Od czatu do infrastruktury badawczej: jak agenci LLM zmieniają analizę orzecznictwa

Relacja eksperymentalna z projektu NormaTrace: modelowe korpusy orzecznictwa z 2018 i 2025 roku, przepływy pracy agentowej i wnioski możliwe do zweryfikowania

Artykuł popularny · NormaTrace · nota badawcza z 8 czerwca 2026 r.

> Popularny, edukacyjny przewodnik — bez marketingu i żargonu technicznego. To projekt badawczy, eksperymentalny, a nie przemysłowy system wyszukiwania prawniczego. Wszystko, co poniżej, opiera się na materiałach jednego konkretnego projektu; na końcu znajduje się lista dokładnie tego, co wykorzystano, oraz które przykłady z projektu trafiły do artykułu.

Skąd to wszystko się bierze: dwa uprzedzenia wobec sieci neuronowych

Siadając do rozmowy z dużym modelem językowym, prawnik zwykle wpada w jedną z dwóch skrajności. Pierwsza to entuzjazm: „maszyna zna prawo, zaraz wszystko przeanalizuje". Druga, jakieś dziesięć minut później, to rozczarowanie: „wymyśliła numer sprawy, powołała się na nieistniejące orzeczenie — nie nadaje się do poważnej pracy".

Znamienne, że obie reakcje wyrastają z tego samego ukrytego założenia — przekonania, że praca z modelem to rozmowa: pytasz, model odpowiada prozą. I to założenie jest właśnie sednem błędu, bo „rozmowa z modelem" ma trzy wady, które są zabójcze dla prawa.

Odpowiedź czatu jest niemożliwa do zweryfikowania: cytat i odesłanie są wyciągane „z głowy modelu" i po prostu mogą nie istnieć. Jest nieodtwarzalna: zadaj to samo pytanie jutro, a dostaniesz inną odpowiedź, bez żadnego zapisanego śladu rozumowania. I jest niemierzalna: model opowie o trzech czy czterech sprawach, które akurat zna, ale nie powie tego, co najważniejsze — jak norma funkcjonuje w całym korpusie praktyki: jaki jest odsetek oddaleń, gdzie orzecznictwo się rozjechało, która izba najrzadziej „przetrwa" kasację.

A przecież to właśnie mierzalne życie normy stanowi sedno pracy naukowej, analitycznej i legislacyjnej. Badacz i ustawodawca potrzebują nie streszczenia, lecz odpowiedzi: czy norma wywołuje skutek, dla którego została uchwalona, a jeśli nie — na czym polega wada i jak ją naprawić.

Projekt NormaTrace, o którym jest ten artykuł, pokazuje, że tę lukę już się zasypuje. Nie dzięki „mądrzejszemu modelowi" — lecz dzięki zmianie gatunku: z rozmowy na infrastrukturę.

Najpierw ustalmy terminologię

Żeby później się nie potykać, wyjaśnijmy kilka pojęć prostym językiem. Nie da się ich uniknąć, ale jest ich niewiele.

LLM (duży model językowy) — program wytrenowany na ogromnych zbiorach tekstu do przewidywania kolejnego słowa; z tego prostego mechanizmu wyrasta zdolność do spójnego pisania, streszczania, klasyfikowania i wnioskowania. Ważne jest jedno: model nie „pamięta" dokumentów tak, jak robi to biblioteka — przechowuje statystyczne prawidłowości języka. Dlatego gdy poprosisz go o zacytowanie orzeczenia, nie sięga po nie z półki, lecz wiarygodnie je rekonstruuje — i czasem się myli, nie odczuwając różnicy między „przypomniałem sobie" a „wymyśliłem". To właśnie słynna halucynacja.

Przepływ agentowy (pipeline agentowy) — sposób organizacji pracy modelu nie jako pojedynczej rozmowy, lecz

jako procesu produkcyjnego z etapami, rolami i punktami kontrolnymi. Różnica jest taka jak między poproszeniem znajomego prawnika o radę na korytarzu a uruchomieniem badania w dziale analitycznym — z planem, podziałem zadań, weryfikacją źródeł, wewnętrzną recenzją i podpisanym raportem. Czat to korytarz; przepływ agentowy to dział.

RAG / przeszukiwanie korpusu (od retrieval-augmented generation) — technika, w której model najpierw wyszukuje odpowiednie dokumenty w dużej bazie danych, a dopiero potem rozumuje wyłącznie na podstawie dostarczonego tekstu, a nie z pamięci. Różnica jest taka jak między „powiedz mi, co wiesz o tej sprawie" a „oto pięćdziesiąt orzeczeń, przeanalizuj je i nie wykraczaj poza nie". To drugie podejście jest fundamentem weryfikowalności.

Silnik cytowań (citation engine) — „dyscyplina odesłań", która fizycznie uniemożliwia modelowi wymyślanie cytatów; jak to dokładnie działa, wyjaśnimy niżej.

Role, umiejętności (skills), Human Review — podział pracy wewnątrz pipeline'u: jedni agenci szukają źródeł, inni analizują, jeszcze inni weryfikują, a czwarty liczy pieniądze. Human Review to etap recenzji naśladujący spojrzenie wymagającego, czepliwego starszego kolegi.

Teraz możemy przejść do meritum.

Na czym to się opiera: modelowe korpusy orzecznictwa

Od razu określmy status projektu, żeby nie budzić zawyżonych oczekiwań. NormaTrace to badawcze, eksperymentalne środowisko, a nie przemysłowy system wyszukiwania prawniczego. Projekt pracuje na modelowych korpusach orzeczeń sądowych z 2018 i 2025 roku — eksperymentalnych bazach danych zebranych po to, by testować hipotezy, dopracowywać metodologię i weryfikować przepływy pracy. To nie jest produkcyjny indeks wyszukiwania ani system zaprojektowany pod przemysłowe obciążenie; to „stanowisko badawcze", na którym testuje się samo podejście.

Ale nawet takie stanowisko zmienia sposób stawiania pytania. Korpus jednego roku to już dziesiątki tysięcy orzeczeń: opisany niżej przypadek obejmuje na przykład 16 460 merytorycznych orzeczeń Sądu Najwyższego za 2025 rok. Przeczytanie tylu orzeczeń ręcznie jest niemożliwe — ani dla jednego prawnika, ani dla całego działu. I to właśnie taki zasięg po raz pierwszy pozwala zadawać pytania, które wcześniej były czysto retoryczne: w jakim odsetku spraw sąd oddala roszczenie na podstawie danej normy? który argument najczęściej staje się przyczyną oddalenia? czy praktyka różnych izb kasacyjnych jest ze sobą spójna?

Oto ważny i szczery szczegół, który sami autorzy projektu podkreślają: wąskim gardłem nie są dane, lecz metodologia. Teksty są już zebrane; trudność polega na przekształceniu dziesiątek tysięcy surowych orzeczeń w wniosek, pod którym można się podpisać z pełnym zaufaniem. Dlatego środek ciężkości projektu to nie „sprytne prompty" ani parsery, lecz dyscyplina: mapa źródeł, macierz dowodowa, silnik cytowań i kontrola jakości.

Cały proces jest podporządkowany jednej formule, zapisanej w pliku kontrolnym projektu (CLAUDE.md), a potem powtarzanej w tuzinie miejsc jako inwariant metodologiczny:

norma → praktyka stosowania → wzorce → problem → klasyfikacja → rozwiązanie → prognoza skutków → raport możliwy do zweryfikowania

To „gatunek infrastruktury" zamknięty w jednym zdaniu. Nie „odpowiedz mi", lecz „przejdź całą drogę od tekstu normy do możliwego do zweryfikowania raportu, zostawiając ślad na każdym kroku".

> **Infografika 2. Cykl życia analizy normy prawnej.** Droga od tekstu normy i korpusu praktyki, przez wybór istotnych spraw, klasyfikację rozstrzygnięć i wykrywanie wzorców — do mapy problemów, wniosków prawnych, rekomendacji i prognozy skutków.

Czym przepływ agentowy różni się od czatu: podział pracy

W zwykłym czacie jeden model robi wszystko naraz — i dlatego nie robi niczego wiarygodnie. W NormaTrace praca jest podzielona na role (w języku projektu: subagentów). Jest ich około tuzina, a ich układ jest przemyślany.

> **Infografika 1. Od czatu do infrastruktury badawczej.** Zwykły czat daje jednorazową, niepopartą dowodami odpowiedź; przepływ agentowy prowadzi pytanie prawnika przez metodologię, przeszukiwanie bazy danych, role agentów, weryfikację źródeł, silnik cytowań i kontrolę jakości — i na każdym etapie zostawia ślad możliwy do zweryfikowania.

Jest koordynator, który planuje badanie i rozgałęzia pracę — i, co znamienne, wyłącznie on ma prawo „powoływać" nowych wykonawców. Jest analityk z żelazną zasadą: każdy wniosek musi być powiązany z wierszem macierzy dowodowej. Są specjaliści od źródeł (osobno dla wewnętrznej bazy danych, osobno dla sieci) oraz inżynier zapytań masowych. I są trzy role „sumienia": strażnik budżetu — jedyny uprawniony do zaciągnięcia hamulca bezpieczeństwa; audytor odtwarzalności, który sprawdza, czy każde zapytanie i każdy cytat da się powtórzyć; oraz autonomiczny recenzent „ludzki" — agent grający rolę czepliwego recenzenta.

Obok ról są umiejętności (skills): powtarzalne procedury w formie list kontrolnych („jak cytować", „jak liczyć pieniądze", „jak weryfikować odtwarzalność"). Jeśli rola to „kto", to umiejętność to „według jakiego regulaminu".

Skąd takie rozdrobnienie? Z tego samego powodu, dla którego istnieje w prawdziwej kancelarii. Jedna osoba robiąca wszystko naraz myli się, nawet tego nie zauważając. Podział pracy tworzy tarcie i punkty kontrolne: ten, kto analizuje, nie jest tym, kto weryfikuje; ten, kto wydaje tokeny, nie jest tym, kto pilnuje budżetu. Błąd jednego agenta ma szansę zostać wychwycony przez innego — i to właśnie sztuczne tarcie odróżnia badanie od gadaniny.

Jak to zadziałało na żywej normie: odpowiedzialność subsydiarna

Abstrakcje najlepiej testować na konkretach. Flagowym przypadkiem projektu jest art. 61 ust. 2 Kodeksu Ukrainy o procedurach upadłościowych: odpowiedzialność subsydiarna (posiłkowa) osób kontrolujących dłużnika („osoby kontrolujące"). To norma „gorąca": gdy upadające przedsiębiorstwo nie może spłacić długów, powstaje pytanie, czy można je odzyskać od osób, które faktycznie nim zarządzały. Pytanie badawcze sformułowano empirycznie: jak ta norma jest faktycznie stosowana przez sądy w 2025 roku?

Pipeline przeszedł całą drogę. Z modelowego korpusu za 2025 rok, na podstawie „sygnatury" normy, wybrano około 260 kandydatów; po deduplikacji zostało 210 unikalnych spraw; po odfiltrowaniu spraw proceduralnych i nieistotnych — 70 orzeczeń „merytorycznych", w których norma została zastosowana co do istoty. Następnie model sklasyfikował rozstrzygnięcia. Wyniki (ściśle jako fakty dotyczące artefaktów tego przebiegu, a nie jako niezależnie zweryfikowany wynik prawny): odpowiedzialność została oddalona w 60,0% spraw (42 z 70), a dominującą podstawą oddalenia był brak udowodnionego związku przyczynowego: 75,6% (34 z 45) w odpowiedniej podgrupie. Cały przebieg kosztował 3,92 USD (666 wywołań modelu „lekkiego").

Warto się przy tym zatrzymać: to nie jest streszczenie trzech spraw, lecz mapa zachowania normy w całym rocznym korpusie, z której wyłania się diagnoza strukturalna — sądy nie mają jednolitego, przewidywalnego testu związku przyczynowego, i to właśnie „nieudowodniony związek przyczynowy" jest głównym kanałem, przez który wycieka odpowiedzialność. Takiego wniosku nie da się uzyskać w czacie — można go tylko zmierzyć.

Ten sam wątek przeprowadzono także po raz drugi, niezależnie — przy pomocy ścisłego zestawu promptów klienckich

opartych na zasadzie „tylko na podstawie dostarczonego tekstu” (rozumuj wyłącznie na podstawie dostarczonego materiału; nie korzystaj z wewnętrznej wiedzy o kodeksie). Oba przebiegi doprowadziły do tej samej diagnozy, a sama zbieżność dwóch niezależnych ścieżek wzmacnia zaufanie do wniosku.

Silnik cytowań: dlaczego model tutaj fizycznie nie może skłamać w cytacie

Teraz — o najważniejszym elemencie, tym, dla którego w istocie zbudowano całą konstrukcję.

Główna obawa prawnika wobec sieci neuronowej: „zacytuje sprawę, która nie istnieje”. NormaTrace rozwiązuje ten problem nie poprzez namawianie modelu, by „nie zmyślał”, lecz architektonicznie.

Oto jak to działa. Wszystkie dokumenty źródłowe są z góry pocięte na numerowane segmenty — każdy akapit otrzymuje stabilny identyfikator (coś w rodzaju P0001) i odcisk skrótu (hash) tekstu. Gdy model przygotowuje wniosek, nie wolno mu wpisać treści cytatu: może jedynie wskazać identyfikator potrzebnego segmentu — „właśnie tutaj, w akapicie P0042”. A rzeczywisty, dosłowny tekst dostarcza nie model, lecz osobny, deterministyczny „resolver”, który pobiera go z rejestru na podstawie identyfikatora.

W schemacie danych zapisano to dosłownie jako zakaz: „model wybiera wyłącznie odesłania; NIE WOLNO mu układać treści cytatu” (evidence_refs.schema.json). W schemacie po prostu nie ma pola na dowolny tekst cytatu — nie ma gdzie wpisać wymysłu.

Metafora jest prosta. Zwykły model to gawędziarz, który cytuje z pamięci i czasem zmyśla. Silnik cytowań zmienia go w kogoś, kto wskazuje palcem linijkę w książce, podczas gdy samą książkę trzyma i czyta ktoś inny, nieprzekupny. Zaufanie przenosi się z modelu na warstwę deterministyczną.

> **Infografika 3. Jak działa silnik cytowań.** Dokument jest z góry podzielony na numerowane akapity; model wybiera tylko identyfikator fragmentu, a dosłowny tekst dostarcza deterministyczny resolver, po czym walidator sprawdza odesłanie. Model fizycznie nie może „ułożyć” cytatu — w schemacie po prostu nie ma pola na dowolny tekst cytatu.

I to działa. Recenzując jeden z finalnych raportów, niezależny zewnętrzny audytor potwierdził: wszystkie 29 odesłań jest na swoim miejscu, wszystkie 29 dosłownych cytatów zgadza się z tekstem kanonicznym, sumy kontrolne się zgadzają. Co więcej, audytor przeliczył kluczowe wskaźniki na podstawie plików źródłowych — „wszystkie wartości zgadzały się co do cyfry”.

Ale i tutaj jest szczera granica, której projekt nie ukrywa. Silnik cytowań dowodzi, że model wskazał na istniejący fragment właściwego dokumentu. Nie dowodzi, że ten fragment jest prawnie wystarczający dla danego wniosku: trafność stanowiska prawnego pozostaje w gestii sumienia człowieka. Odesłanie jest poprawne — ale czy jest trafne, decyduje prawnik.

Kontrola błędów, halucynacji, pieniędzy i jakości

Weryfikowalność cytatów to tylko jedna linia obrony. Jest ich kilka, a razem tworzą „dyscyplinę wokół modelu”.

Pieniądze. Jedyną twardą podstawą do zatrzymania całego systemu jest ryzyko przekroczenia limitu 50 USD dziennie na zapytania do modelu. Wszystko inne (nawet klasyfikacja prawna i wewnętrzna recenzja) system wykonuje autonomicznie, ale pieniądze traktuje jak świętość: przed dużym przebiegiem obowiązkowe jest wstępne oszacowanie kosztów, a jeśli cena jest nieznana — stop.

A oto być może najbardziej praktyczna technika projektu: „najpierw regex, potem LLM”. Większość powtarzalnej pracy

(na przykład klasyfikacja rozstrzygnięcia na podstawie standardowego brzmienia sentencji) wykonuje nie drogi model, lecz tanie, deterministyczne dopasowywanie wzorców; model jest angażowany dopiero do „reszty" — przypadków niejednoznacznych — oraz do kontroli końcowej. Efekt jest oszałamiający: analiza 16 460 orzeczeń Sądu Najwyższego kosztowała około 1,5 USD zamiast szacowanych 233 USD, gdyby każde orzeczenie przepuścić przez model. Oszczędność 150-krotna — a przy tym wyższa odtwarzalność, bo regex zawsze daje ten sam wynik.

> **Infografika 4. Ekonomia przepływu pracy: najpierw regex, potem LLM.** Zamiast przepuszczać wszystkie dokumenty przez drogi model, tani deterministyczny filtr usuwa większość, a model pracuje tylko nad spornymi przypadkami; potem następuje kontrola jakości. Liczby podano jako przykład z przebiegu eksperymentalnego, a nie jako metrykę produktową.

Kontradyktoryjna kontrola jakości. Wewnętrzna recenzja jest prowadzona przez kilku niezależnych recenzentów, a każdy komentarz blokujący jest ponownie sprawdzany głosowaniem sceptyków. To nie jest dekoracja: w jednym z przebiegów taka kontrola wychwyciła realne zawyżenie — „8 z 8" zamiast poprawnego „7 z 7" dla jednego z analogów-donorów — a audytor potwierdził następnie, że poprawka została wprowadzona.

Główny szczery wynik. A teraz — to, co odróżnia dojrzały projekt od marketingowego hasła. Gdy jeden z flagowych raportów przeszedł wewnętrzną kontrolę ze statusem „wystarczający", wysłano go do niezależnego zewnętrznego audytu, który wydał werdykt: „DO POPRAWKI" — warunkowo pozytywny, z pięcioma komentarzami blokującymi. Wewnętrzna bramka powiedziała „wystarczająco dobre", zewnętrzna powiedziała „jeszcze nie".

Ta rozbieżność jest być może najcenniejszą lekcją z całego projektu: recenzja wewnątrz modelu nie zastępuje zewnętrznej ekspertyzy. Agent grający rolę czepliwego recenzenta to ten sam model; jego niezależność jest deklarowana, ale nie zagwarantowana architektonicznie. Prawdziwe spojrzenie z zewnątrz na razie pozostaje po stronie człowieka.

Co jeszcze pokazały eksperymenty

Oprócz przypadku flagowego projekt przeprowadził kilka merytorycznie odmiennych badań — a każde z nich dodaje coś do obrazu całości.

Skala i regex-first: stabilność kasacji. Pytanie — jaki odsetek orzeczeń sądów niższej instancji „przetrwał" kasację Sądu Najwyższego w 2025 roku. Metoda — deterministyczne parsowanie brzmień sentencji względem taksonomii 16 kodeksów z pokryciem 98%; model jedynie kontrolował próbkę. Wynik: 58,2% orzeczeń przetrwało, 41,8% zostało zdestabilizowanych, przy dużym rozrzucie między izbami — od 68,2% w Wielkiej Izbie do 43,2% w kasacji administracyjnej. Zgodność kontroli modelu z regexem wyniosła niemal 95%. To, co ręcznie zajęłoby miesiące, zajęło kilka minut i około półtora dolara.

Przypadek metodologicznej ostrożności: sędziowie sprawozdawcy. Na już oznaczonym korpusie obliczono stabilność dla poszczególnych sędziów sprawozdawców (nie kosztowało to nic — wyniki po prostu połączono z gotowymi danymi). Cenniejsze od samych liczb jest jednak ostrzeżenie, które system sam zapisał w raporcie: wskaźnika „stopnia stabilności" nie wolno używać jako oceny jakości pracy sędziego ani jako podstawy wniosków dyscyplinarnych bez uwzględnienia struktury spraw. Sprawy złożonych kategorii częściej „padają" w kasacji — a to mówi o kategorii, nie o sędzim. System, który sam ogranicza zastosowanie własnych metryk, zasługuje na znacznie większe zaufanie niż taki, który beztrosko wystawia rankingi.

Przypadek praktyki ponadnarodowej: ETPC przez HUDOC. Do argumentacji prawnoporównawczej projekt wydobył 79 stanowisk Europejskiego Trybunału Praw Człowieka z 8 spraw, każde z precyzyjnym odesłaniem do akapitu. I znów

znamienna jest nie skala, lecz dyscyplina: sprawy niemożliwe do zweryfikowania (nieustalony numer skargi, tekst wyłącznie po francusku) zostały wykluczone, a wczesne cytaty „z pamięci” wychwycono w pętli kontroli jakości i zastąpiono weryfikowalnymi. To bezpośredni prototyp przygotowywania materiałów dla trybunałów międzynarodowych.

Mikropilotaż. Przed uruchomieniem dużych korpusów ścieżkę przetwarzania masowego przetestowano na ośmiu syntetycznych dokumentach: poprawny format wyjścia — 8 z 8, trafność selekcji — 8 z 8, koszt — poniżej centa. Nudne — ale właśnie tak wygląda inżynierska staranność: najpierw przetestuj rurę wodą, dopiero potem puść nią właściwy przepływ.

Gdzie system wciąż się potyka

Uczciwość projektu polega na tym, że dokumentuje swoje ograniczenia równie dokładnie jak osiągnięcia. Wymieńmy najważniejsze — to ważniejsze niż jakikolwiek entuzjazm.

„Pozorna precyzja”. Czytając „60,0% oddaleń” albo „75,6% w sprawie związku przyczynowego”, łatwo uwierzyć w precyzję do dziesiątej części procenta. Ale za tymi liczbami stoi jedno przejście taniego modelu (średnia pewność około 0,7), a nie wszystkie sprawy zostały zweryfikowane dosłownie. Właściwym sposobem prezentacji jest podanie przedziałów („około 60%”), a nie dziesiątych części procenta, plus ręczny audyt próbki 20–30 orzeczeń. Liczba, która wygląda na precyzyjną, i liczba, która rzeczywiście jest precyzyjna, to nie to samo.

Cienka baza analogii. Żeby zaproponować rozwiązanie „jak w podobnej konstrukcji prawnej”, trzeba mieć te konstrukcje pod ręką. Pełny przegląd 220 stanowisk prawnych dał jednak tylko 2 prawdziwe komparatory — co oznacza, że część wniosków prawnoporównawczych opiera się bardziej na niedostatku materiału niż na wiarygodnie przeniesionym standardzie. Ryzyko fałszywej analogii jest realne; projekt uczciwie je odnotowuje, ale go nie eliminuje.

Kompletność i jakość danych. Część metadanych jest skąpa, jurysdykcja bywa wnioskowana z pośrednich sygnałów (z nazwy sądu), a korpus dla konkretnej normy może być niekompletny (263 sprawy tam, gdzie teoretycznie mogło być ~1000). Dosłowne zakotwiczenie cytatów nie zostało wykonane dla wszystkich źródeł — dynamicznych stron oficjalnych portali prawnych nie da się po prostu „przypiąć”.

Granica autonomii. Limit 50 USD był przestrzegany (do ósmego zadania narosło około 33 z 50 USD), ale nie ma jednego skonsolidowanego licznika kosztów — sumy zbiera się z osobnych plików i nieco się one rozjeżdżają. I subtelniej: limit obejmuje wyłącznie zapytania do modelu, a praca samego orkiestratora nie jest w nim uwzględniona — pełny koszt obliczeniowy jest wyższy niż zarejestrowany.

Dojrzałość warstwy „produktowej”. Na bazie badawczego rdzenia budowany jest interfejs, przez który prawnik mógłby uruchomić pipeline zapytaniem w języku naturalnym. To jednak na razie demonstracja, nie produkt: autoryzacja to atrapa, część adapterów to „zaślepki”, sprawdzenie gotowości do wdrożenia zwróciło wynik „niepowodzenie”. Nie da się tego wdrożyć produkcyjnie.

I osobne zastrzeżenie, ważne dla dokładności: nazwy modeli w projekcie („GPT-5.4”, „Mini”) to role routingowe w ramach metodologii, a nie potwierdzone zewnętrzne identyfikatory; co faktycznie działa pod tymi nazwami, zależy od środowiska.

Dlaczego to już jest przydatne dla prawnika — i gdzie bez człowieka się nie obejdzie

Jeśli odłożyć na bok zarówno entuzjazm, jak i sceptycyzm, zostaje trzeźwy obraz: takie systemy dają prawnikowi już dziś to, czego nigdy nie dawał ani czat, ani klasyczne wyszukiwanie prawnicze.

Dają zasięg nieosiągalny ręcznie: nie próbkę kilkunastu spraw wybranych „na wyczucie", lecz powtarzalny przebieg przez cały roczny korpus. Pokazują strukturę praktyki, a nie anegdoty: gdzie dominuje jedno podejście, gdzie praktyka jest podzielona, która izba jest najmniej stabilna — to materiał na artykuł naukowy, na pismo procesowe do sądu i na uzasadnienie inicjatywy ustawodawczej. Pomagają formułować i testować hipotezy o przyczynie dysfunkcji normy: system nie tylko liczy, ale klasyfikuje wadę (lukę, kolizję przepisów, wadę techniki legislacyjnej, systemową niespójność). I wreszcie tworzą raporty „gotowe do podpisu", w których każdy wniosek jest powiązany z wierszem macierzy dowodowej i możliwym do zweryfikowania cytatem, a w załączniku znajduje się wszystko potrzebne do powtórzenia przebiegu — identyfikator przebiegu, wersje promptów, zapytania, koszt. To usuwa główny zarzut wobec AI w prawie — „a skąd to wziąłeś?".

Ale ten sam obraz wyznacza też miejsce człowieka. Maszyna mierzy — człowiek decyduje, czy fragment jest wystarczający dla wniosku prawnego. Maszyna znajduje wzorzec — człowiek ocenia, czy nie jest to artefakt metody (jak właśnie owo „2 z 220"). Maszyna wystawia procent — człowiek przekłada go na uczciwy przedział, pamiętając, że za dziesiątymi częściami procenta może kryć się jedno niepewne przejście modelu. I, jak pokazał zewnętrzny audyt, to właśnie człowiek pozostaje ostateczną instancją jakości: wewnętrzna recenzja to użyteczne sito, ale nie podpis eksperta.

Stąd główna teza, bez patosu. Wartość tworzy nie sam model, lecz dyscyplina wokół niego: metodologia, ustrukturyzowane prompty, weryfikowalne źródła, odtwarzalne przebiegi, kontrola jakości, Human Review i uczciwy opis granic. Usuń dyscyplinę — a infrastruktura z powrotem zapadnie się w gadatliwy czat.

Co jeszcze warto zbadać

NormaTrace jest przydatny również dlatego, że wyraźnie pokazuje front dalszej pracy. Zanim takie wnioski trafią do publikacji lub do sądu, warto: przeprowadzić ręczny audyt próbki klasyfikacji z jednego przejścia, potwierdzając procenty rozstrzygnięć; niezależnie sprawdzić cienką bazę analogii (czy „2 z 220" to cecha praktyki, czy artefakt metody?); zweryfikować kompletność korpusu dla danej normy; policzyć pełny koszt, łącznie z pracą orkiestratora. Osobnym, dużym wątkiem jest to, jak sprawić, by zewnętrzna recenzja była niezależna architektonicznie, a nie tylko deklaratywnie, oraz jak doprowadzić demonstracyjną warstwę produktową do stanu przydatnego do realnej pracy.

To nie jest lista zarzutów, lecz mapa dojrzewającej technologii — takiej, która przestała już być zabawką, ale jeszcze nie stała się narzędziem, któremu można ufać bez patrzenia mu na ręce. I chyba najzdrowszą rzeczą, jaką można dziś powiedzieć o agentach LLM w prawie, jest to: są już wystarczająco dobrzy, by radykalnie wzmocnić prawnika, i wciąż niewystarczająco dobrzy, by go zastąpić — a uczciwe systemy, takie jak NormaTrace, są cenne właśnie dlatego, że pokazują, gdzie przebiega ta granica.

Załącznik: wykorzystane materiały i przykłady

Jakie materiały projektu wykorzystano. Artykuł opiera się na nocie badawczej projektu NormaTrace (metaanaliza materiałów projektu z 8 czerwca 2026 r.), a za jej pośrednictwem — na możliwych do zweryfikowania artefaktach: pliku kontrolnym CLAUDE.md (inwariant metodologiczny, limit 50 USD dziennie); schemacie silnika cytowań schemas/evidence_refs.schema.json (zakaz układania treści cytatu); rejestrze ról i umiejętności w .claude/agents/ i .claude/skills/; raporcie o strukturze korpusu badawczego reports/court_decisions_schema_analysis.md; przebiegach w artifacts/runs/ (flagowy przebieg dotyczący art. 61 ust. 2 Kodeksu o procedurach upadłościowych, stabilność kasacji Sądu Najwyższego, sędziowie sprawozdawcy, dyscyplina sędziowska / Wyższa Rada Sprawiedliwości, mikropilotaż); wnioskach niezależnego zewnętrznego audytu reports/audit/audit_conclusion_kuzpb_st61_v5.md; ocenach gotowości i szacunkach kosztów w odpowiednich plikach readiness_brief.md i cost_report.md.

Które przykłady z NormaTrace trafiły do artykułu. (1) Odpowiedzialność subsydiarna osób kontrolujących na podstawie art. 61 ust. 2 Kodeksu

o procedurach upadłościowych jako „głęboki przypadek” ze zmierzonymi wynikami (60,0% oddaleń; 75,6% „nieudowodniony związek przyczynowy”; lejek 260→210→70; koszt 3,92 USD) oraz ewolucja raportu pod wpływem zewnętrznego audytu. (2) Stabilność kasacji Sądu Najwyższego w 2025 roku jako „przypadek skali i regex-first” (16 460 orzeczeń; stabilność 58,2%; rozrzut między izbami 68,2%→43,2%; ok. 1,49 USD wobec szacowanych ok. 233 USD). (3) Sędziowie sprawozdawcy jako „przypadek metodologicznej ostrożności” (ostrzeżenie o niestosowalności metryki; 0 USD). (4) Stanowiska ETPC przez HUDOC jako „przypadek weryfikowalnego cytowania” (79 stanowisk z 8 spraw; wykluczenie tych niemożliwych do zweryfikowania). (5) Zewnętrzny audyt z werdyktem „do poprawki” jako ilustracja rozbieżności między kontrolą wewnętrzną a zewnętrzną.

Ten polski artykuł jest adaptacją anglojęzycznej wersji, która z kolei jest adaptacją ukraińskiego oryginału; liczby, cytaty i wnioski zostały zachowane.