Від чату до дослідницької інфраструктури: як LLM-агенти змінюють аналіз судової практики

Експериментальний досвід проєкту NormaTrace: модельні корпуси судової практики за 2018 і 2025 роки, агентські workflow та перевірювані висновки

Популярно-просвітницький розбір без реклами й без техножаргону. Ідеться про дослідницький, експериментальний проєкт — а не про промислову правову пошукову систему. Усе, що описано нижче, спирається на матеріали конкретного проєкту; наприкінці перелічено, що саме використано і які приклади з проєкту увійшли до статті.

З чого все починається: два упередження щодо нейромережі

Сідаючи за розмову з великою мовною моделлю, юрист зазвичай впадає в одну з двох крайнощів. Перша — захват: «машина знає право, зараз усе проаналізує». Друга, хвилин за десять, — розчарування: «вигадала номер справи, послалася на неіснуючу постанову — для серйозної роботи непридатна».

Цікаво, що обидві реакції виростають з однієї прихованої засновки — нібито робота з моделлю це **розмова**: ви запитуєте, модель відповідає текстом. І ця засновка — головна омана, бо в «розмови з моделлю» є три вади, фатальні для права.

Відповідь у чаті **неперевірювана**: цитата й посилання взяті «з голови» моделі й можуть просто не існувати. Вона **невідтворювана**: поставте те саме питання завтра — отримаєте іншу відповідь, а хід міркування ніде не збережеться. І вона **невимірна**: модель перекаже три-чотири відомі їй справи, але не скаже головного — *як норма працює на всьому масиві практики*: яка частка відмов, де практика розкололася надвоє, яка палата найрідше «встоює» в касації.

А саме це — вимірне життя норми — і є ядром наукової, аналітичної та законопроєктної роботи. Дослідникові й законодавцеві потрібен не переказ, а відповідь: *чи дає норма той ефект, заради якого її ухвалювали, і якщо ні — у чому дефект і як його лагодити.*

Проєкт **NormaTrace**, про який ітиметься, показує: цей розрив уже долається. Але не за рахунок «розумнішої моделі», а за рахунок зміни жанру — від розмови до **інфраструктури**.

Спершу домовимося про слова

Щоб далі не спотикатися, пояснимо кілька термінів простою мовою. Без них не обійтися, але їх небагато.

LLM (велика мовна модель) — програма, навчена на величезних обсягах тексту передбачати наступне слово; із цього простого механізму виростає здатність зв’язно писати, узагальнювати, класифікувати й міркувати. Важливо зрозуміти одне: модель не «пам’ятає» документи як бібліотека — вона зберігає статистичні закономірності мови. Тому, коли ви просите процитувати постанову, вона не дістає її з полиці, а *правдоподібно реконструює* — і часом помиляється, не відчуваючи різниці між «згадав» і «вигадав». Це і є знаменита **галюцинація**.

Агентський workflow (агентський конвеєр) — спосіб організувати роботу моделі не як одну розмову, а як виробничий процес з етапами, ролями й контрольними точками. Різниця як між тим, щоб запитати поради у знайомого юриста в коридорі, і тим, щоб запустити дослідження в аналітичному відділі — з планом, розподілом завдань, перевіркою джерел, внутрішньою рецензією та підписаним звітом. Чат — це коридор; агентський workflow — відділ.

RAG / пошук по корпусу (від *retrieval-augmented generation* — «генерація з опорою на знайдене») — прийом, за якого модель спершу *знаходить* релевантні документи у великій базі, а потім міркує

лише за поданим текстом, а не за пам’яттю. Різниця як між «розкажи, що ти знаєш про справу» і «ось тобі п’ятдесят рішень, проаналізуй і не виходь за їхні межі». Друге — основа перевірюваності.

Citation engine (механізм цитування) — «дисципліна посилань», яка фізично не дає моделі вигадувати цитати; як саме — розберемо нижче.

Ролі, скіли, Human Review — поділ праці всередині конвеєра: одні агенти шукають джерела, інші аналізують, треті перевіряють, четвертий рахує гроші. **Human Review** («людська перевірка») — етап рецензування, що імітує погляд прискіпливого старшого колеги.

Тепер можна говорити по суті.

Що лежить в основі: модельні корпуси судової практики

Одразу позначимо статус проєкту, щоб не виникало завищених очікувань. NormaTrace — це **дослідницьке, експериментальне середовище**, а не промислова правова пошукова система. Воно працює з **модельними корпусами судових рішень за 2018 і 2025 роки** — експериментальними базами, зібраними для перевірки гіпотез, відпрацювання методології й тестування workflow. Це не продуктовий пошуковий індекс і не система, розрахована на промислове навантаження; це «дослідницький стенд», на якому перевіряють сам підхід.

Але навіть такий стенд змінює постановку задачі. Один річний корпус — це вже десятки тисяч рішень: наприклад, у кейсі нижче фігурують 16 460 істотних постанов Верховного Суду за 2025 рік. Прочитати стільки руками неможливо — ні юристові, ні цілому відділу. А саме охоплення вперше дозволяє ставити питання, які раніше були суто риторичними: *у якому відсотку справ суд відмовляє за цією нормою? який аргумент найчастіше стає причиною відмови? чи збігається практика різних касаційних палат?*

Тут важлива й чесна деталь, яку автори проєкту підкреслюють самі: **вузьке місце — не дані, а методологія**. Тексти вже зібрані; складність у тому, щоб перетворити десятки тисяч сирих рішень на висновок, якому можна довіряти настільки, щоб поставити під ним підпис. Тому центр ваги проєкту — не «розумні промпти» і не парсери, а дисципліна: карта джерел, матриця доказів, механізм цитування та контроль якості.

Увесь процес підпорядкований одній формулі, прописаній у керівному файлі проєкту (`CLAUDE.md`) і повтореній згодом у десятку місць як методологічний інваріант:

> **норма → практика застосування → патерни → проблема → класифікація → рішення → прогноз впливу → перевірюваний звіт.**

Це і є «жанр інфраструктури» в один рядок. Не «дай мені відповідь», а «пройди весь шлях від тексту норми до перевірюваного звіту, залишаючи сліди на кожному кроці».

Інфографіка 2. Життєвий цикл аналізу правової норми. Шлях від тексту норми та корпусу практики через відбір релевантних справ, класифікацію результатів і виявлення патернів — до проблемної карти, юридичних висновків, рекомендацій і прогнозу впливу.

Чим агентський конвеєр відрізняється від чату: поділ праці

У звичайному чаті одна модель робить усе одразу — і тому не робить нічого надійно. У NormaTrace роботу розкладено на ролі (мовою проєкту — *субагентів*). Їх близько дюжини, й влаштовані вони продумано.

Інфографіка 1. Від чату — до дослідницької інфраструктури. Звичайний чат видає разову, нічим не підкріплену відповідь; агентський workflow проводить питання юриста через методологію, пошук по базі, ролі-агенти, перевірку джерел, citation engine і QA — і залишає перевірюваний слід на кожному кроці.

Є **координатор**, який планує дослідження й розгалужує роботу, — і характерно, що право «породжувати» нових виконавців надано лише йому. Є **аналітик** із залізним правилом: кожен висновок зобов'язаний бути прив'язаний до рядка матриці доказів. Є **спеціалісти з джерел** (окремо по внутрішній базі, окремо по інтернету) та **інженер масових запитів**. І є три ролі-«совісті»: **білінг-контролер** — єдиний, кому дозволено натиснути стоп-кран; **аудитор відтворюваності**, який перевіряє, чи можна повторити кожен запит і кожну цитату; і **автономний «людський» рецензент** — агент, що грає роль прискіпливого рецензента.

Поряд із ролями — **скіли**: повторно використовувані процедури-чеклісти («як цитувати», «як рахувати гроші», «як перевіряти відтворюваність»). Якщо роль — це «хто», то скіл — «за яким регламентом».

Навіщо таке дроблення? Затим же, навіщо воно існує в реальній юрфірмі. Одна людина, яка робить усе одразу, помиляється непомітно для себе. Поділ праці створює **тертя і перевірки**: той, хто аналізує, не той, хто перевіряє; той, хто витрачає токени, не той, хто стежить за бюджетом. Помилка одного агента має шанс бути спійманою іншим — і саме ці штучні тертя відрізняють дослідження від балаканини.

Як це працювало на живій нормі: субсидіарна відповідальність

Абстракції краще перевіряти на конкретиці. Флагманський кейс проєкту — частина 2 статті 61 Кодексу України з процедур банкрутства: субсидіарна відповідальність осіб, що контролюють боржника (КДЛ). Норма «гаряча»: коли збанкрутіле підприємство не може розрахуватися, постає питання, чи можна стягнути борги з тих, хто ним фактично керував. Питання дослідження звучало емпірично: *як ця норма реально застосовується судами у 2025 році?*

Конвеєр пройшов увесь шлях. З модельного корпусу 2025 року за сигнатурою норми відібрали близько 260 кандидатів; після дедуплікації залишилося 210 унікальних справ; після відсіву процесуальних і нерелевантних — **70 «істотних» актів**, де норма застосовувалася по суті. Далі модель класифікувала результати. Результати (суворо *як факти про артефакти прогону*, а не як самостійно переперевірений правовий результат): **відмова у притягненні до відповідальності — у 60,0% справ (42 з 70)**, а домінівна причина відмови — «причинний зв’язок не доведено»: **75,6% (34 з 45)** у відповідній підгрупі. Увесь прогін коштував **3,92 долара** (666 звернень до «легкої» моделі).

Варто вдуматися: це не переказ трьох справ, а **карта поведінки норми** на корпусі цілого року, з якої видно структурний діагноз — у судів немає єдиного передбачуваного критерію причинного зв’язку, і саме «недоведеність причинності» працює головним каналом, яким відповідальність утікає. Такий висновок не можна отримати в чаті — його можна лише *виміряти*.

Цей самий сюжет пройшли і другим, незалежним способом — за суворим набором клієнтських промптів із принципом *grounded-only* («міркуй лише за поданим текстом, внутрішні знання про Кодекс не використовуй»). Два прогони зійшлися до одного діагнозу, а сходження двома незалежними шляхами саме по собі зміцнює довіру до висновку.

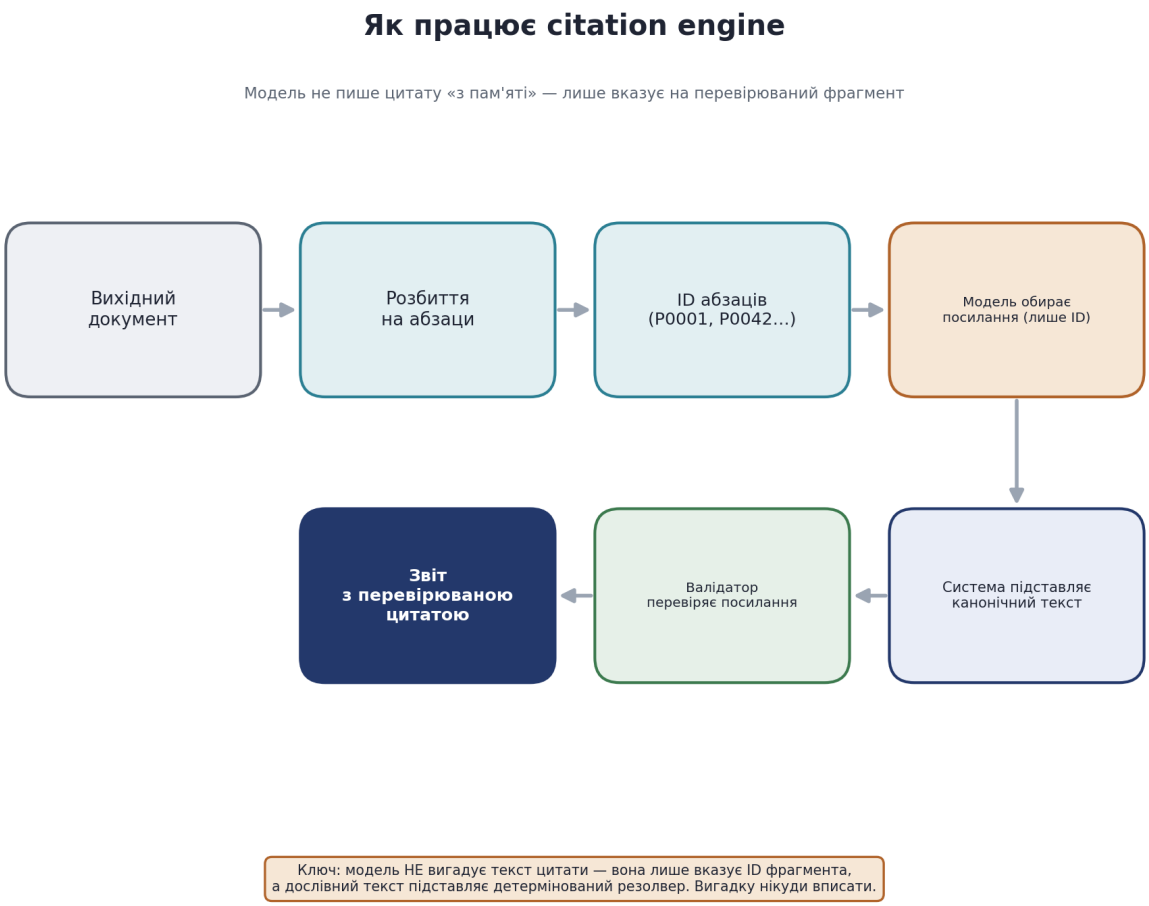
Citation engine: чому модель тут фізично не може збрехати в цитаті

Тепер — про найважливіший вузол, заради якого, по суті, й затівалася вся конструкція. Головний страх юриста перед нейромережею: «вона пошлеться на справу, якої немає». NormaTrace розв’язує цю проблему не вмовляннями моделі «не вигадуй», а **архітектурно**.

Працює це так. Усі вихідні документи заздалегідь нарізаються на пронумеровані сегменти — кожному абзацу присвоюється стабільний ідентифікатор (на кшталт `P0001`) і хеш-відбиток тексту. Коли модель готує висновок, їй **заборонено писати текст цитати**: вона може лише *вказати на ідентифікатор* потрібного сегмента — «ось тут, в абзаці P0042». А реальний, дослівний текст підставляє вже не модель, а окремий детермінований резолвер, що дістає його з реєстру за ідентифікатором.

У схемі даних це записано буквально як заборона: *«модель обирає лише посилання; вона НЕ ПОВИННА вигадувати текст цитати»* (`evidence_refs.schema.json`). Поля для вільного тексту цитати в схемі просто немає — вигадку нікуди вписати.

Метафора проста. Звичайна модель — переказувач, який цитує з пам’яті й часом прибріхує. Citation engine перетворює її на того, хто **показує пальцем на рядок у книзі**, а саму книгу тримає й читає хтось інший, непідкупний. Довіра переноситься з моделі на детермінований шар.



Інфографіка 3. Як працює citation engine. Документ заздалегідь розбивається на пронумеровані абзаци; модель обирає лише ідентифікатор фрагмента, а дослівний текст підставляє детермінований резолвер, після чого посилання перевіряє валідатор. Модель фізично не може «вигадати» цитату — поля для вільного тексту цитати в схемі просто немає.

І це працює. Перевіряючи один із підсумкових звітів, **незалежний зовнішній аудитор** підтвердив: усі 29 посилань на місці, усі 29 дослівних цитат збігаються з канонічним текстом, контрольні суми сходяться. Ба більше, аудитор заново перерахував ключові показники з первинних файлів — «усі значення збіглися до цифри».

Але тут же — чесна межа, яку проєкт не приховує. Citation engine доводить, що модель указала на *наявний* фрагмент *правильного* документа. Він **не** доводить, що фрагмент юридично достатній для висновку: релевантність правової позиції лишається на совісті людини. Посилання правильне — але чи доречне воно, вирішує юрист.

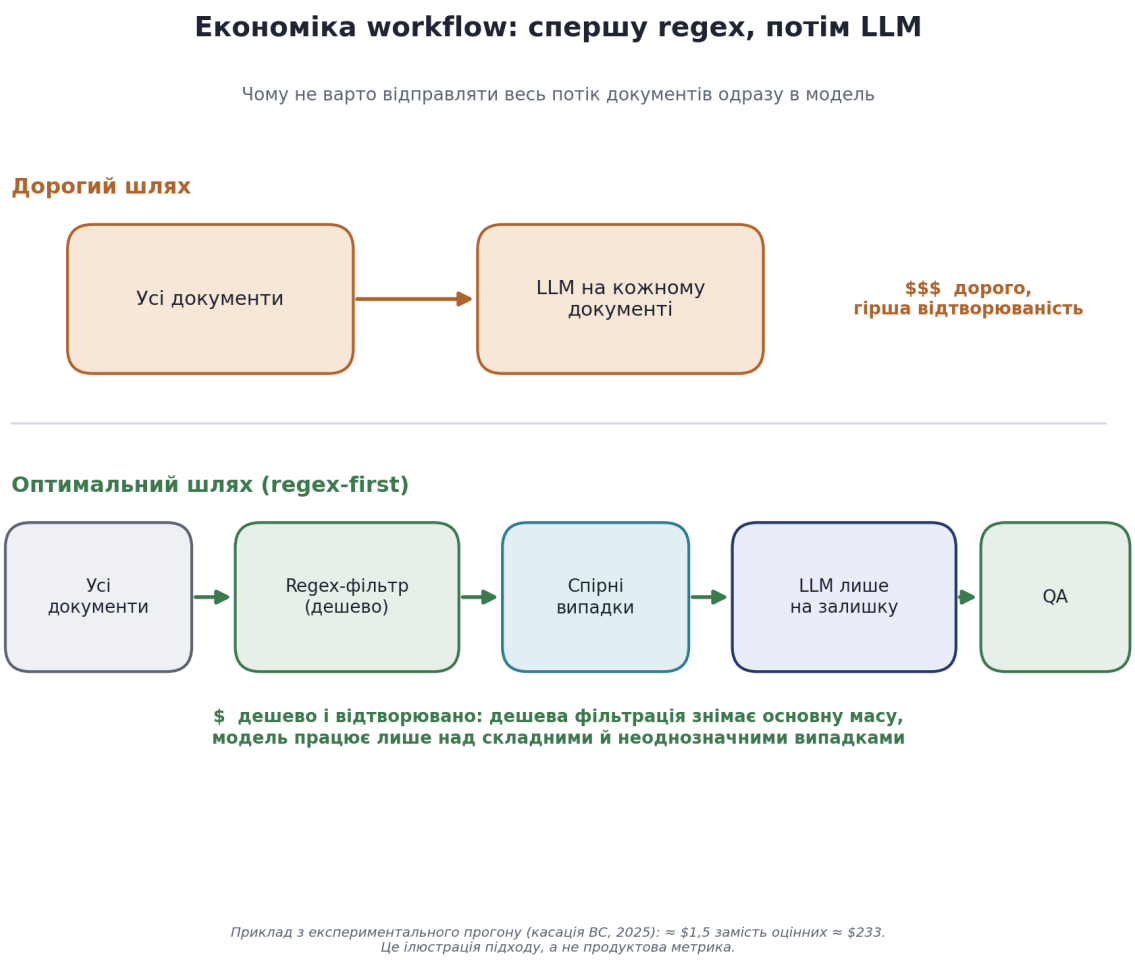
Контроль помилок, галюцинацій, грошей і якості

Перевірюваність цитат — лише один контур захисту. Їх кілька, і разом вони й утворюють «дисципліну навколо моделі».

Гроші. Єдина жорстка підстава зупинити всю систему — ризик перевищити ліміт **50 доларів на добу** на модельні запити. Усе інше (навіть правову класифікацію і внутрішню рецензію) система робить автономно, але до грошей ставиться свято: перед великим прогоном обов’язковий попередній розрахунок вартості, а якщо ціна невідома — стоп.

І тут — мабуть, найпрактичніший прийом проєкту: **«спершу regex, потім LLM»**. Більшу частину однотипної роботи (наприклад, класифікацію результату за стандартним формулюванням

резолютивної частини) робить не дорога модель, а дешевий детермінований пошук за шаблоном; модель підключається лише до «залишку» — неоднозначних випадків — і до фінальної перевірки. Ефект приголомшливий: аналіз **16 460 постанов** Верховного Суду обійшовся приблизно у **1,5 долара** замість оцінних **233**, якби кожне рішення гнали через модель. Економія в півтораста разів — і заодно **вища відтворюваність**, адже regex завжди дає той самий результат.



Інфографіка 4. Економіка workflow: спершу regex, потім LLM. Замість того щоб проганяти через дорогу модель усі документи, дешевий детермінований фільтр знімає основну масу, а модель працює лише над спірними випадками; потім — QA. Цифри наведено як приклад з експериментального прогону, а не як продуктову метрику.

Змагальна якість. Внутрішню рецензію ведуть кілька незалежних рецензентів, і кожне блокувальне зауваження переперевіряється голосуванням скептиків. Це не декорація: в одному прогоні така перевірка спіймала реальне завищення — «8 з 8» замість коректних «7 з 7» по одному з донорів-аналогів, і аудитор потім підтвердив, що виправлення внесли.

Головний чесний результат. А тепер — те, що відрізняє зрілий проєкт від рекламної презентації. Коли один із флагманських звітів пройшов внутрішній контроль зі статусом «достатньо», його віддали на **незалежний зовнішній аудит**, і той виніс вердикт: **«ПОВЕРНУТИ НА ДООПРАЦЮВАННЯ»** — умовно-позитивно, з п’ятьма блокувальними зауваженнями. Внутрішній гейт сказав «годиться», зовнішній — «ще ні».

Це розходження — мабуть, найцінніший висновок усього проєкту: **внутрішньомодельна рецензія не замінює зовнішню експертизу.** Агент, який грає роль прискіпливого рецензента, — це та сама модель; її незалежність *декларується, але не забезпечена архітектурно*. Справжній погляд збоку

поки що лишається за людиною.

Що ще показали експерименти

Окрім флагмана, проєкт прогнав ще кілька змістовно різних досліджень — і кожне додає штрих до картини.

Масштаб і regex-first: стабільність касації. Питання — яка частка рішень нижчих судів «встоює» в касації Верховного Суду за 2025 рік. Метод — детермінований розбір резолютивних формулювань за таксономією з 16 кодів з покриттям **98%**; модель лише контрольно перевіряла вибірку. Результат: **58,2% рішень встояло, 41,8% дестабілізовано**, причому розкид по палатах великий — від 68,2% у Великій Палаті до **43,2% в адміністративній касації**. Згода модельної перевірки з regex — майже 95%. Те, що вручну потребувало б місяців, зайняло хвилини і близько півтора долара.

Кейс методологічної обережності: судді-доповідачі. На вже розміченому корпусі порахували стабільність по окремих суддях-доповідачах (це не коштувало взагалі нічого — результати приєднали до готових даних). Але цінніше за цифри тут **попередження**, яке система сама вписала у звіт: показник «відсоток стабільності» **не можна** використовувати як оцінку якості судді або підставу для дисциплінарних висновків без поправки на склад справ. Складні категорії справ руйнуються в касації частіше — і це говорить про категорію, а не про суддю. Система, яка сама обмежує застосовність своїх метрик, заслуговує куди більшої довіри, ніж та, що жваво видає рейтинги.

Кейс наднаціональної практики: ЄСПЛ через HUDOC. Для порівняльно-правової аргументації проєкт видобув 79 позицій Європейського суду з прав людини з 8 справ, кожну з точним зазначенням параграфа. І знову показова не величина, а дисципліна: неперевірювані справи (невстановлений номер заяви, текст лише французькою) були **виключені**, а ранні цитати «з пам’яті» QA-петля спіймала й замінила на перевірювані. Це прямий прообраз підготовки матеріалів для міжнародних інстанцій.

Мікропілот. Перш ніж ганяти великі корпуси, шлях масової обробки обкатали на восьми синтетичних документах: коректний формат на виході — 8 з 8, точність відбору — 8 з 8, вартість — менше цента. Нудно — але саме так виглядає інженерна акуратність: спершу перевір трубу на воді, потім пускай потік.

Де система поки що спотикається

Чесність проєкту в тому, що свої обмеження він документує так само ретельно, як досягнення. Перелічимо головні — це важливіше за будь-які захвати.

«Уявна точність». Читаючи «60,0% відмов» або «75,6% по причинності», легко повірити в точність до десятих часток. Але за цими цифрами стоїть **один прохід** недорогої моделі (середня впевненість близько 0,7), і дослівно звірені не всі випадки. Коректний спосіб подачі — **діапазони** («близько 60%»), а не десяті частки, плюс ручний аудит вибірки у 20–30 актів. Цифра, що виглядає точною, і цифра точна — не одне й те саме.

Тонка база аналогій. Щоб запропонувати рішення «як у схожій конструкції», потрібні самі ці конструкції. Але суцільний перегляд 220 правових позицій дав лише **2** справжні компаратори — тобто частина порівняльно-правових висновків спирається радше на *дефіцит* матеріалу, ніж на надійно перенесений стандарт. Ризик хибної аналогії реальний; проєкт його чесно фіксує, але не усуває.

Повнота і якість даних. Частина метаданих розріджена, юрисдикція місцями виводиться за непрямими ознаками (за назвою суду), а корпус за конкретною нормою може бути неповним (263 справи там, де гіпотетично могла бути ~тисяча). Дослівна прив’язка цитат зроблена не для всіх джерел — динамічні сторінки офіційних правових порталів так просто не «пришпилити».

Межа автономії. Ліміт у 50 доларів дотримувався (до восьмого завдання набігло близько 33 з 50), але

єдиного зведеного лічильника витрат немає — суми збираються з окремих файлів і трохи розходяться. І тонше: ліміт рахує лише запити до моделі, а **робота самого оркестратора в нього не входить** — повна обчислювальна ціна вища за відображену.

Зрілість «продуктового» шару. Поверх дослідницького ядра будується інтерфейс, через який юрист міг би смикати конвеєр запитом природною мовою. Але це поки що **демонстрація, а не продукт**: авторизація — заглушка, частина адаптерів — «пустушки», перевірка готовності до розгортання поверталася «провалено». Викочувати в продакшен не можна.

І окреме застереження, важливе для точності: назви моделей у проєкті («GPT-5.4», «Mini») — це **ролі маршрутизації** всередині методології, а не підтверджені зовнішні ідентифікатори; що під ними фізично працює, залежить від оточення.

Чому це вже корисно юристові — і де без людини не обійтися

Якщо відкинути і захват, і скепсис, лишається тверезa картина: такі системи **вже** дають юристові те, чого не давали ні чат, ні класичний правовий пошук.

Вони дають **охоплення**, недосяжне вручну: не вибірку з десятка справ «навпомацки», а відтворюваний прогін по всьому корпусу року. Вони показують **структуру практики, а не анекдоти**: де домінує один підхід, де практика розколота, яка палата найменш стабільна — це матеріал і для наукової статті, і для записки суду, і для обґрунтування законодавчої ініціативи. Вони допомагають **формулювати й перевіряти гіпотези** про причину дисфункції норми: система не просто рахує, а класифікує дефект (прогалина, колізія, дефект законодавчої техніки, системна неузгодженість). І, нарешті, вони продукують **звіти «під підпис»**, де кожен висновок прив’язаний до рядка доказової матриці й до перевірюваної цитати, а в додатку лежить усе для повтору — ідентифікатор прогону, версії промптів, запити, вартість. Це знімає головне заперечення проти ШІ в праві — «а звідки ви це взяли?».

Але ця сама картина окреслює й **місце людини**. Машина вимірює — людина вирішує, *чи достатній* фрагмент для правового висновку. Машина знаходить патерн — людина оцінює, чи не артефакт це методу (як ті самі «2 з 220»). Машина видає відсоток — людина переводить його у чесний діапазон, пам’ятаючи, що за десятими може ховатися один невпевнений прохід моделі. І, як показав зовнішній аудит, саме людина лишається останньою інстанцією якості: внутрішня рецензія — корисне сито, але не підпис експерта.

Звідси головна теза, без пафосу. Цінність створює **не модель сама по собі, а дисципліна навколо неї**: методологія, структуровані промпти, перевірювані джерела, відтворювані прогони, контроль якості, людська перевірка й чесний опис меж. Заберіть дисципліну — й інфраструктура знову схлопнеться в балакучий чат.

Що варто дослідити далі

NormaTrace корисний ще й тим, що ясно показує фронт робіт. Перш ніж виносити такі висновки в публікацію або в суд, варто: провести **ручний аудит вибірки** одно-прохідних класифікацій, підтверджуючи відсотки результатів; **незалежно перевірити тонку базу аналогій** («2 з 220» — властивість практики чи артефакт методу?); **звірити повноту корпусу** за нормою; **порахувати повну вартість**, включно з роботою оркестратора. Окремий великий сюжет — як зробити зовнішню рецензію незалежною архітектурно, а не декларативно, і як довести демонстраційний продуктовий шар до придатного для реальної роботи.

Це не перелік претензій, а карта дорослішання технології, яка вже перестала бути іграшкою, але ще не стала інструментом, якому можна довіряти не дивлячись. І, мабуть, найздоровіше, що можна сказати про LLM-агентів у праві сьогодні: вони достатньо хороші, щоб радикально посилити юриста, і недостатньо хороші, щоб його замінити, — а чесні системи на кшталт NormaTrace цінні тим, що показують, *де* проходить ця межа.

Додаток: використані матеріали та приклади

Які матеріали проєкту використано. Стаття спирається на дослідницький бриф по проєкту NormaTrace (мета-аналіз матеріалів проєкту від 8 червня 2026 р.) і через нього — на перевірювані артефакти: керівний файл `CLAUDE.md` (методологічний інваріант, ліміт 50 USD/добу); схему механізму цитування `schemas/evidence_refs.schema.json` (заборона на вигадування тексту цитати); реєстр ролей і скілів у `.claude/agents/` та `.claude/skills/`; звіт про структуру дослідницького корпусу `reports/court_decisions_schema_analysis.md`; прогони в `artifacts/runs/` (флагман по ч.2 ст.61 КУзПБ, стабільність касації ВС, судді-доповідачі, дисципліна суддів/ВРП, мікропілот); висновок незалежного зовнішнього аудиту `reports/audit/audit_conclusion_kuzpb_st61_v5.md`; готовності та розрахунки вартості у відповідних `readiness_brief.md` і `cost_report.md`.

Які приклади з NormaTrace увійшли до статті. (1) Субсидіарна відповідальність КДЛ по ч.2 ст.61 КУзПБ як «глибокий кейс» з виміряними результатами (60,0% відмов; 75,6% «причинність не доведено»; воронка 260→210→70; вартість $3,92) та еволюцією звіту під зовнішній аудит. (2) Стабільність касації ВС-2025 як «кейс масштабу і regex-first» (16 460 постанов; 58,2% стабільності; розкид палат 68,2%→43,2%; ~$1,49 проти оцінних ~$233). (3) Судді-доповідачі як «кейс методологічної обережності» (попередження про незастосовність метрики; $0). (4) Позиції ЄСПЛ через HUDOC як «кейс верифіковуваного цитування» (79 позицій з 8 справ; виключення неперевірюваних). (5) Зовнішній аудит з вердиктом «повернути на доопрацювання» як ілюстрація розходження внутрішнього і зовнішнього контролю.