From Chat to Research Infrastructure: How LLM Agents Are Changing the Analysis of Case Law

An experimental account of the NormaTrace project: model corpora of case law from 2018 and 2025, agentic workflows, and verifiable conclusions

A popular, educational walkthrough — no marketing, no tech jargon. This is a research, experimental project, not an industrial legal search system. Everything described below rests on the materials of one specific project; at the end you will find a list of exactly what was used and which examples from the project made it into the article.

Where it all begins: two prejudices about neural networks

Sitting down to talk with a large language model, a lawyer usually falls into one of two extremes. The first is enthusiasm: "the machine knows the law, it will analyze everything right now." The second, about ten minutes later, is disappointment: "it invented a case number, cited a ruling that doesn't exist — unfit for serious work."

It is telling that both reactions grow out of the same hidden premise — the idea that working with a model is a **conversation**: you ask, the model answers in prose. And that premise is the core misconception, because a "conversation with a model" has three flaws that are fatal for the law.

A chat answer is **unverifiable**: the quotation and the citation are pulled "out of the model's head" and may simply not exist. It is **unreproducible**: ask the same question tomorrow and you will get a different answer, with no record of the reasoning kept anywhere. And it is **unmeasurable**: the model will retell three or four cases it happens to know, but it will not tell you the main thing — *how the norm works across the entire body of practice*: what the share of dismissals is, where case law has split in two, which chamber least often "survives" in cassation.

Yet this — the measurable life of a norm — is the very core of scholarly, analytical, and legislative work. The researcher and the legislator need not a retelling but an answer: *does the norm produce the effect for which it was enacted, and if not — what is the defect and how do we fix it.*

The **NormaTrace** project, which is what this article is about, shows that this gap is already being bridged. But not through a "smarter model" — rather through a change of genre, from conversation to **infrastructure**.

First, let's agree on terms

So that we don't stumble later, let us explain a few terms in plain language. There is no avoiding them, but there are only a few.

LLM (large language model) — a program trained on vast volumes of text to predict the next word; out of this simple mechanism grows the ability to write coherently, summarize, classify, and reason. One thing is important to grasp: the model does not "remember" documents the way a library does — it stores statistical regularities of language. So when you ask it to quote a ruling, it does not pull it off a shelf but *plausibly reconstructs* it — and sometimes errs, feeling no difference between "I recalled it" and "I made it up." This is the famous **hallucination**.

Agentic workflow (agentic pipeline) — a way of organizing the model's work not as a single conversation but as a production process with stages, roles, and checkpoints. The difference is like that between asking a lawyer acquaintance for advice in the hallway and launching a study in an analytics department — with a plan, a division of tasks, source checking, internal peer review, and a signed report. Chat is the hallway; the agentic workflow is the department.

RAG / corpus search (from *retrieval-augmented generation*) — a technique in which the model first *re-*

trieves relevant documents from a large database and then reasons **only over the supplied text**, not from memory. The difference is like that between "tell me what you know about the case" and "here are fifty decisions, analyze them and do not go beyond them." The latter is the foundation of verifiability.

Citation engine — the "discipline of references" that physically prevents the model from inventing quotations; exactly how, we will unpack below.

Roles, skills, Human Review — the division of labor inside the pipeline: some agents search for sources, others analyze, others verify, a fourth one counts the money. **Human Review** is a review stage that imitates the eye of a fault-finding senior colleague.

Now we can get down to substance.

What it rests on: model corpora of case law

Let us state the project's status up front, so that no inflated expectations arise. NormaTrace is a **research, experimental environment**, not an industrial legal search system. It works with **model corpora of court decisions from 2018 and 2025** — experimental databases assembled to test hypotheses, refine methodology, and validate workflows. This is not a product search index and not a system designed for industrial load; it is a "research bench" on which the approach itself is tested.

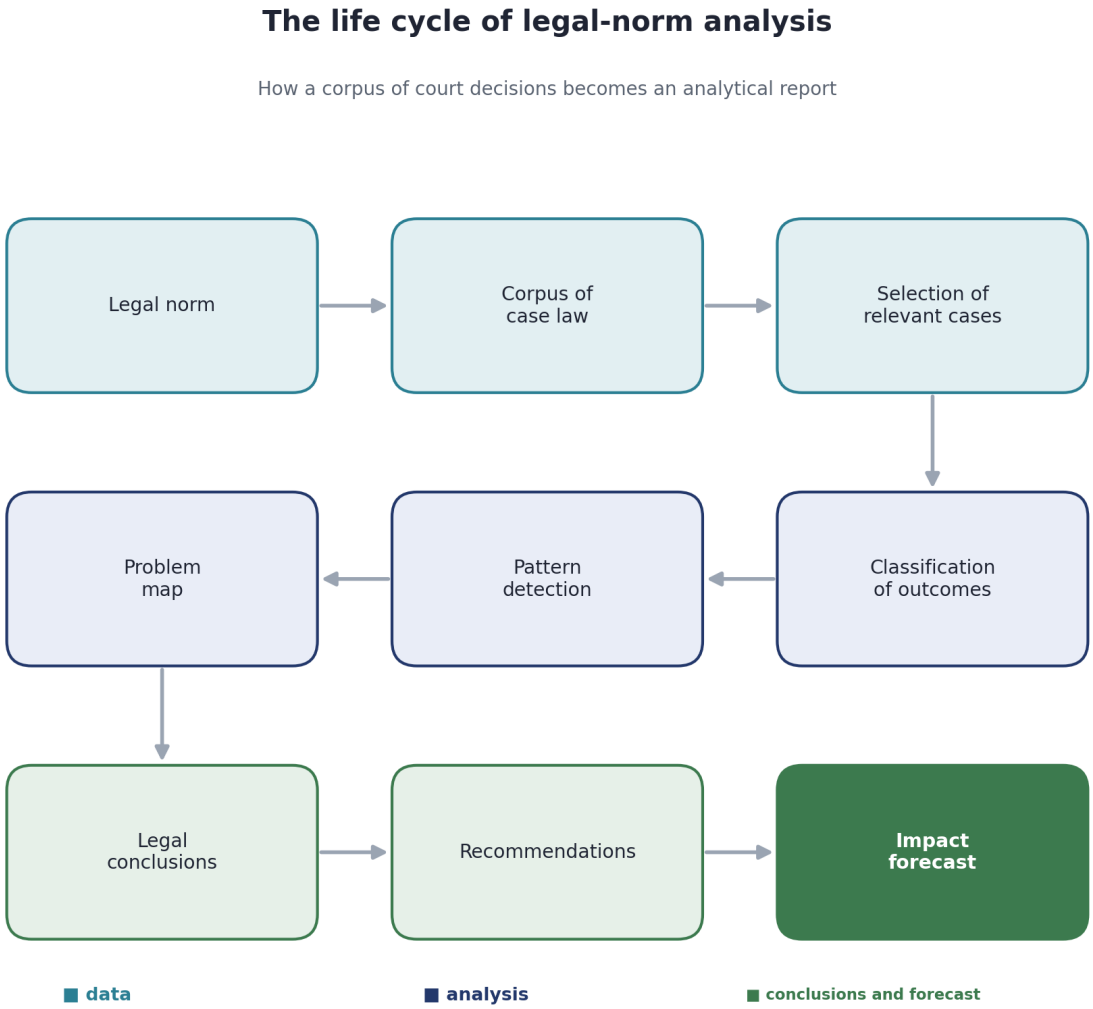
But even such a bench changes how the task is framed. A single year's corpus is already tens of thousands of decisions: for example, the case below involves 16,460 substantive rulings of the Supreme Court for 2025. Reading that many by hand is impossible — for a lawyer or for an entire department. And it is precisely this coverage that, for the first time, lets you ask questions that used to be purely rhetorical: *in what percentage of cases does the court dismiss under this norm? which argument most often becomes the reason for dismissal? does the practice of different cassation chambers agree?*

Here is an important and honest detail that the project's authors emphasize themselves: **the bottleneck is not the data but the methodology**. The texts are already collected; the difficulty is turning tens of thousands of raw decisions into a conclusion you can trust enough to put your signature under it. That is why the project's center of gravity is not "clever prompts" and not parsers, but discipline: a source map, an evidence matrix, a citation engine, and quality control.

The whole process is subordinated to a single formula, written into the project's control file (`CLAUDE.md`) and then repeated in a dozen places as a methodological invariant:

> **norm → application practice → patterns → problem → classification → solution → impact forecast → verifiable report.**

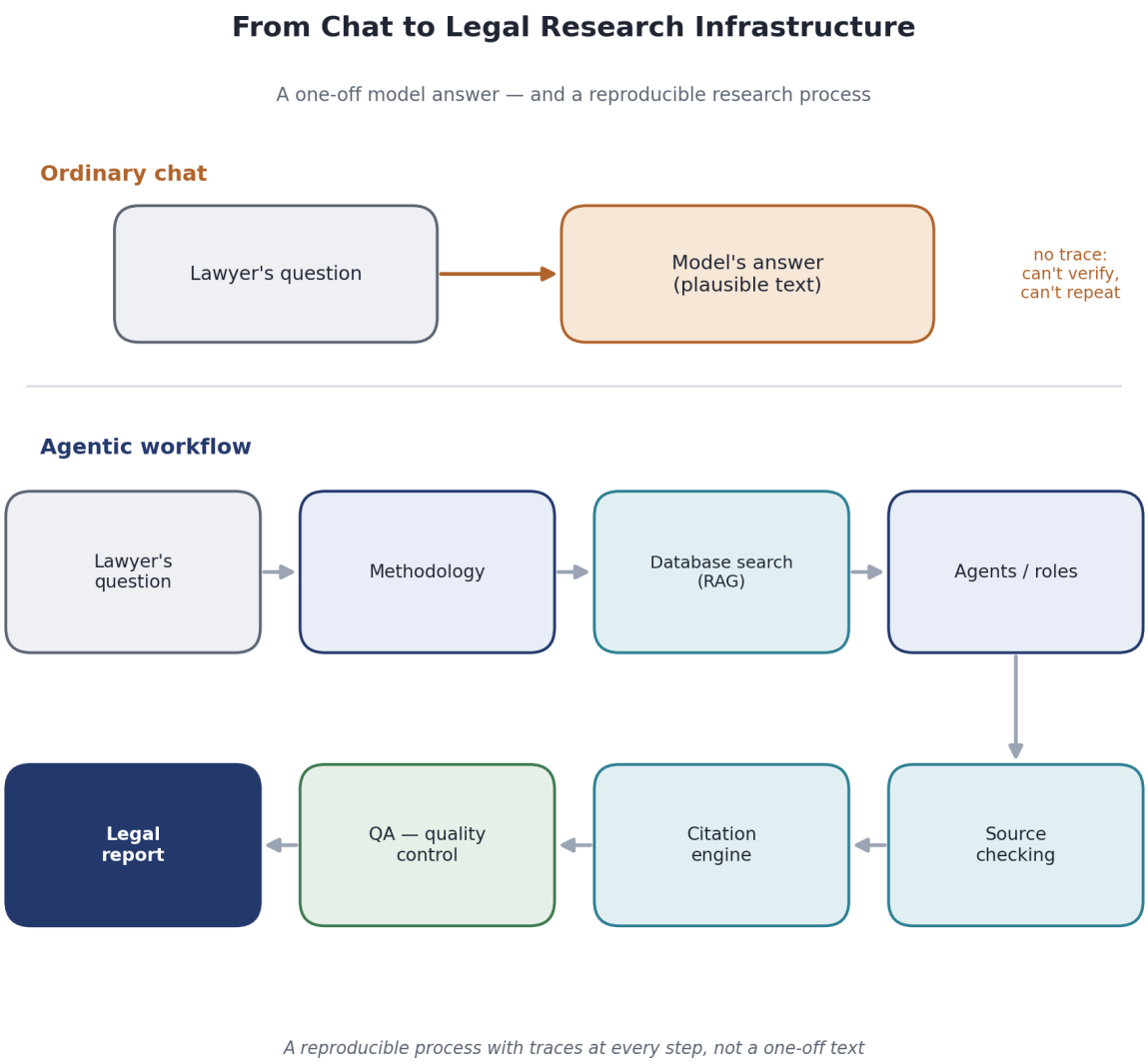
This is the "genre of infrastructure" in a single line. Not "answer me" but "walk the entire path from the text of the norm to a verifiable report, leaving traces at every step."



Infographic 2. The life cycle of legal-norm analysis. The path from the text of a norm and a corpus of practice, through the selection of relevant cases, the classification of outcomes, and the detection of patterns — to a problem map, legal conclusions, recommendations, and an impact forecast.

How an agentic pipeline differs from chat: the division of labor

In an ordinary chat, one model does everything at once — and therefore does nothing reliably. In Norma-Trace, the work is broken down into roles (in the project's language, *subagents*). There are about a dozen of them, and they are arranged thoughtfully.



Infographic 1. From chat to research infrastructure. An ordinary chat produces a one-off, unsupported answer; an agentic workflow takes the lawyer's question through methodology, database search, agent roles, source checking, the citation engine, and QA — and leaves a verifiable trace at every step.

There is a **coordinator** that plans the research and branches the work — and, tellingly, the right to "spawn" new performers is granted to it alone. There is an **analyst** with an iron rule: every conclusion must be tied to a row of the evidence matrix. There are **source specialists** (separately for the internal database and separately for the web) and a **bulk-request engineer**. And there are three "conscience" roles: the **billing guardian** — the only one allowed to pull the emergency brake; the **reproducibility auditor**, who checks whether every query and every quotation can be repeated; and the **autonomous "human" reviewer** — an agent playing the part of a fault-finding peer reviewer.

Alongside the roles are the **skills**: reusable checklist procedures ("how to cite," "how to count the money," "how to verify reproducibility"). If a role is the "who," then a skill is "by what regulation."

Why such fragmentation? For the same reason it exists in a real law firm. One person doing everything at once errs without noticing. The division of labor creates **friction and checks**: the one who analyzes is not the one who verifies; the one who spends tokens is not the one who watches the budget. One agent's error stands a chance of being caught by another — and it is precisely this artificial friction that distinguishes research from chatter.

How it worked on a live norm: subsidiary liability

Abstractions are better tested on specifics. The project's flagship case is Part 2 of Article 61 of the Code of Ukraine on Bankruptcy Procedures: the subsidiary (vicarious) liability of persons controlling the debtor ("controlling persons"). The norm is a "hot" one: when a bankrupt enterprise cannot pay its debts, the question arises whether those debts can be recovered from the people who actually ran it. The research question was framed empirically: *how is this norm actually applied by the courts in 2025?*

The pipeline walked the entire path. From the 2025 model corpus, about 260 candidates were selected by the norm's signature; after deduplication, 210 unique cases remained; after filtering out procedural and irrelevant ones — **70 "substantive" rulings** in which the norm was applied on the merits. The model then classified the outcomes. The results (strictly *as facts about the run's artifacts*, not as an independently re-verified legal result): **liability was denied in 60.0% of cases (42 of 70)**, and the dominant ground for denial was "causation not proven": **75.6% (34 of 45)** in the relevant subgroup. The whole run cost **$3.92** (666 calls to the "light" model).

It is worth pausing on this: this is not a retelling of three cases but a **map of the norm's behavior** across an entire year's corpus, from which a structural diagnosis emerges — the courts have no single, predictable test for causation, and it is precisely "unproven causation" that serves as the main channel through which liability leaks away. Such a conclusion cannot be obtained in a chat — it can only be *measured*.

This same storyline was also run a second, independent way — through a strict set of client prompts on a *grounded-only* principle ("reason only over the supplied text; do not use internal knowledge of the Code"). The two runs converged on a single diagnosis, and convergence by two independent paths strengthens confidence in the conclusion in itself.

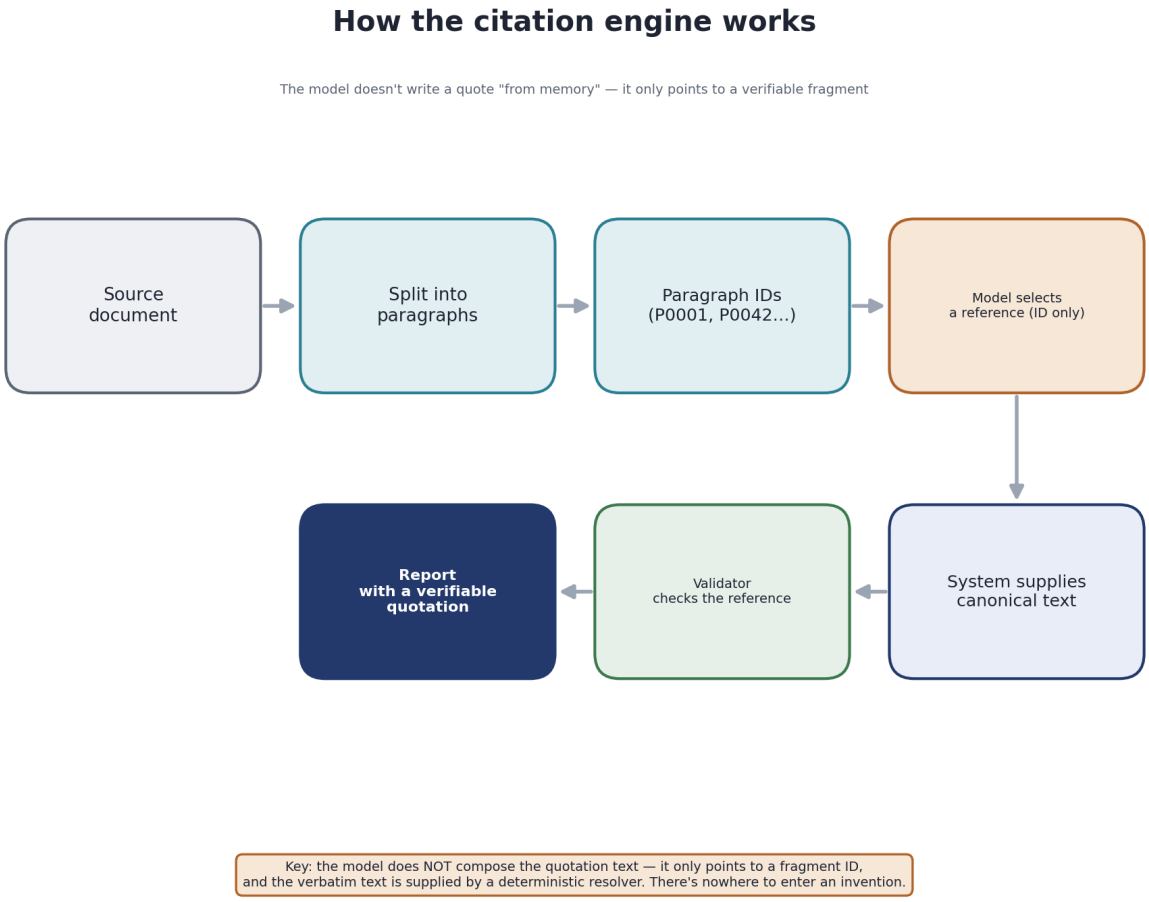
Citation engine: why the model here physically cannot lie in a quotation

Now — about the most important node, the one for which, in essence, the whole construction was undertaken. The lawyer's chief fear of a neural network: "it will cite a case that doesn't exist." NormaTrace solves this problem not by coaxing the model to "not make things up" but **architecturally**.

Here is how it works. All source documents are sliced in advance into numbered segments — each paragraph is assigned a stable identifier (something like `P0001`) and a hash fingerprint of the text. When the model prepares a conclusion, it is **forbidden to write the text of the quotation**: it may only *point to the identifier* of the segment it needs — "right here, in paragraph P0042." And the real, verbatim text is supplied not by the model but by a separate deterministic resolver that retrieves it from the registry by identifier.

In the data schema this is written down literally as a prohibition: *"the model selects only references; it MUST NOT compose the text of the quotation"* (`evidence_refs.schema.json`). There is simply no field for free-text quotation in the schema — there is nowhere to enter an invention.

The metaphor is simple. An ordinary model is a reteller, quoting from memory and occasionally fibbing. The citation engine turns it into someone who **points a finger at a line in a book**, while the book itself is held and read by someone else, incorruptible. Trust is shifted from the model to the deterministic layer.



Infographic 3. How the citation engine works. The document is split in advance into numbered paragraphs; the model selects only the identifier of a fragment, while the verbatim text is supplied by a deterministic resolver, after which a validator checks the reference. The model physically cannot "compose" a quotation — there is simply no field for free-text quotation in the schema.

And it works. Reviewing one of the final reports, an **independent external auditor** confirmed: all 29 references are in place, all 29 verbatim quotations match the canonical text, the checksums reconcile. Moreover, the auditor recomputed the key indicators from the primary files — "all values matched to the digit."

But here too is an honest boundary the project does not hide. The citation engine proves that the model pointed to an *existing* fragment of the *correct* document. It does **not** prove that the fragment is legally sufficient for the conclusion: the relevance of the legal position remains on the human's conscience. The reference is correct — but whether it is apt is decided by the lawyer.

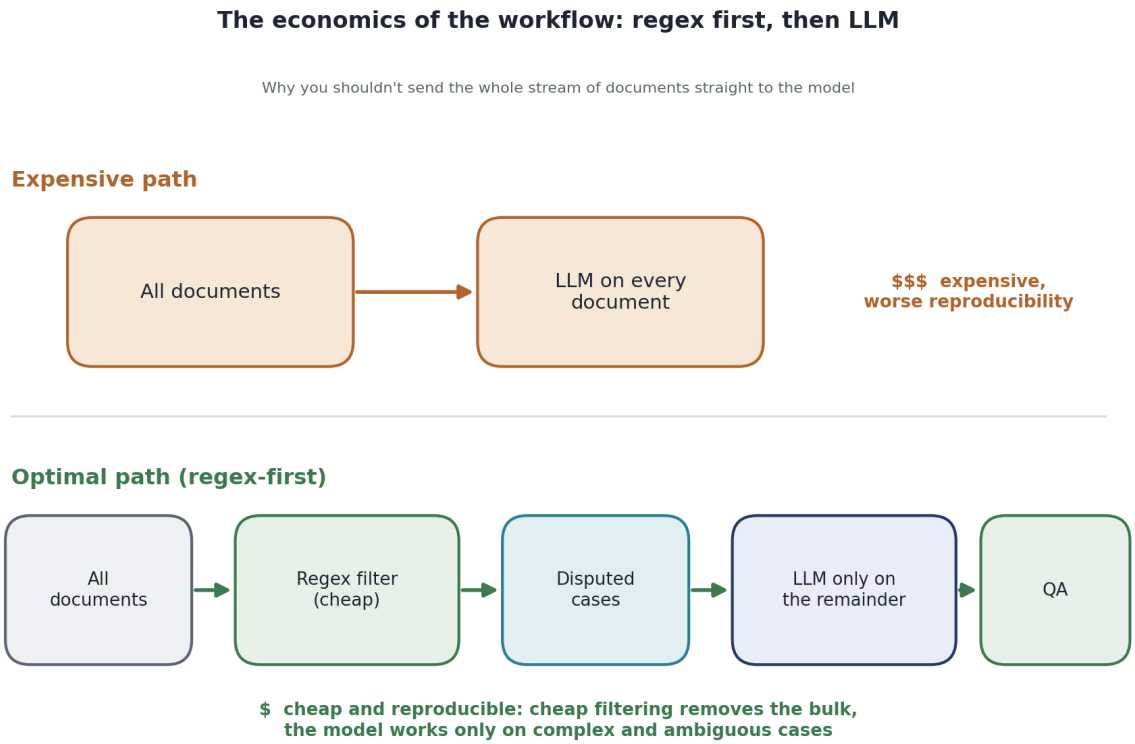
Control of errors, hallucinations, money, and quality

The verifiability of quotations is only one line of defense. There are several, and together they form the "discipline around the model."

Money. The only hard ground for stopping the entire system is the risk of exceeding the limit of **$50 per day** on model requests. Everything else (even legal classification and internal peer review) the system does autonomously, but it treats money as sacred: before a large run a preliminary cost estimate is mandatory, and if the price is unknown — stop.

And here is perhaps the project's most practical technique: **"regex first, then LLM."** Most of the repetitive work (for example, classifying the outcome from the standard wording of the operative part) is done not by an expensive model but by cheap deterministic pattern matching; the model is brought in only for the

"remainder" — the ambiguous cases — and for the final check. The effect is staggering: analyzing **16,460 rulings** of the Supreme Court cost roughly **$1.5** instead of an estimated **$233** had every decision been pushed through the model. A 150-fold saving — and at the same time **higher reproducibility**, since regex always yields the same result.



Infographic 4. The economics of the workflow: regex first, then LLM. Instead of running all documents through an expensive model, a cheap deterministic filter removes the bulk, and the model works only on the disputed cases; then comes QA. The figures are given as an example from an experimental run, not as a product metric.

Adversarial quality. The internal review is conducted by several independent reviewers, and every blocking comment is re-checked by a vote of skeptics. This is not decoration: in one run such a check caught a real overstatement — "8 of 8" instead of the correct "7 of 7" for one of the donor analogues — and the auditor then confirmed that the correction had been entered.

The main honest result. And now — the thing that distinguishes a mature project from a marketing pitch. When one of the flagship reports had passed internal control with a "sufficient" status, it was sent for an **independent external audit**, which returned a verdict: **"RETURN FOR REVISION"** — conditionally positive, with five blocking comments. The internal gate said "good enough," the external one said "not yet."

This discrepancy is perhaps the most valuable takeaway of the entire project: **in-model review does not replace external expertise.** The agent playing the part of a fault-finding reviewer is the same model; its independence is *declared but not architecturally guaranteed*. A genuine outside view, for now, remains with the human.

What else the experiments showed

Besides the flagship, the project ran several substantively different studies — and each adds a stroke to the picture.

Scale and regex-first: cassation stability. The question — what share of lower-court decisions "survives" in the cassation of the Supreme Court in 2025. The method — deterministic parsing of operative-part wordings against a taxonomy of 16 codes with **98%** coverage; the model only checked the sample for control. The result: **58.2% of decisions survived, 41.8% were destabilized**, with a wide spread across chambers — from 68.2% in the Grand Chamber to **43.2% in administrative cassation**. The agreement of the model check with regex was almost 95%. What would have taken months by hand took minutes and about a dollar and a half.

A case of methodological caution: reporting judges. On the already-annotated corpus, stability was computed for individual reporting judges (this cost nothing at all — the outcomes were joined to the ready data). But more valuable than the figures here is the **warning** the system itself wrote into the report: the "stability rate" indicator **must not** be used as an assessment of a judge's quality or a basis for disciplinary conclusions without adjusting for case mix. Complex categories of cases collapse in cassation more often — and that speaks to the category, not the judge. A system that itself limits the applicability of its metrics deserves far more trust than one that briskly issues rankings.

A case of supranational practice: the ECtHR via HUDOC. For comparative-law argumentation, the project extracted 79 positions of the European Court of Human Rights from 8 cases, each with a precise paragraph reference. And again, what is telling is not the size but the discipline: unverifiable cases (an unestablished application number, text only in French) were **excluded**, while early "from memory" quotations were caught by the QA loop and replaced with verifiable ones. This is a direct prototype of preparing materials for international tribunals.

Micro-pilot. Before running large corpora, the bulk-processing path was tried out on eight synthetic documents: correct output format — 8 of 8, selection accuracy — 8 of 8, cost — less than a cent. Boring — but this is exactly what engineering care looks like: first test the pipe with water, then let the flow through.

Where the system still stumbles

The project's honesty lies in documenting its limitations as thoroughly as its achievements. Let us list the main ones — this matters more than any enthusiasm.

"Spurious precision." Reading "60.0% dismissals" or "75.6% on causation," it is easy to believe in precision to the tenth of a percent. But behind these figures stands **a single pass** of an inexpensive model (average confidence around 0.7), and not all cases were verbatim-checked. The correct way to present this is **ranges** ("about 60%"), not tenths of a percent, plus a manual audit of a sample of 20–30 rulings. A figure that looks precise and a figure that is precise are not the same thing.

A thin base of analogies. To propose a solution "as in a similar construction," you need those constructions in the first place. But a full review of 220 legal positions yielded only **2** genuine comparators — meaning that some comparative-law conclusions rest more on a *shortage* of material than on a reliably transferred standard. The risk of false analogy is real; the project honestly records it but does not eliminate it.

Completeness and quality of the data. Some metadata is sparse, jurisdiction is in places inferred from indirect signs (from the court's name), and the corpus for a particular norm may be incomplete (263 cases where there could hypothetically have been ~1000). Verbatim anchoring of quotations was not done for all sources — the dynamic pages of official legal portals cannot simply be "pinned down."

The boundary of autonomy. The $50 limit was observed (by the eighth task about $33 of 50 had accrued), but there is no single consolidated cost counter — the sums are gathered from separate files and diverge slightly. And more subtly: the limit counts only requests to the model, while **the work of the orchestrator itself is not included in it** — the full compute cost is higher than the one recorded.

The maturity of the "product" layer. On top of the research core, an interface is being built through which a lawyer could pull the pipeline with a natural-language query. But this is so far a **demonstration, not a product**: authorization is a stub, some adapters are "dummies," the deployment-readiness check returned "failed." It cannot be rolled into production.

And a separate caveat, important for accuracy: the model names in the project ("GPT-5.4," "Mini") are **routing roles** within the methodology, not confirmed external identifiers; what physically runs under them depends on the environment.

Why this is already useful to a lawyer — and where there is no doing without a human

If you set aside both the enthusiasm and the skepticism, what remains is a sober picture: such systems **already** give a lawyer what neither chat nor classical legal search ever did.

They provide **coverage** unattainable by hand: not a sample of a dozen cases picked "by feel," but a reproducible run across an entire year's corpus. They show the **structure of practice, not anecdotes**: where one approach dominates, where practice is split, which chamber is least stable — this is material for a scholarly article, for a brief to a court, and for justifying a legislative initiative. They help **formulate and test hypotheses** about the cause of a norm's dysfunction: the system not only counts but classifies the defect (a gap, a conflict of laws, a defect of legislative technique, a systemic inconsistency). And, finally, they produce **reports "ready for signature,"** in which every conclusion is tied to a row of the evidence matrix and to a verifiable quotation, while the appendix holds everything needed to repeat the run — the run identifier, prompt versions, queries, cost. This removes the chief objection to AI in the law — "and where did you get that from?"

But this same picture outlines **the human's place**. The machine measures — the human decides whether the fragment is *sufficient* for a legal conclusion. The machine finds a pattern — the human assesses whether it is an artifact of the method (like that very "2 of 220"). The machine issues a percentage — the human translates it into an honest range, remembering that behind the tenths a single uncertain pass of the model may be hiding. And, as the external audit showed, it is precisely the human who remains the final instance of quality: internal review is a useful sieve, but not an expert's signature.

Hence the main thesis, without pathos. Value is created **not by the model in itself, but by the discipline around it**: methodology, structured prompts, verifiable sources, reproducible runs, quality control, Human Review, and an honest description of the boundaries. Remove the discipline — and the infrastructure collapses back into a talkative chat.

What is worth researching further

NormaTrace is also useful in that it clearly shows the front of work ahead. Before taking such conclusions into a publication or into a court, it is worth: conducting a **manual audit of a sample** of single-pass classifications, confirming the outcome percentages; **independently checking the thin base of analogies** (is "2 of 220" a property of the practice or an artifact of the method?); **verifying the completeness of the corpus** for the norm; **computing the full cost**, including the work of the orchestrator. A separate, large storyline is how to make external review architecturally independent rather than declaratively so, and how to bring the demonstration product layer up to something fit for real work.

This is not a list of grievances but a map of a technology coming of age — one that has already ceased to be a toy but has not yet become an instrument you can trust without looking. And perhaps the healthiest thing one can say about LLM agents in the law today: they are good enough to radically empower a lawyer, and not good enough to replace one — and honest systems like NormaTrace are valuable precisely because they show *where* that boundary runs.

Appendix: materials and examples used

Which project materials were used. The article draws on the research brief for the NormaTrace project (a meta-analysis of the project's materials dated June 8, 2026) and, through it, on verifiable artifacts: the control file `CLAUDE.md` (the methodological invariant, the 50 USD/day limit); the citation-engine schema `schemas/evidence_refs.schema.json` (the prohibition on composing quotation text); the registry of roles and skills in `.claude/agents/` and `.claude/skills/`; the report on the structure of the research corpus `reports/court_decisions_schema_analysis.md`; the runs in `artifacts/runs/` (the flagship on Part 2 of Article 61 of the Bankruptcy Procedures Code, the stability of Supreme Court cassation, reporting judges, judicial/High Council of Justice discipline, the micro-pilot); the conclusion of the independent external audit `reports/audit/audit_conclusion_kuzpb_st61_v5.md`; the readiness assessments and cost estimates in the corresponding `readiness_brief.md` and `cost_report.md`.

Which examples from NormaTrace made it into the article. (1) The subsidiary liability of controlling persons under Part 2 of Article 61 of the Bankruptcy Procedures Code as a "deep case" with measured outcomes (60.0% dismissals; 75.6% "causation not proven"; the funnel 260→210→70; cost $3.92) and the evolution of the report under external audit. (2) The stability of Supreme Court cassation in 2025 as a "case of scale and regex-first" (16,460 rulings; 58.2% stability; chamber spread 68.2%→43.2%; ~$1.49 versus an estimated ~$233). (3) Reporting judges as a "case of methodological caution" (a warning about the metric's inapplicability; $0). (4) ECtHR positions via HUDOC as a "case of verifiable citation" (79 positions from 8 cases; exclusion of unverifiable ones). (5) The external audit with a "return for revision" verdict as an illustration of the divergence between internal and external control.