buses, taxis, and limousines, whereas the claim arises under 42 U.S.C. § 12182, the public-accommodations provision (Plaintiff's Opposition, ECF No. 76, at 17).[20] The paradigmatic concern about pro se AI use is the fabricated citation: authority that does not exist, generated by a model and filed by a litigant. Here, the verifiable citation error appears in a submission filed by counsel, and it was identified in a submission filed by a self-represented litigant using an orchestration system whose documented design includes citation validation against external sources. Orchestration as defined in Part III bundles two mechanisms, cross-model disagreement and validation against external sources, and a transposed section number is the kind of defect the second is built to catch. Multi-model debate is not required for it. By the author's account the error was surfaced by the workflow rather than by the litigant: one model flagged the mismatch between the section cited and the provision the claim arises under, the consolidating model checked it and confirmed it, and the critic checked it again. The litigant is not a lawyer and could not have found the error or verified the correction himself. The property implicated is the verification component, not the bundle as a whole. The direction of the correction runs opposite to the pattern the literature leads one to expect; its weight in the motion's disposition is for the court. The error is independently verifiable by any reader: the two provisions lie two sections apart in the same subchapter of the U.S. Code, and the defense memorandum's citation of § 12184 is visible on the face of the page. The two provisions carry parallel operative language, each guaranteeing "full and equal enjoyment" within its domain, and the substitution is therefore the kind a reader absorbs without noticing at reading speed and cannot miss when every citation is checked against its source. What distinguishes them is coverage: § 12184 governs specified public transportation provided by private entities, not places of public accommodation. The plaintiff's opposition pressed the miscitation as "not merely clerical" because it appeared in the memorandum's treatment of a dispositive element; a transposed section number is not fabricated authority.

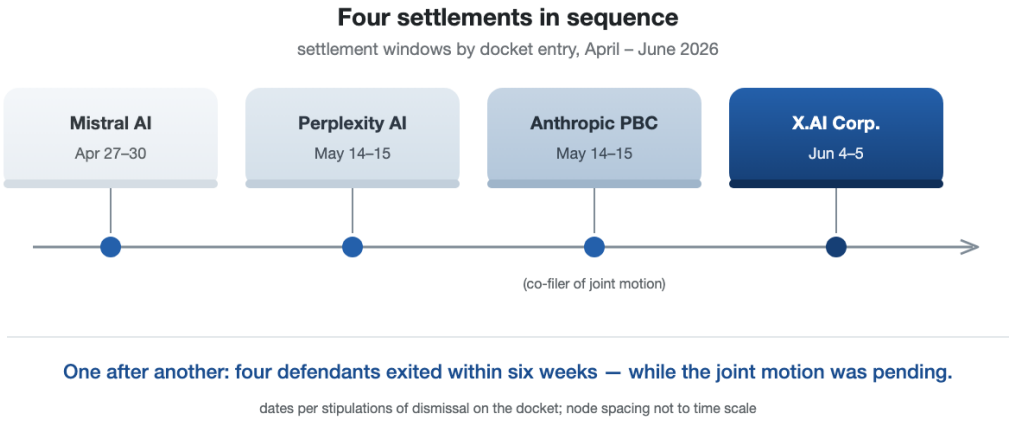
[20]Compare Memorandum in Support of Joint Motion to Dismiss, ECF No. 69-1, with Plaintiff's Opposition, ECF No. 76, at 17.

Detecting a misplaced statutory citation in an opposing brief is the inverse of the failure mode that dominates the literature on pro se AI use, namely generating misplaced citations in one's own.

Attribution is a separate matter from the error itself. The literature's implicit rule assigns errors by AI-assisted pro se litigants to the category, as evidence of what the technology does, and errors by counsel to the individual, as an oversight in an otherwise reliable practice. The episode documented here offers a test of that rule, and it is best stated as a question: had a self-represented litigant using generative AI cited the transportation provision of the ADA in place of the public-accommodations provision, on a dispositive element, would the error have been read as an isolated oversight, or as one more entry in the catalogue of what happens when laypeople litigate with chatbots? The two readings would be unlikely to match, and this asymmetry of attribution, not the error itself, is part of what the architecture-dependence hypothesis asks the literature to examine: whether conduct is being coded by who produced it rather than by how it was produced. The asymmetry, moreover, persists against a moving baseline: as Part II records, the documented hallucination incidents are no longer predominantly a self-represented phenomenon, yet the framing of systemic threat remains attached to the self-represented side of the docket.

E. Early settlements.

Figure 7. *Sequence of the four settlements.*



Source: Public docket, ECF Nos. 83–86 and 94 (stipulations of dismissal with prejudice; ECF Nos. 83 and 84 are two same-day filings of the Mistral stipulation, the order of dismissal citing No. 84).

Four defendants that had entered the litigation resolved their disputes with the plaintiff and were dismissed with prejudice within a compressed period (Figure 7). As is standard, only the stipulations of dismissal appear on the public docket; the settlement terms are confidential, the

reasons are not in the record, and no cause is inferred. This subsection puts on the record a settlement sequence that is ordinarily invisible, and thereby makes it verifiable.

Two features of the sequence are worth stating, because they are visible on the public docket. In at least two instances a defendant that had aligned with the coordinated defense subsequently left it by stipulation of dismissal, one settling after the opposition was filed and before the coalition's reply; the other after first incorporating the coalition's joint memorandum into a motion of its own. The chronological pattern is a matter of public filing: defendants that had entered a common front exited it, one at a time, while the motion they had joined remained pending.

The other feature is that settlements between corporate defendants and a self-represented litigant are ordinarily confidential and leave little trace in the public record. Here the stipulations of dismissal are themselves docket entries, which makes the sequence unusually legible; the evidentiary value of that record lies in its verifiability.

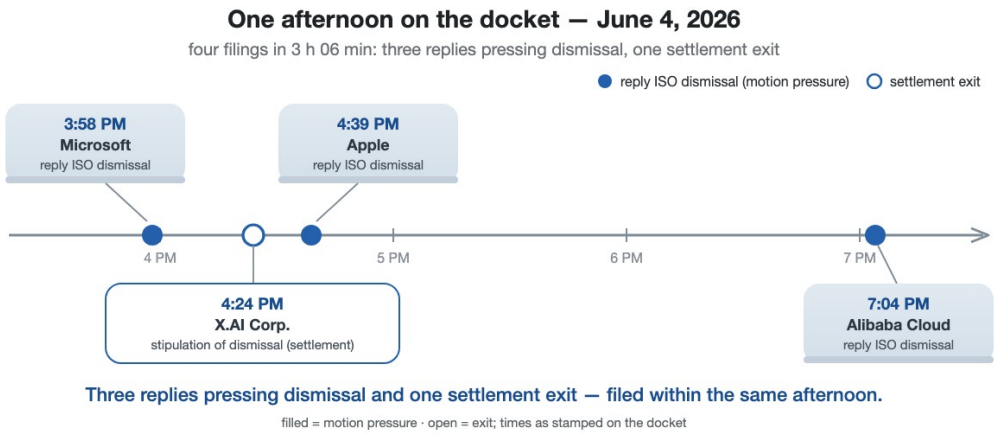
Table 3. *Status of the thirteen named defendants (as of June 2026).*

Status	Defendants
Did not appear	Samsung; DeepSeek; OpenAI
Settled and dismissed	Mistral AI; Perplexity AI; Anthropic PBC; X.AI Corp.
Continuing	Microsoft; Apple; Google; Meta; Amazon; Alibaba Cloud

Source: Public docket, as of June 2026.

Figure 8 reproduces the docket sequence of a single day, June 4, 2026, as a primary record. It is included so that the reader can inspect the timing directly. What configurations of AI-assisted litigation produce records of this shape, and how often? That question belongs to the systematic study Part VII proposes.

Figure 8. *Docket sequence on June 4, 2026.*



Source: Public docket; filing times on a single day, ECF Nos. 93-96.

These features together describe a self-represented litigant sustaining a multi-party federal action through a full briefing cycle, meeting deadlines, responding in proportion, and identifying an opponent's error, against a coordinated, represented defense. That is not the pattern the prevailing account describes. That account, of undifferentiated volume and unreliable output, characterizes a distribution, and a single case cannot contradict a distribution; but the pattern documented here falls outside what the aggregate studies measure, which is why it has remained invisible to them. Whether the orchestration architecture is responsible for the pattern, and whether the pattern recurs across cases, are questions the next Parts frame for systematic inquiry.

VI. Doctrinal Implications

The harder questions are doctrinal, and two of them are moving through the federal courts right now. The aim is to show how an orchestration-based, quality-oriented account of pro se AI use reframes each. Courts already possess a developed framework for evaluating AI outputs as evidence (Grimm, Grossman & Cormack 2021). The questions here are adjacent ones: they are not about whether machine output is admissible; both concern who owns the process that produced the output and whether that process practices law.

A. Work product and the pro se litigant.

Whether a litigant's interactions with an AI system are protected from discovery as work product is newly contested. Within a seven-week span in early 2026, three federal courts reached the intersection of AI use and work-product protection, and they did not speak with one voice.[21] In *Warner v. Gilbarco, Inc.*, a magistrate judge denied a motion to compel a pro se employment plaintiff to produce her generative-AI queries and the responses she received, holding that the materials were work product under Federal Rule of Civil Procedure 26(b)(3): a self-represented litigant acts as her own counsel, the materials reflected her mental impressions and preparation, and disclosure to a generative AI system, a tool rather than a person, was not the adversarial disclosure that waiver requires. The holding sits in the doctrine's oldest stratum, since what Hickman v. Taylor shielded most strongly was the lawyer's mental impressions, conclusions, and theories: the tier of work product courts treat as nearly absolute (Hickman v. Taylor, 329 U.S. 495, 510–11 (1947)). The queries a litigant frames, the drafts he iterates, and the arguments he discards are that tier's pro se analogue. One week later, in *United States v. Heppner*, a court reached the opposite result: a represented criminal defendant had used a public AI assistant on his own initiative, without any involvement of counsel, and the court ordered the resulting files produced, emphasizing that he had acted without direction from counsel and that the platform's consumer terms permitted the provider to retain and use his inputs. *Morgan v. V2X, Inc.*, the third and most elaborate decision of the line, extended Warner's core holding, concluding that a pro se litigant's use of AI in litigation preparation "closely resembles the kind of confidential, strategy-laden iterative work product that Rule 26(b)(3) was designed to protect." The same order, however, required the litigant to disclose the name of any AI tool he had used in connection with material the other side had designated confidential, and amended the protective order to restrict the use of mainstream AI services with such material.[22] The distinction the court drew is analytically important, because it explains why Warner and Heppner are not in conflict: a represented party who bypasses counsel creates a gap between party and advocate, the gap that proved fatal in Heppner, whereas a pro se litigant "is simultaneously the party and the advocate," so no such gap exists. The protection was not unlimited: the court held that the identity of the AI tool was not

[21]Warner v. Gilbarco, Inc., No. 2:24-cv-12333, 2026 WL 373043 (E.D. Mich. Feb. 10, 2026) (Patti, M.J.); United States v. Heppner, No. 25-cr-503, 2026 WL 436479 (S.D.N.Y. Feb. 17, 2026) (Rakoff, J.); Morgan v. V2X, Inc., No. 25-cv-01991-SKC-MDB, 2026 WL 864223 (D. Colo. Mar. 30, 2026).

[22]Morgan v. V2X, Inc., 2026 WL 864223, at *3 (D. Colo. Mar. 30, 2026).

itself shielded, and it conditioned AI use on confidential discovery materials on contractual safeguards against training, retention, and third-party disclosure. The three decisions thus converge on applying the traditional doctrine to AI rather than carving AI out of it, but diverge on who may claim the protection and on what counts as waiver. The line they draw is an architectural one, though the vocabulary is this Article's: the opinions speak of agency, confidentiality, and waiver. What doomed the Heppner claim was not the use of AI but the configuration used: a consumer tool whose terms permitted retention and training. Morgan, by contrast, conditions protection on a configuration contractually barred from training, retention, and third-party disclosure. Confidentiality posture, like verification, is a property of the architecture of assistance rather than of AI as such. The Article's variable appears here inside the doctrine itself (Table 4).

Table 4. The early-2026 work-product triad, by configuration.

Case	Litigant	Agency gap?	Tool configuration	Holding
Warner	Pro se plaintiff	None — party and advocate are one	Generative AI used as her own preparation tool	Protected as work product
Heppner	Represented defendant, acting without counsel	Yes — party bypassed advocate	Consumer tool; terms permitted provider retention and training	Production ordered
Morgan	Pro se plaintiff	None — "simultaneously the party and the advocate"	Use on confidential materials conditioned on contractual bars: no training, retention, or third-party disclosure	Protected, with conditions; tool identity not shielded

Source: Warner, 2026 WL 373043; Heppner, 2026 WL 436479; Morgan, 2026 WL 864223. The decisions converge on applying the traditional doctrine and diverge on configuration facts; the architectural reading of the final two columns is the Article's, not the courts' — architecture changes the facts the doctrine weighs, not the law.

Morgan was decided on March 30, 2026, the day the joint motion to dismiss was filed in the case studied here: the doctrine was forming while the case was being briefed.

Read together, the early-2026 decisions make the architecture of the tool increasingly dispositive: who directed it, what it did, how its record was kept. The orchestration account sharpens what is at stake in that line. If AI assistance to a self-represented litigant were merely mechanical text generation, the case for treating it as protected strategy would be weak; an orchestration workflow, in which a litigant frames questions, weighs divergent model outputs, and iterates toward a position, is difficult to distinguish, in function, from the strategy-laden iterative

process the work-product doctrine has always protected in the hands of counsel. A litigant directing a panel of divergent analytical engines resembles nothing so much as a lead attorney directing a team of junior associates. The doctrine's underlying purpose, preventing one party from free-riding on another's litigation preparation, applies with equal force whether that preparation is performed by an associate or by a litigant working through a multi-model system. *Morgan* points in this direction, and the orchestration framing supplies the functional rationale.

Work-product protection depends on confidentiality, and it can be waived by voluntary public disclosure, at least as to what is disclosed. Publishing an account of one's own pending case therefore risks waiving, as to what it discloses, the protection *Morgan* recognizes.

B. Unauthorized practice of law.

The UPL concern in the pro se setting is, on inspection, oddly framed. The unauthorized-practice-of-law rules reserve the practice of law to licensed attorneys in order to protect the public from unqualified representation of others.[23] Those rules are creatures of state law, and their definitions of practice vary by jurisdiction; what follows concerns the framework's shared logic rather than any one state's boundary. The genuine UPL question raised by consumer legal AI is whether the tool, or the company operating it, engages in unauthorized practice by advising the litigant. That question turns substantially on the reliability of the advice: an instrument that confidently fabricates authority looks like the incompetent practice UPL rules exist to prevent.

A pro se litigant, however, represents no one but himself. He has a statutory right to do so (28 U.S.C. § 1654), and the sanction for a defective filing falls on him rather than on a client. When such a litigant uses an AI system, the UPL framework must locate the "practice of law" somewhere: in the tool, in its developer, or nowhere, and each of those answers is awkward. What actually animates the concern is not authorization but reliability, the fear that an unlicensed, unaccountable system will supply confidently wrong law to a person with no capacity to check it. Reliability is not the whole of it, because courts police the practice of law through accountability as much as competence: a licensed attorney can be sanctioned, suspended, or disbarred, and an AI system can be none of these. The accountability gap is real, and the orchestration account addresses

[23]See ABA Model Rules of Prof'l Conduct r. 5.5; Joseph J. Avery, Patricia Sánchez Abril & Alissa del Riego, ChatGPT, Esq.: Recasting Unauthorized Practice of Law in the Era of Generative AI, 26 Yale J.L. & Tech. 64 (2024).

reliability, not accountability. If an orchestrated system erred systematically, responsibility could fall in one of three places: on the operator who signs the filing (the only actor Rule 11 presently reaches), on the provider of the architecture, or on no one. Which of the three the law selects will determine whether the accountability gap narrows or merely relocates.

Here the single-model/orchestration distinction does real work. The reliability deficit that animates the UPL concern is characteristic of the single-model configuration, and an architecture built around cross-model verification and citation validation is directed at that deficit. That does not dissolve the UPL question, since a highly reliable tool may still be said to "practice law." What it does is separate two things the debate tends to merge: reliability and authorization are analytically distinct questions, and conflating them obscures that an architectural fix can address the first without touching the second. If reliability is the true concern, then architecture, and not licensure alone, becomes part of the analysis, and a system designed to surface and resolve disagreement rather than assert a lone answer addresses the concern on its own terms. If reliability is architectural, it is also designable and certifiable, the kind of object on which proposals to reinvent the regulation of legal services depend (Hadfield 2017). An architecture that declines to emit what it cannot verify, and that routes irreducible uncertainty to the human, begins to function as compliance by design rather than as an unlicensed adviser.

C. A reframing, not a resolution.

Both doctrinal areas remain unsettled. The dominant volume-and-hallucination account of pro se AI use has shaped these doctrinal debates, and a quality-oriented, orchestration-based account reframes them. Work-product protection looks more coherent when AI use is understood as iterative strategy rather than mechanical output, and the UPL debate looks different when reliability is treated as an architectural variable rather than an inherent defect.

VII. Limitations and Future Work

Four limits bound the design. First, and most fundamentally, this is a single case (n = 1) authored by a party to it. The findings in Part V are existence claims rather than distributional ones. What such a case can do is what qualitative methodology has long recognized for extreme or revelatory cases: expose a phenomenon that standard designs have not yet reached, and generate the hypotheses that systematic study must then test.

Second, the author is litigant, developer and analyst; the constraints in Part IV exist because of it.

Third, causation cannot be isolated. The record cannot separate the architecture's contribution from the litigant's own skill, prior preparation, the incentives of an unbilled party, or other confounders. The architectural discussion in Part III supplies a plausible mechanism, not a demonstrated one.

Several rival explanations remain plausible. The litigant may simply be capable: educated, motivated, and unbilled, a pro se party can spend hundreds of hours on a case that counsel would staff in tens. The three-day opposition may measure preparation rather than processing (Part V.B). Selection operates at two levels. At the case level, this case is the subject of an article because its record is favorable, and cases in which orchestration-assisted litigants fail generate no articles. At the episode level the selection is the author's own: the docket contains more events than the five this Article describes, and the author chose which to present. Against that selection, the completeness statement is docket-verifiable: as of June 2026, the record contains no order sanctioning the plaintiff, no order striking any of his filings, and no deadline he missed. The nearest the record comes to an adverse procedural note lies at the case's threshold, where the court denied the plaintiff's application to proceed in forma pauperis (the filing fee was then paid within the time allowed) and observed that summonses could not issue until service addresses were supplied, which they then were. The statement is falsifiable by anyone with PACER access; what episode-level selection can still conceal is emphasis, not adverse rulings. The study proposed below is designed to adjudicate among these rival explanations.

One further limitation goes to reach rather than validity. Part II frames the justice gap in terms of the low-income Americans the Legal Services Corporation studies, while the litigant here is a Ph.D.-level AI-systems architect operating a system of his own design, a self-represented plaintiff drawn from the far tail of the distribution of technical capability. The existence proof is not weakened by that fact, but its scope is narrower than the population Part II centers: nothing in this case shows that orchestration-assisted litigation is practically available to that population, for whom cost, digital access, and technical fluency are themselves barriers. Whether the architecture's benefits survive the transfer from an expert operator to a typical self-represented litigant is an empirical question, and for the assistive framing of Part II it is arguably the decisive one; the disability arm of the study proposed below is designed to reach it. The transfer is not only a matter of technical setup. Orchestration relocates the operator's cognitive burden from drafting

to arbitration: reading competing outputs, weighing surfaced disagreement, deciding which line survives. Arbitration is itself a demanding skill. Whether that burden can be carried by interface design rather than by operator expertise, and at what cost to litigants with cognitive, visual, or educational barriers, is as much a design question as an empirical one.

Fourth, the case is pending; no finding here depends on how it ends.

The hypothesis this case generates, that orchestration architecture is associated with higher-quality pro se litigation conduct than single-model prompting, is testable. A systematic study could compare litigation conduct (deadline adherence, citation accuracy, motion outcomes) across a sample of pro se filings coded by mode of AI assistance, using the AI-consistent-drafting indicators the empirical literature has begun to develop and LegalBench, a collaboratively built benchmark for legal reasoning (Guha et al. 2023). One task family, seeded-error detection, follows directly from Part V. Test briefs are salted with verifiable defects of the kinds documented there: a transposed statutory section, a misquoted authority, a fabricated citation. Detection rates are then compared across assistance modes, from single-model prompting to orchestrated verification, with lawyer and non-lawyer operators as separate arms. The architecture's components (external-source verification and cross-model disagreement) are ablated separately, so that the contribution of each mechanism is identified rather than credited to the bundle (Table 5). Such a design, conducted by disinterested researchers across many cases, could convert the existence claim advanced here into a distributional one, or refute it. Stronger designs are available and should be preferred: controlled comparisons of single-model and orchestrated assistance on identical legal tasks, with pre-registered protocols, open materials, and blind-set validation of the system's own claims, including its citation-verification rate; quasi-experimental identification, such as difference-in-differences designs around the public availability of multi-model tools; and a dedicated arm for participants with motor, visual, or cognitive disabilities, for whom the assistive framing of Part II is not hypothetical. Randomized evaluation of legal assistance has precedent: the access-to-justice literature has run randomized studies of offers of representation and of unbundled assistance (Greiner, Pattanayak & Hennessy 2013), and the assistance-mode comparison proposed here extends that design tradition to the architecture variable. More broadly, the hypothesis suggests a standard the field has not yet adopted: that AI systems be evaluated not only by the quality of their outputs, but also by the quality of their reasoning architecture. Nothing in it is specific to law:

wherever consequential work is delegated to generative systems, in medicine, engineering, or finance, the same question arises, and the same variable goes unmeasured. It also reframes a variable that access-to-justice analysis has long treated as fixed. The plaintiff's entry cost was the $405 filing fee, and a defense attorney's formal entry cost was the $200 pro hac vice fee (ECF Nos. 19, 52, 60). Both are in the low hundreds of dollars, and they stand at the two ends of what has been an effectively unbridgeable resource gap; the comparison concerns formal entry fees only, not the cost of representation. What may have changed is not the price of entering a courtroom but the marginal cost of the litigation capability required to remain in one. Whether that cost has in fact fallen, and by how much, is an empirical question for the proposed study; that it has become a variable rather than a constant is itself a hypothesis worth testing.

Table 5. The proposed study, by arm.

Arm	Configuration	What it isolates
1	Human only, no AI (control)	Baseline detection without assistance
2	Single-model prompting	The mode the literature documents
3	Orchestration, full	The bundle as deployed
4	Orchestration without external-source verification	Contribution of citation checking
5	Orchestration without cross-model disagreement	Contribution of surfaced divergence

Note: Each arm is run with lawyer and non-lawyer operators as separate groups, and a separate disability arm, run outside the five numbered arms, for participants with motor, visual, or cognitive disabilities, and with outcome measures pre-registered before sessions; see text.

VIII. Conclusion

The prevailing account of artificial intelligence in pro se litigation is a story about volume and unreliability: more litigants, filing faster, with a heightened risk of fabricated authority. It assumes, without saying so, that AI assistance means one litigant prompting one general-purpose model, and that assumption no longer holds.

The case examined here was built, by the author's account, on a different architecture: several independent models were queried in parallel, and their disagreement was surfaced and resolved rather than suppressed. The description of the litigation conduct rests entirely on a verifiable public record; that record contains conduct the volume-and-hallucination account neither describes nor predicts: a rapid reconstruction of a defective filing, a substantive opposition filed far ahead of

schedule, and the identification of a statutory-citation error in a coordinated defense brief. The Article then showed how understanding AI assistance as orchestration, and as a matter of quality, reframes two doctrinal debates now before the federal courts. For litigants who cannot type, see, or organize at the speed litigation demands, the architectural difference is a question not of efficiency but of access in the Americans with Disabilities Act's sense of full and equal enjoyment: whether the way AI assistance is configured is what makes self-representation possible at all.

The claim is that the architecture of AI assistance, not only its availability, belongs in the analysis of what self-represented litigants can now do. That claim only grows more pressing as foundation models converge in raw capability, because when the models themselves differ less, the way they are combined will differ more. The architecture-dependence hypothesis is that architecture is no implementation detail: it is a determinant of the quality of what people, and courts, can do. Whether AI participates in legal work is no longer in question; what remains open is which architectures do what. On this account the pro se AI problem is a design problem, not an intelligence problem, and design problems have design answers. The work ahead is to specify particular architectures and test them against one another instead of waiting for better individual models. The Appendix states one such architecture precisely enough to be run and criticized, and a reference implementation is published with it.

Appendix. One Assistance Architecture, Specified So It Can Be Run

The argument of this Article is that "AI assistance" names a family of architectures with different failure properties. The claim is easier to test once at least one architecture is stated precisely enough to be run and criticized. The protocol below is such a statement. It is called here the *blind-consolidation protocol*[24], after the stage that most plainly separates it from multi-agent debate, in which the models read one another.[25] It uses only public models, and anyone with consumer or API access to three vendors can reproduce it.

The protocol below states, at the level of architecture, the family described in Part III: heterogeneous models queried in isolation, a consensus formed under explicit rules with the disagreement preserved rather than averaged away, and every citation checked against its source. It differs from the system described in Part III in how that last check is made: there it is a component that validates against a live database, here it is search-enabled models followed by a person, and constraint three below says plainly that the models' approval is not the check. It is stated here as a specification, not as evidence: the Article offers no filing in the case study as its output and draws no inference from it (what is withheld: note 24). What the protocol contributes to the argument is the run manifest. The architecture variable in this Article is self-reported (Part IV); a manifest is what an outside reader would need before it could be anything else. The filings in Part V carry none.

The problem it addresses.

A self-represented litigant can reach several frontier models through free consumer tiers, and cost, digital access and technical fluency remain barriers (Part VII): the high-stakes column below is paid API access, and the confidentiality condition set out later excludes most free tiers for the material that most needs protection. The common practice is to ask one, read the answer and file it. The documented failure is a confident model producing a citation that does not exist, with nothing in the process positioned to catch it. The protocol replaces the single ask with a common

[24]This Article describes only publicly filed documents and architectural features already disclosed to the defendants and the public. It withholds the prompts used, the system's decision logic on any particular question, and all draft work product.

[25]docprep, a reference implementation of the blind-consolidation protocol, https://github.com/vadimchernets/docprep (MIT license); the version described here is release v1.0.0. The repository holds the runnable pipeline in one file, the configuration of Table 6, the manifest format, and offline tests that run without accounts or keys, together with this Appendix as a standalone specification under CC BY 4.0. Model identifiers and access must be configured for the operator's own accounts.

task statement and four stages that follow it. It leaves each model exactly as accurate as it was; it arranges several so that one system's error has to survive contact with systems trained by other organizations, and it records what each of them said.

Roles.

The protocol is defined by what each role is allowed to see. Four rules fix it:

1. Generators see the task statement and nothing else. None sees another generator's draft.
2. The consolidator sees every draft under neutral labels R1 to Rn, in a random order. It never learns which vendor produced which draft.
3. The first critic reads inside the consolidator's own family, in a fresh context. It sees the task statement and the consolidated document, and nothing of the consolidation that produced it.
4. The last reader is a critic from a different vendor than the consolidator. It sees the task statement, the consolidated document and the first critic's report, and its verdict closes the run.

Any set of frontier models from different vendors that satisfies the four rules runs the protocol. As an illustration, these are public models that were available in September 2026 and could fill the roles:

Table 6. Public models available in September 2026 that could fill the four roles.

Role	Routine documents	High-stakes documents
Generators, isolated	claude-sonnet-5 with search; gpt-5.6-terra with search; gemini-3.1-pro-high with search; an open-weight model	claude-opus-5 with search; gpt-6-astra with search; gemini-3.1-pro-high with search; Grok (xAI); an open-weight model
Consolidator, blind	claude-sonnet-5	claude-fable-5-1
Critic 1, consolidator's family, fresh context	claude-sonnet-5	claude-fable-5-1
Critic 2, different vendor, reads last	gpt-5.6-terra	gpt-6-astra

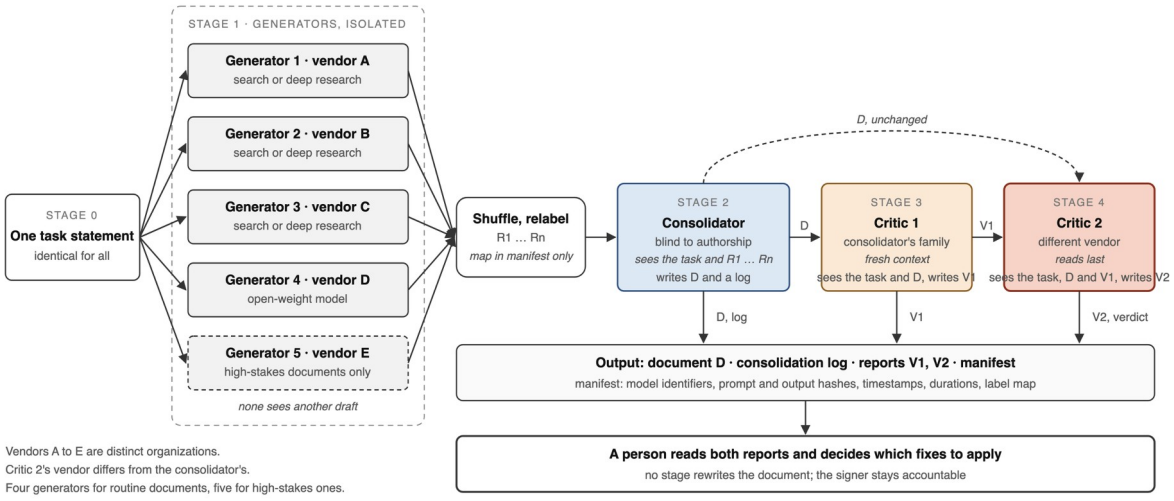
Source: Vendor documentation as of September 2026. The assignment of models to roles is the author's; the roles, not the identifiers, define the protocol.

The table is an illustration of what was available, and the identifiers will age. The roles will not. Grok is named by vendor and the open-weight model by class, because their hosted identifiers vary from host to host, and gemini-3.1-pro-high is the alias of a hosted interface rather than the

model code in the vendor's API documentation. The high-stakes column adds a fifth generator, and its consolidator is a model that wrote no draft, so it never judges its own text under a label. Figure 9 draws the protocol by role.

The assignment of families to roles rests on the author's use, not on a measurement. Across roughly two thousand hours of work with these systems in 2025 and 2026, a model placed in a fresh context as a blind critic returned the strictest and most useful verification when it came from the same family as the consolidator, and it did so even where the text under review had been produced by that family. Critics from other vendors were not weaker readers; they found different things, which is why the protocol keeps one of them and gives it the closing verdict. The four rules are about what each role may see, not about which vendor fills it, and a reader who prefers a different family at the consolidator or at the first critic can substitute one and still be running the protocol.

Figure 9. *The blind-consolidation protocol, by role: isolated generation, blind consolidation, and two verifications.*



Source: Author's diagram of the protocol specified in this Appendix.

This is a ramp. It was built for someone who cannot take the stairs: who cannot hold five interfaces open at once, read them against each other, and carry the answer back. A ramp is used by everyone who reaches the door, not only by the person it was built for. The formula above is written down so that anyone can take it.

Stages.

Stage 0. One task statement. A single written statement sets out the objective, the jurisdiction, the required format, what counts as a source, and what to do with a claim that cannot be verified. Every model receives it unchanged, so that differences between drafts come from the models.

Stage 1. Isolated generation. Each generator produces a complete draft in parallel, with web search enabled wherever the interface or the API carries it, and with deep research where the task calls for depth. The consumer interfaces of these vendors search by default; some programmatic endpoints carry no search tool, and a generator running on one of those is drafting from weights, which the run record must show. Each is required to cite a verifiable source for every factual and legal proposition, to mark what it could not verify as unverified, and to close with its open questions. Isolation is the mechanism. Models built by different organizations on different corpora invent different things, and a claim that several of them produce independently, each with its own citation, stands in a different evidential position from a claim one of them produced alone. A drafter that sees another draft agrees with it for reasons unrelated to whether the claim is true.

Stage 2. Blind consolidation. One model receives the drafts as R1 to Rn, in an order chosen at random and recorded only in the run manifest, and merges them under explicit rules. Claims supported by several drafts or by a verifiable citation are preferred for inclusion, which is a rule about what enters the draft and not a finding that they are true. Material found in a single draft is kept when it carries a verifiable citation. Conflicts are resolved with evidence or flagged in the text. Agreement between drafts is recorded as agreement and never promoted to proof. New material enters only if the consolidator verifies it itself, and is marked as added. Each substantive claim carries a tag naming the drafts that support it. The consolidator also writes a log: what the drafts agreed on, what they disagreed on and how each conflict was resolved, and what was dropped and why. The log is where the disagreement survives. Because the label map is kept, a later reader can set the tags against it and measure how often the consolidator preferred its own vendor's draft.

Stage 3. Internal verification. A critic from the consolidator's family reads the merged document in a fresh context, with no memory of the consolidation. Its task is verification, and it returns no improved document. It checks every citation for existence and accuracy, tests the

reasoning for gaps and overstatement, confirms the task statement was met, and reports findings ranked Critical, Major and Minor with a verdict: approve, approve with fixes, or reject.

Stage 4. Cross-vendor verification. A critic from a different vendor reads the same document and the first report. It confirms, disputes or extends each finding, adds its own, and gives the closing verdict. The stage exists because of the weakness of the two before it: a consolidator checked by its own family shares that family's blind spots, and a reader trained elsewhere is the most direct route to them.

What leaves the pipeline.

The deliverable is the consolidated document, the consolidation log, both verification reports and a manifest. No stage overwrites the document with a critic's output. A person reads the reports and decides which fixes to apply. In litigation the person signing the filing is accountable for its contents, and a pipeline that silently rewrote the text would obscure who decided what.

The manifest records, for every stage, the model identifier, the prompt hash, the output hash, the start time and the duration, together with the label map. Without it, a document prepared this way is indistinguishable from one produced by asking a single model twice. With it, a reader who was not present can audit the operator's record of the run: which identifier was configured for which role, in what order the drafts were labeled, what each stage received and returned. It is a structured self-report, not a proof. The identifiers come from the configuration rather than from an authenticated response, a stage run by hand carries the operator's word for what happened in the interface, and a record of a search tool being offered is not a record of a source being checked. Making the variable observed rather than reported takes third-party custody of the run, which is what the design in Part VII supplies and what a file format cannot.

Reference implementation.

A reference implementation, docprep, is published under an open license (note 25): some five hundred lines of Python in one file, with the configuration of Table 6, the manifest format, and offline tests that run without accounts or keys.

The implementation separates what it enforces from what it only asks. The four role rules are enforced. A generator's prompt is assembled from the task statement alone; the drafts reach the consolidator shuffled and relabeled, with the map written only to the manifest; and a configuration

whose first critic sits outside the consolidator's family, or whose last critic sits inside it, is refused before any model is called (Figure 10). The merging rules of Stage 2 are only asked: they are instructions in the consolidator's prompt, and a model can disobey them. One is checked afterwards, since a consolidation returned without its log stops the run.

Figure 10. *A role rule as a precondition of the run: the check on the closing critic, and the refusal it prints.*

```
home = family(cfg, cons)              # the consolidator's vendor family
    . . .
if critics and family(cfg, critics[-1]) == home:
    problems.append(f"{critics[-1]['name']} is from the consolidator's family "
                    f"({home}); the last reader must come from another vendor")
    . . .
if problems and enforce:
    sys.exit(f"formula '{formula_name}' breaks the protocol:\n  - "
             + "\n  - ".join(problems))


$ python3 docprep.py run --formula manual
formula 'manual' breaks the protocol:
  - critic_cross is from the consolidator's family (anthropic);
    the last reader must come from another vendor
```

Source: docprep, docprep.py, function validate(), release v1.0.0 (note 25); elisions marked. Below the code, the output of the command shown, run on the example configuration with the closing critic reassigned to the consolidator's vendor. The repository's offline tests assert the same refusal.

Enforcement has the limits the manifest has. A role's vendor family is declared in the configuration, not authenticated. A stage run by hand is blind and fresh on the operator's word. Relabeling removes names, not style. And the refusal can be switched off, in which case the manifest records that it was and lists the rules broken. The rules are enforced against an operator's mistake, not against an operator.

The legal envelope.

An architecture that lowers error correlation by adding vendors raises, by the same act, the number of places where the material leaves. For a self-represented litigant in a United States court that trade is not abstract.

The position in the United States as of September 2026 has three parts: what a filer must disclose, what may be sent to a provider at all, and whether the work survives as work product. Part II and Table 1 carry the disclosure regime; Part VI.A carries the work-product analysis and its authorities. No generally applicable federal procedural rule requires a filer to disclose the use of AI or to name a model: in April 2026 the Advisory Committee on Civil Rules dropped from its agenda two proposals directed at fabricated authority, agreeing that rulemaking would not be

appropriate at that time.[26] Local rules and individual standing orders do require it, and they differ. The Northern District of Texas asks for the fact of AI use on the first page of a brief and does not ask which model. Judge Vaden of the Court of International Trade asks for the program used, for the specific portions of text it produced, and for a certification that the use disclosed no confidential or business proprietary information to an unauthorized party.[27]

Two rulemaking actions of spring 2026 are easy to conflate, and the conflation changes the answer. The Advisory Committee on Civil Rules took up two proposals to amend Rule 11 to address citations of authority that do not exist, agreed that rulemaking would not be appropriate at that time, and dropped the item on April 14, 2026. Three weeks later the Advisory Committee on Evidence Rules took up proposed Federal Rule of Evidence 707, which governs the admissibility of machine-generated evidence rather than the disclosure of AI use; it revised the draft and kept it under study rather than advancing it. The first action concerns what a filer certifies, the second what a court may admit, and neither created a duty to disclose. The two are linked in the record itself, which is why they are read as one: the April memorandum of the Civil Rules committee quotes the draft of Rule 707, and its minutes note that the Evidence Rules committee had that rule under study.

Material another party has designated confidential is the harder case, and it reaches this architecture directly. In *Morgan* the court allowed a self-represented plaintiff to put such material into an AI service only where the provider is contractually bound not to store or use the input to train or improve its model, not to disclose it to third parties except where disclosure is essential to delivering the service and the recipient is bound by equivalent protections, and to remove or delete it on request; it required the plaintiff to name any tool used in connection with that material and to

[26]Advisory Comm. on Civil Rules, Draft Minutes of Meeting of Apr. 14, 2026, at 18, reprinted in Comm. on Rules of Practice & Procedure, Agenda Book 383 (June 3–4, 2026) (item 17, "Artificial Intelligence Hallucinations"); see also Advisory Comm. on Civil Rules, Agenda Book 438–50 (Apr. 14, 2026) (Suggestions 25-CV-S and 25-CV-V, both directed at Rule 11); Report of the Advisory Comm. on Civil Rules 60 (May 6, 2026), in Agenda Book, supra, at 364 (matters dropped from the agenda). The minutes are in draft pending approval at the committee's fall 2026 meeting. Proposed Federal Rule of Evidence 707, which governs the admissibility of machine-generated evidence rather than disclosure, is a separate matter: the Advisory Committee on Evidence Rules revised it on May 7, 2026 and kept it under study rather than advancing it.

[27]N.D. Tex. Civ. R. 7.2(f) (disclosure of the use of generative artificial intelligence on the first page of a brief, with further disclosure of the portions so prepared if the presiding judge directs); Order on Artificial Intelligence, U.S. Court of International Trade (Vaden, J.) (June 8, 2023) (requiring identification of the program used, of the specific portions of text it produced, and certification that the use disclosed no confidential or business proprietary information to an unauthorized party).

keep written documentation of those safeguards. The court observed that the condition excludes most mainstream low-to-no-cost AI. Those are the terms of one protective order rather than a national rule, and they are the most restrictive the author has located; the protocol adopts them as its own design policy.

Three constraints follow. They are design rules of this protocol, not statements of what the law requires of a litigant.

1. Public filings and material the operator is free to transmit may go to the models. Sealed material, material under a protective order, and another party's confidential discovery do not, unless the order permits it and the provider's terms meet the condition *Morgan* imposed. The restriction covers every stage and every recipient, not the generators alone: the consolidator and both critics receive the merged draft, and the retained artifacts travel with it. Ownership of a document does not by itself discharge an obligation of confidentiality attached to it. The manifest makes the exposure auditable; it does not make it lawful.
2. The manifest supports the disclosure; it is not the disclosure. It records which model identifier produced which artifact and when, which is the material a court asking for the program used will want, and it is worth keeping whether or not the presiding judge asks. It does not file the notice a standing order requires, identify the portions of the final document a rule such as Judge Vaden's reaches, or supply any certification.
3. Verification of every citation against its source is done by a person. Rule 11(b) reaches an unrepresented party by its terms, and its standard is reasonable inquiry under the circumstances rather than any particular procedure; Part II sets out what the record of fabricated citations shows. Agreement among models and approval by two critics are not verification.

Requirements vary by district and by judge, and the standing order of the presiding judge governs.

Limits.

The deep-research modes of the consumer interfaces have no exact equivalent in the vendor APIs. A stage that needs one is run by hand: the prompt is written to a file, the operator runs it in the interface and pastes the answer back, and the manifest records that the stage was run that way. That fallback is manual, and Part II sets out why a manual workflow across several interfaces is not merely slower for a litigant with a motor or visual impairment. A protocol that depends on it inherits that limit.

Agreement between models is evidence of correlation as much as of truth. Models trained on overlapping corpora share errors, and a unanimous set of drafts can be wrong together. The protocol lowers the correlation it can reach and records the rest.

The protocol has not been compared with single-model prompting. The comparison that would settle its value is a blind assessment: the same tasks prepared both ways, the resulting documents scored by readers who do not know which is which, on citation accuracy, completeness against the task statement, and errors a court would notice. The protocol is stated here so that this comparison can be run.

The repository named above holds the pipeline, the configuration and the offline tests.

Selected References and Authorities

ABA Model Rules of Professional Conduct r. 5.5.

Americans with Disabilities Act, 42 U.S.C. §§ 12181–12188.

Ashley, Kevin D., Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age (Cambridge Univ. Press 2017).

Avery, Joseph J., Patricia Sánchez Abril & Alissa del Riego, ChatGPT, Esq.: Recasting Unauthorized Practice of Law in the Era of Generative AI, 26 Yale J.L. & Tech. 64 (2024).

Blanck, Peter, eQuality: The Struggle for Web Accessibility by Persons with Cognitive Disabilities (Cambridge Univ. Press 2014).

Cappelletti, Mauro & Bryant Garth, Access to Justice: The Newest Wave in the Worldwide Movement to Make Rights Effective, 27 Buff. L. Rev. 181 (1978).

Carpenter, Anna E., Colleen F. Shanahan, Jessica K. Steinberg & Alyx Mark, Judges in Lawyerless Courts, 110 Geo. L.J. 509 (2022).

Centers for Disease Control and Prevention, Disability Impacts All of Us (2022 BRFSS data).

Chan, Chi-Min, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu & Zhiyuan Liu, ChatEval: Towards Better LLM-Based Evaluators Through Multi-Agent Debate, arXiv:2308.07201 (2023).

Charlotin, Damien, AI Hallucination Cases Database, https://www.damiencharlotin.com/hallucinations/ (last visited September 2026).

Chernets, Vadym, docprep: A Reference Implementation of the Blind-Consolidation Protocol, release v1.0.0 (2026), https://github.com/vadimchernets/docprep (MIT license; the protocol specified in the Appendix, and the code that runs it).

Chernets v. Google LLC et al., No. 1:25-cv-05691-RER-RML (E.D.N.Y.) (public docket); PolyHelper.AI LLC v. United States of America, No. 1:25-cv-07770 (S.D.N.Y. 2025) (earlier action, different defendant).

Cohen-Sasson, Or, Stochastic Justice: Legal Inconsistency by Humans and AI, 105 Neb. L. Rev. (forthcoming 2026).

Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum & Igor Mordatch, Improving Factuality and Reasoning in Language Models through Multiagent Debate, Proceedings of the 41st International Conference on Machine Learning (ICML 2024), arXiv:2305.14325.

Engstrom, Nora Freeman & Aviv Caspi, Opening the Courthouse Door—or Just Lowering the Threshold?, JOTWELL: Legal Profession (July 3, 2026).

Fed. R. Civ. P. 26(b)(3).

Greiner, D. James, Cassandra Wolos Pattanayak & Jonathan Hennessy, The Limits of Unbundled Legal Assistance: A Randomized Study in a Massachusetts District Court and Prospects for the Future, 126 Harv. L. Rev. 901 (2013).

Grimm, Paul W., Maura R. Grossman & Gordon V. Cormack, Artificial Intelligence as Evidence, 19 Nw. J. Tech. & Intell. Prop. 9 (2021).

Guha, Neel, et al., LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models, Advances in Neural Information Processing Systems 36 (2023).

Hadfield, Gillian K., Rules for a Flat World: Why Humans Invented Law and How to Reinvent It for a Complex Global Economy (Oxford Univ. Press 2017).

Hannaford-Agor, Paula, Scott Graves & Shelley Spacek Miller, The Landscape of Civil Litigation in State Courts (National Center for State Courts 2015).

Hickman v. Taylor, 329 U.S. 495 (1947).

Katz, Daniel Martin, Michael James Bommarito, Shang Gao & Pablo Arredondo, GPT-4 Passes the Bar Exam, 382 Phil. Trans. R. Soc. A 20230254 (2024).

Legal Services Corporation, The Justice Gap: The Unmet Civil Legal Needs of Low-Income Americans (2022).

Liang, Tian, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi & Zhaopeng Tu, Encouraging Divergent Thinking in Large Language Models Through Multi-Agent Debate, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (2024).

Magesh, Varun, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools, Stanford HAI (2024).

Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023).

Morgan v. V2X, Inc., No. 25-cv-01991-SKC-MDB, 2026 WL 864223 (D. Colo. Mar. 30, 2026).

New York State Bar Association, Pro Se Advocacy in the AI Era: Benefits, Challenges, and Ethical Implications (2026).

Park v. Kim, 91 F.4th 610 (2d Cir. 2024).

Sandefur, Rebecca L., Access to What?, 148 Daedalus 49 (2019).

Shah, Anand V. & Joshua Y. Levy, Access to Justice in the Age of AI: Evidence from U.S. Federal Courts (Mar. 20, 2026) (unpublished manuscript), https://ssrn.com/abstract=6766859.

Sharma, Mrinank, et al., Towards Understanding Sycophancy in Language Models, ICLR (2024).

Simshaw, Drew, Access to A.I. Justice: Avoiding an Inequitable Two-Tiered System of Legal Services, 24 Yale J.L. & Tech. 150 (2022).

Surden, Harry, Machine Learning and Law, 89 Wash. L. Rev. 87 (2014).

System Documentation (on file with author).

U.S. Bureau of Labor Statistics, People with a Disability: Labor Force Characteristics — 2025, USDL-26-0364 (Mar. 3, 2026).

United States v. Heppner, No. 25-cr-503, 2026 WL 436479 (S.D.N.Y. Feb. 17, 2026).

Warner v. Gilbarco, Inc., No. 2:24-cv-12333, 2026 WL 373043 (E.D. Mich. Feb. 10, 2026).