

Czy odpowiedzi modeli AI są poprawnie ugruntowane w polskim prawie?

Benchmark 02: dziesięć krótkich pytań prawnych, osiem modeli, bez narzędzi i z wyszukiwaniem.

Każde powołanie przepisu i orzeczenia sprawdzone w źródle urzędowym na dzień stanu prawnego 1.10.2026. Te same odpowiedzi ocenione przez pięciu sędziów-modeli i przez warstwę weryfikacji powołań. Zbiór kalibracyjny: dziesięć pytań, jedna seria, reguły oceny zapisane z datą.

OpenAI

GPT-6.1 Sol

GPT-6 Astra

GPT-5.6 Terra

Codex CLI, rozumowanie medium

Anthropic

Claude Opus 5.5

Claude Sonnet 5.5

Claude Haiku 5.5

Claude Code CLI

Google

Gemini 4 Argon

Gemini 3.8 Flash

Vertex AI; sędziowie także Gemini 3.1 Pro

160

odpowiedzi: 10 pytań × 8 modeli × 2 warunki (W0 bez narzędzi, W1 z wyszukiwaniem)

91

odpowiedzi z co najmniej jedną halucynacją (51 w W0, 40 w W1)

0–7 z 10

odpowiedzi poprawnie ugruntowanych na model (46 ze 160)

≤ 0,90

najlepszy sędzia-model (Gemini 4 Argon) na kartach sprawdzonych przed sędziami, n = 79

≥ 170,70 USD

pełny koszt badania: odpowiedzi ≥ 15,78, sędziowie ≥ 68,85, warstwa weryfikacji 86,07 (bez Codex z subskrypcji)

BLOK A · RZETELNOŚĆ / BLOK B · PRZYDATNOŚĆ / KLASY BŁĘDU F · T · W · S

Adam Piotrowski, Andrzej Koman (LexAlpha Lab)

Benchmark 02: pytania prawne. Czy odpowiedzi modeli AI są poprawnie ugruntowane w polskim prawie — dziesięć pytań, osiem modeli, dwa warunki biegu

Adam Piotrowski, Andrzej Koman (LexAlpha Lab) · wersja 1.0 · 27.10.2026

ABSTRAKT

Benchmark 02 przenosi format z Benchmarku 01 (jeden kazus terminu sankcji kredytu darmowego, dziesięć systemów) na krótkie pytania prawne w postaci, w jakiej prawnik wpisuje je do okna czatu: z literówkami, skrótami, a w sześciu z dziesięciu pytań z wbudowaną pułapką. Zbiór jest kalibracyjny: reguły oceny ustalano w trakcie pracy z kartami, a każdą decyzję zapisywaliśmy z datą. Osiem modeli trzech dostawców otrzymało dziesięć pytań: dziewięć z czterech dziedzin prawa (postępowanie cywilne, postępowanie administracyjne, prawo wykroczeń, prawo autorskie) i jedno o AI Act, każdy w dwóch warunkach: bez narzędzi (W0) i z wyszukiwaniem w internecie (W1), razem 160 odpowiedzi. W źródłach urzędowych sprawdziliśmy 1 436 z 1 454 powołań przepisów i orzeczeń na dzień stanu prawnego 1 października 2026 r.; niesprawdzone powołania nie obniżały oceny. Odpowiedź uznajemy za **poprawnie ugruntowaną**, gdy ma pełny rdzeń i nie zawiera żadnego twierdzenia o źródle prawa niezgodnego z tekstem urzędowym. Takich odpowiedzi było 46 ze 160 (19 w W0, 27 w W1), a 91 zawierało co najmniej jedną halucynację (51 w W0, 40 w W1). Najczęstszą klasą błędu nie było zmyślane źródło (8 odpowiedzi), lecz prawdziwy przepis lub wyrok z treścią, której w nim nie ma (66), oraz brzmienie przepisu z innej daty podane jako obowiązujące (35). Wyszukiwanie podniosło średni wynik bloku obiektywnego u wszystkich ośmiu modeli; liczba odpowiedzi z halucynacją spadła u pięciu modeli (o 4, 3, 2, 1 i 1) i nie zmieniła się u trzech. Te same odpowiedzi oceniło pięciu sędziów-modeli; na pięciu pytaniach, których karty sprawdzono przed uruchomieniem sędziów (jedna etykieta zmieniona później po zawężeniu reguły), najwyższą zgodność z kartą prawnika w etykiecie halucynacji osiągnął Gemini 4 Argon (κ Cohena 0,90, n = 79), najniższą Gemini 3.8 Flash (0,58, n = 80). Sędziowie różnili się kierunkiem błędów: Gemini 3.1 Pro i Gemini 3.8 Flash pomijały halucynacje zapisane w kartach (8 i 16), a zarazem nie wykonały żadnego wyszukiwania przy 25 i 9 kartach; GPT-6.1 Sol zgłosił 7 halucynacji, których karty nie przyjęły. Raport z warstwy weryfikacji (kod sprawdzający powołania w źródłach urzędowych i model oceniający poparcie tezy) podany sędziemu jako wiążący obniżał zgodność obu badanych sędziów; podany jako pomocniczy podniósł zgodność GPT-6.1 Sol z 0,84 do 0,97 i obniżył zgodność Gemini 4 Argon z 0,95 do 0,88. Wycena użycia modeli i narzędzi według cenników z 18 października 2026 r. (nie wydatek) wyniosła co najmniej 170,70 USD, z czego odpowiedzi ocenianych modeli co najmniej 15,78 USD, a ich ocena (sędziowie i warstwa weryfikacji) resztę. Liczby pochodzą z generatorów czytających karty ocen; wszystkie karty sprawdził jeden prawnik.

Słowa kluczowe: benchmark prawniczy, halucynacje, poprawne ugruntowanie, sędzia-model, weryfikacja powołań, brzmienie przepisu na datę, prawo polskie

01 / WSTĘP

Wstęp

Cel raportu, związek z Benchmarkiem 01 i z artykułem „Correctly grounded”, status zbioru.

Benchmark 01 [8] pokazał na jednym kazusie, że rozdzielanie kryteriów obiektywnych od ocennych różnicuje odpowiedzi systemów AI i że rzetelna ocena wymaga sprawdzenia każdego powołania u źródła przez prawnika znającego dziedzinę. Zakończyliśmy go listą kroków przed pełnym benchmarkiem: więcej zadań, zadania podzielone na typy, powtórzone uruchomienia i drugi oceniający. Benchmark 02 zwiększa liczbę zadań i wprowadza drugi rodzaj zadania obok kazusu: krótkie pytanie prawne, w trzech typach (zamknięte, obliczenie, weryfikacja). Pytanie mierzy co innego niż kazus: nie rozumowanie łańcuchowe na rozbudowanym stanie faktycznym, lecz to, czy odpowiedź ma prawidłową konkluzję, nie przypisuje źródłom treści, której w nich nie ma, i koryguje błędną przesłankę pytania.

Drugim punktem odniesienia jest artykuł „*Hallucination Is Not Enough: Why Legal LLM Benchmarks Should Measure „Correctly Grounded”*” [7], w którym na pilotażu z pięciu pytań proponujemy zastąpić pytanie „czy model halucynował?” pytaniem „czy odpowiedź jest poprawnie ugruntowana?”. Powołanie jest poprawnie ugruntowane, gdy oznaczenie istnieje, przypisana mu treść rzeczywiście się w nim znajduje, a brzmienie obowiązywało w dniu, na który pytamy. Punktem wyjścia jest operacjonalizacja zespołu Stanford RegLab [1], którą dostosowujemy do prawa stanowionego. Niniejszy raport jest polską, roboczą wersją materiału empirycznego do wersji 1.2 tej pracy; nie jest jej tłumaczeniem. Obejmuje pełny zbiór dziesięciu pytań, osiem modeli w dwóch warunkach, ocenę pięciu sędziów-modeli i pierwsze wyniki warstwy weryfikacji.

Raport ma cztery cele:

- zmierzyć, ile odpowiedzi ośmiu modeli na krótkie pytania z polskiego prawa jest poprawnie ugruntowanych, z wyszukiwaniem i bez, oraz jakie klasy błędów przeważają;
- sprawdzić, czy sędzia-model z dostępem do wyszukiwarki odtwarza ocenę prawnika i czego nie widzi;
- sprawdzić, czy raport z weryfikacji powołań wykonanej kodem w źródłach urzędowych poprawia ocenę sędziego;
- zapisać reguły oceny wraz z przykładami granicznymi i datami decyzji, aby następny zbiór można było ocenić według zamrożonej wersji.

Zastrzeżenie dotyczące statusu wyników powtarzamy za specyfikacją oceny: pytania P01–P13 i wszystkie dotychczasowe serie służą przede wszystkim do ustalenia kryteriów. Reguły zmienialiśmy po lekturze odpowiedzi (rozdział 4.5), próba jest mała, każdy model odpowiadał raz, a karty wypełniał wstępnie model z rodziny części ocenianych systemów. Wyniki należy czytać z tymi zastrzeżeniami: jako ilustrację kryteriów na małej próbie, a nie jako ostateczne porównanie systemów.

Struktura raportu jest następująca. Rozdział 2 opisuje zbiór pytań i klucze, rozdział 3 systemy i warunki biegu, rozdział 4 metodę oceny i kalendarz decyzji. Rozdział 5 zawiera wyniki modeli, rozdział 6 ocenę sędziów-modeli, rozdział 7 warstwę weryfikacji. Rozdział 8 określa ograniczenia, rozdział 9 wnioski i dalsze kroki. Aneksy zawierają klucze w skrócie (A), wyniki w układzie pytanie × system (B), zgodność i ustawienia sędziów (C) słownik pojęć (D) oraz ma-

cierze pomyłek i przedziały niepewności dla każdej k
(E).

02 / ZBIÓR PYTAŃ: DZIESIĘĆ PYTAŃ Z PRAKTYKI, KLUCZE ZATWIERDZANE PRZED BIEGIEM

Zbiór pytań: dziesięć pytań z praktyki, klucze zatwierdzone przed biegiem

Dziesięć pytań, pułapki, para lustrzana, data stanu prawnego i zmiany kluczy po lekturze źródeł.

2.1. Skład zbioru

Pytania pochodzą z katalogu pytań zebranych w praktyce kancelarii, w brzmieniu, w jakim prawnik wpisał je do czatu. Zbiór ma dwie części. Pięć pytań z postępowania cywilnego (P05, P06, P07, P12,

P13) stanowiło pilotaż opisany w artykule [7]. Pięć kolejnych (P01, P04, P08, P09, P10) dołączyliśmy 18 października 2026 r., aby objąć inne dziedziny: prawo wykroczeń, postępowanie administracyjne, prawo autorskie i AI Act. Typ pytania i cechy klucza zestawia tabela 1.

Tabela 1. Dziesięć pytań zbioru (typ i kalibracja według kluczy; liczby według kart ocen, 16 odpowiedzi na pytanie)

PYTANIE	DZIEDZINA	TYP	PULAPKA	KALIBRACJA	HALUCYNACJE (Z 16)	POPRAWNIE UGRUNTOWANE (Z 16)	A-R = 2 (Z 16)	ŚR. A %
P01	prawo wykroczeń	obliczenie	tak	kategoryczna	12	4	13	69,5
P04	postępowanie administracyjne	zamkniete	tak	kategoryczna	9	0	0	43,0
P05	postępowanie cywilne	zamkniete	tak	kategoryczna	10	4	7	52,1
P06	postępowanie cywilne	zamkniete	nie	warunkowa	12	3	6	39,9
P07	postępowanie cywilne	zamkniete	nie	kategoryczna	10	5	9	55,7
P08	AI Act	weryfikacja	tak	warunkowa	3	2	2	71,6
P09	prawo autorskie	zamkniete	nie	warunkowa	9	7	15	78,2
P10	prawo autorskie	zamkniete	tak	warunkowa	9	7	15	83,0
P12	postępowanie cywilne	zamkniete	tak	warunkowa	6	9	13	79,1
P13	postępowanie cywilne	zamkniete	nie	kategoryczna	11	5	13	68,5

Osiem pytań ma typ zamkniete (trafienie: tak, nie albo „zależy od X”), P01 typ obliczenie (kwota progu), a P08 typ weryfikacja (czy system sprawdza cudze twierdzenie, czy je powtarza). Pułapka, czyli błędna albo nieaktualna przesłanka, którą system ma nazwać i skorygować, jest wbudowana w sześć py-

tań: P01 (powiązanie progu wykroczenia z płacą minimalną, obowiązujące do 14 listopada 2018 r.), P04 (postanowienie zamiast decyzji i klauzula zamiast zaświadczenia), P05 (doręczenie przez portal osobie fizycznej po zmianie art. 131^{1a} § 1 k.p.c. z 1 czerwca 2026 r.), P08 (przesłanka, że głos, który „niczego nie

udaje”, nie jest deepfake; klucz v4 każe ją zweryfikować wobec kryterium „mogą realnie istnieć” z wytycznych KE, a dodatkowo oczekuje wskazania, że obowiązek oznaczania treści syntetycznych z art. 50 ust. 2 AI Act nie zależy od kwalifikacji jako deepfake — ten element rubryki wykracza poza dosłowne pytanie i przyjęliśmy go jako test typu „weryfikacja”), P10 (przesłanka, że ogólna formuła o „wszelkich polach eksploatacji” wystarcza) i P12 (tryb klauzuli przeciwko małżonkowi „bez ograniczeń” i błędny numer przepisu w pytaniu). W pięciu pytaniach klucz oczekuje kalibracji warunkowej: poprawna odpowiedź wiąże konkluzję z warunkiem, którego pytanie nie przesądza.

Pytania P09 i P10 tworzą **parę lustrzaną**: dotyczą tej samej kwestii, ale P09 sugeruje odpowiedź prawidłową („ogólny zapis nie wystarcza?”), a P10 błędną („wystarcza i nie trzeba wyliczać?”). Para służy sprawdzeniu, czy odpowiedź systemu zależy od sugestii zawartej w pytaniu. Klucz P10 ma karę K2 za rdzeń sprzeczny z odpowiedzią tego samego systemu na P09. W biegu żaden z 16 systemów nie otrzymał kary K2, a karę K1 za przytaknięcie przesłance albo odpowiedź wymijającą otrzymał jeden (Claude Haiku 5.5 bez narzędzi: „zwykle nie”). Pełny rdzeń w P09 i w P10 miało po 15 z 16 odpowiedzi.

2.2. Data stanu prawnego i zasady klucza

Każdy klucz ma pole `data_stanu_prawnego`; we wszystkich dziesięciu pytaniach jest to 1 października 2026 r. Odpowiedź ocenia się według prawa na ten dzień, także gdy system odpowiadał później. Klucz zawiera rdzeń odpowiedzi, podstawy wymagane, wzorce cytatów przepisane z tekstu urzędowego, elementy rubryki (punkty dodatnie), kary (punkty ujemne), oczekiwaną kalibrację i pole `weryfikacja`, w którym zapisujemy, czego nie sprawdzono. Orzeczenie wchodzi do klucza z tezą przepisaną z tekstu orzeczenia i ze wskazaniem bazy oraz zakresu odczytu; omówienia i komentarze nie są źródłem tezy. Klucz obejmuje tylko to, o co pyta pytanie.

W badaniu stosujemy wyłącznie klucze zatwierdzone. Klucze P12 i P13 zatwierdzono 10 października 2026 r., klucze P01 (wersja 3), P04 (wersja 3), P08 (wersja 4), P09 (wersja 3) i P10 zatwierdzono 18 października 2026 r., przed biegiem rozszerzenia. Klucze P05, P06 i P07 pochodzą z 10 października 2026 r. i nie mają pola `status`, które wprowadziła zasada z tego samego dnia; potwierdzenie zatwierdzenia jest w rejestrze decyzji (specyfikacja oceny), nie w samym kluczu.

2.3. Zmiany kluczy po lekturze źródeł

Cztery klucze zmieniliśmy w trzech opisanych niżej przypadkach, gdy analiza źródeł ujawniła uproszczenia w kluczu. Zmiany zapisujemy jako nowe wersje z datą i powodem; kart według starszej wersji nie porównujemy z nowymi bez przeliczenia.

- **P04 (wersje 1–2 → 3, przed biegiem).** Klucz przygotowany przez model wskazywał „klauzulę ostateczności, nie prawomocności”. Prawnik wskazał z praktyki, że decyzja ministra o zezwoleniu na nabycie nieruchomości przez cudzoziemca jest najpierw ostateczna (art. 16 § 1, art. 127 § 3 k.p.a.), a potem prawomocna (art. 16 § 3 k.p.a.; art. 52 § 3 i art. 53 § 1 p.p.s.a.), a stan ten potwierdza zaświadczenie organu (art. 217 § 2 pkt 2 w zw. z art. 218 § 1 k.p.a.). Model piszący klucz przeniósł do niego to samo uproszczenie („administracja — ostateczność, sądy — prawomocność”), które benchmark miał wykrywać.
- **P08 (wersja 3 → 4, przed biegiem).** Klucz zakładał, że głos niepodobny do żadnej istniejącej osoby nie jest deepfake. Wytyczne Komisji Europejskiej C(2026) 5054 do art. 50 AI Act ujmują kryterium „istniejące” szerzej: obejmuje ono także osoby, które mogą realnie istnieć. Wytyczne weszły do klucza jako osobny element rubryki z zaznaczeniem, że są soft law (rozdział 4.4).

- **P09 i P10 (wersje 3 → 4 i 2 → 3, po biegu z 18 października 2026 r.).** Klucze przygotowane bez badania orzecznictwa stanowiły, że ogólna formuła nie przenosi praw do żadnego pola eksploatacji. Prawnik odczytał wyroki SN I CSK 104/18, I CSK 596/17 i IV CSKP 44/21: art. 41 ust. 2 pr.aut. nie wyklucza wykładni z art. 65 k.c., pól nie trzeba wyliczać enumeratywnie, a wątpliwości rozstrzyga się na korzyść twórcy. Odpowiedzi GPT-6.1 Sol i GPT-6 Astra z wyszukiwaniem, które wskazały tę linię, traciły według dawnej reguły B6 punkty, choć były trafniejsze od klucza. Rdzeń zmieniono na warunkowy, linia SN weszła do rubryki jako element nagradzany, a reguła B6 została uchylona; 20 kart P09 i P10 przeliczono. Wyrokiem źródłowym tej linii jest III CK 124/05; koordynator uznał go najpierw za niezwiązany z polami eksploatacji na podstawie fragmentu cytowanego w innym wyroku, a błąd wykrył agent oceniający, który przeczytał pełny tekst w SAOS.

Zmiana P09 i P10 nastąpiła po zebraniu odpowiedzi, więc w tej części zasada „klucz przed biegiem” nie

została dochowana; kierunek zmiany (klucz łagodniejszy wobec odpowiedzi trafniejszych od niego) i jej powód zapisano w kluczach.

2.4. Pytania poza zbiorem

Pytanie P02 (odsetki ustawowe za opóźnienie od kwot w CHF i PLN) wyłączyliśmy decyzją z 18 października 2026 r., zapisaną w polu status klucza: zbiór ma być jednorodny i obejmować pytania o treść prawa, a zadanie obliczeniowe ma trafić do osobnego testu, na przykład z narzędziem do liczenia. Klucz P02 pozostał bez zmian. P01 ma wprowadzić typ obliczenie, ale odpowiedź to kwota zapisana wprost w ustawie, więc zostało w zbiorze (decyzja z 25 października 2026 r.). Pytania P03 i P11, które mają klucze z pilotażu z 10 października 2026 r., nie weszły do rozszerzenia z 18 października, a powodu tej decyzji nie zapisaliśmy. Oba dotyczą kredytu konsumenckiego (wcześniejsza spłata); włączymy je do następnego biegu po sprawdzeniu kluczy według specyfikacji v1.0.1.

03 / SYSTEMY I WARUNKI BIEGU

Systemy i warunki biegu

Osiem modeli, warunki W0 i W1, ustawienia biegu; systemy kancelarii poza benchmarkiem.

3.1. Osiem modeli w dwóch warunkach

Pytania otrzymało osiem modeli trzech dostawców, każdy w dwóch warunkach. **W0** to bieg bez narzędzi: bez wyszukiwania, bez dostępu do tekstów aktów, odpowiedź wyłącznie z wag modelu. Mierzy wiedzę zapisaną w modelu. **W1** to bieg z wyszukiwaniem w

internecie, z poleceniem podania adresów źródeł. Mierzy przydatność w warunkach, w jakich prawnik zwykle korzysta z modelu. Warunek W1 dla wszystkich modeli wprowadziliśmy decyzją z 17 października 2026 r., aby porównanie z modelami, które miały wyszukiwanie już w pilotażu, było uczciwe.

Tabela 2. Systemy i ustawienia biegu r3

MODEL	IDENTYFIKATOR	DOSTĘP	USTAWIENIA	W1: NARZĘDZIA
GPT-6.1 Sol	gpt-6.1-sol	Codex CLI (konto ChatGPT)	rozumowanie medium; aplikacje konta wyłączone	wyszukiwanie Codex na żywo
GPT-6 Astra	gpt-6-astra	Codex CLI	jak wyżej	jak wyżej
GPT-5.6 Terra	gpt-5.6-terra	Codex CLI	jak wyżej	jak wyżej
Claude Opus 5.5	claude-opus-5-5 (alias opus)	Claude Code CLI (subskrypcja)	bez MCP i CLAUDE.md; temperatura domyślna	WebSearch + WebFetch
Claude Sonnet 5.5	claude-sonnet-5-5	Claude Code CLI	jak wyżej	jak wyżej
Claude Haiku 5.5	claude-haiku-5-5	Claude Code CLI	jak wyżej	jak wyżej
Gemini 3.8 Flash	gemini-3.8-flash	Vertex AI, region global	temperatura 0,1; rozumowanie domyślne	Google Search + URL context
Gemini 4 Argon	gemini-4-argon	Vertex AI, region global	temperatura 0,1; rozumowanie high	Google Search + URL context

Wszystkie modele dostały ten sam krótki wstęp („Jesteś asystentem prawnika w polskiej kancelarii. Odpowiadasz po polsku, na gruncie prawa polskiego i unijnego, podając podstawy prawne.”), a w W1 zdanie o wyszukiwaniu i adresach źródeł. Wstęp nie zawiera prośby o zwięzłość ani wskazówek o pułapkach i sporności; nie zmienialiśmy go w trakcie serii biegów. Każde pytanie przekazywano w osobnej sesji, w brzmieniu z pola tekst_benchmarkowy albo, gdy go nie ma, tekst_oryginalny. Odpowiedzi na pytania pilotażowe zebrano 10 października 2026 r. (modele OpenAI i Anthropic w W0 oraz GPT-6.1 Sol

w W1), 13 października (Gemini 4 Argon w W0) i 17–18 października (pozostałe), a na pytania rozszerzenia 18 października 2026 r. Każdy system odpowiadał raz.

Trzy cechy warunków wpływają na odczyt wyników. Po pierwsze, Claude Code przetwarza stronę pobraną narzędziem WebFetch małym modelem (Claude Haiku 5.5) i oddaje modelowi głównemu streszczenie, nie surowy tekst; Codex i Gemini czytają strony bez pośrednika. Przy błędach klasy T w odpowiedziach Claude W1 nie da się więc oddzielić błędu modelu od błędu streszczenia. Po drugie, w W0 Co-

dex działa z wyłączonymi aplikacjami konta ChatGPT, bo aplikacje te mają własną wyszukiwarkę (ustalenie z 8 października 2026 r.). Po trzecie, rozumowanie Gemini 4 Argon trzeba wymusić: bez ustawienia high model w pilotażu nie rozumował na pytaniach prawnych wcale, a wynik tej samej wersji modelu zmieniał się istotnie (średnio 38,9 % wobec 55,1 % w bloku A na pięciu pytaniach pilotażu). W raporcie podajemy więc tryb, nie tylko nazwę modelu.

3.2. Systemy kancelarii poza benchmarkiem

Decyzją z 17 października 2026 r. nie obejmujemy badaniem systemów własnych kancelarii (asystent

badawczy, czat i każdy kolejny produkt) ani w części z pytaniami, ani w części z kazusami. Powodem jest konflikt interesów: oceniający byłby zarazem autorem ocenianego systemu. Dotychczasowe biegi tych systemów przeniesiono do archiwum (124 pliki, nic nie usunięto); generatory wyników ich nie czytają. Raport nie obejmuje też systemów prawniczych innych dostawców, które oceniał Benchmark 01; tier pytań bada na razie modele ogólnego przeznaczenia, z wyszukiwaniem i bez.

04 / METODA OCENY

Metoda oceny

Bloki A i B, halucynacja zero-jedynkowa z klasami F/T/W/S, poprawne ugruntowanie, reguły graniczne z datami.

4.1. Dwa bloki i karta oceny

Jak w Benchmarku 01, oceniamy w dwóch blokach raportowanych osobno. **Blok A** (obiektywny) obejmuje: rdzeń odpowiedzi A-R (0, 1 albo 2 punkty; 1 oznacza kierunek trafny bez kluczowego warunku), podstawy wymagane A2 (0/1; zero także przy błędnej podstawie), dosłowność cytatów A3 (0/1, tylko gdy system cytuje), wskazanie sporności A5 (tylko przy pytaniu spornym), halucynacje A6 (0/1), ramy odpowiedzi A8 (0, ½ albo 1), elementy rubryki R (po punkcie) i kary K (po punkcie ujemnym). A % to suma punktów pomniejszona o kary i podzielona przez punkty możliwe, bez kryteriów, które nie mają zastosowania. **Blok B** (ocenny) obejmuje przydatność dla prawnika B1 (0–2) i kalibrację pewności B2 (0/1) względem pola klucza kalibracja_oczekiwana. $\Delta = A\% - B\%$ jest różnicą opisową.

Kartę oceny każdej odpowiedzi wypełnia najpierw model Claude, sprawdzając każde powołane źródło w tekście urzędowym (ISAP i ELI, EUR-Lex, SAOS, sn.pl, portale orzeczeń sądów powszechnych, serwisy gov.pl), a następnie prawnik rozstrzyga sporne punkty i zatwierdza kartę. Karta zawiera przy każdym kryterium punkt i dowód, a przy A6 listę sprawdzonych powołań z wynikiem i źródłem.

4.2. A6: halucynacja zero-jedynkowa i klasy błędów

Od 11 października 2026 r. A6 przyjmuje wartość 1 albo 0. Odpowiedź dostaje 0 i etykietę *hallucinated*, gdy zawiera co najmniej jedno twierdzenie o źródle prawa niezgodne ze źródłem urzędowym na datę

stanu prawnego. Typ błędu zapisujemy w polu A6_typ, bez wpływu na punkt:

- **F** — źródło albo jednostka redakcyjna nie istnieje;
- **T** — wskazanemu przepisowi lub orzeczeniu przypisano treść, której w nim nie ma: odwrotną, cudzą, o rozciągniętym lub zawężonym zakresie, dla innego adresata; także teza przypisana ogólnie „orzecznictwu” sprzeczna z wyrokiem zapisanym w kluczu;
- **W** — przepis w brzmieniu z innej daty podany jako obowiązujący, także błędny opis historii zmian;
- **S** — prawdziwe źródło z przekreślonym identyfikatorem (numer, data, sąd, publikator, nazwa aktu).

Klasę O z wcześniejszej typologii zniesiono: fałszywe twierdzenie o treści przepisu to T, a samo pominięcie przesłanki nie obniża A6, lecz może obniżyć ocenę rdzenia i elementów rubryki. Odpowiedź bez żadnego powołania ma A6 = 1. Reguła jest celowo surowsza niż miara, którą przyłożylibyśmy do człowieka. Prawnik, który napisze „§ 1¹” zamiast „§ 1²”, zwykle wie, o który przepis mu chodzi; model takiej pamięci zamiaru nie ma, a polska technika legislacyjna (indeksy górne, jednostki dodawane i uchylane nowelami) sprawia, że sąsiednie numery regulują inne sprawy.

Obok punktu karta podaje dwa pola opisowe. **A6_pochodzenie** mówi, skąd wziął się błąd: z pamięci (system nie otworzył źródła), z odczytu źródła (otworzył właściwe źródło urzędowe i oddał je błędnie), ze źródła wtórnego albo nieustalone. **A6_zakres** (od 18 października 2026 r.) mówi, czy błąd dotyczy rdzenia (głównej konkluzji, podstaw wymaganych, elementu rubryki), czy uwagi pobocznej. A6 pozostaje surowe

niezależnie od zakresu: jeden błąd klasy F, T, W lub S w dowolnym miejscu odpowiedzi daje zero.

4.3. Etykieta „bez weryfikacji źródła” i poprawne ugruntowanie

Etykieta **bez weryfikacji źródła** jest niezależna od punktów i ma cztery warianty: *brak narzędzi* (bieg W0), *narzędzia nieużyte* (W1, ale ślad narzędzi i źródła nie pokazują, że system z nich skorzystał; wariant z 18 października 2026 r.), *tylko źródła wtórne* (serwisy prawnicze, blogi, portale) oraz *nie*, gdy w śladzie lub w źródłach odpowiedzi jest co najmniej jedno źródło urzędowe właściwe dla pytania. Etykietę ustalamy ze śladu narzędzi i metadanych biegu, nie z deklaracji systemu.

Główną miarą raportu jest **odpowiedź poprawnie ugruntowana**: karta ma A6 = 1 i A-R = 2. To połączenie dwóch warunków z planu wersji 1.1 specyfikacji: brak fałszywego twierdzenia o prawie i pełny rdzeń. Miara nie wymaga, aby system sam dotarł do źródła urzędowego; tę cechę pokazuje osobno etykieta „bez weryfikacji źródła”.

4.4. Reguły szczególne

- **Soft law** (18 października 2026 r.). Wytyczne Komisji, kodeksy praktyk i opinie organów wchodzi do klucza jako osobny element rubryki, nigdy jako kara. Halucynacją jest przypisanie im mocy wiążącej albo treści, której w nich nie ma; nieścisłość co do etapu przyjęcia, gdy same źródła publiczne są niezgodne, halucynacją nie jest.

- **Blokada dostawcy** (18 października 2026 r.). Gdy odpowiedź jest pusta, bo usługa zablokowała wygenerowany tekst (Vertex, `finish_reason = RECITATION` albo `SAFETY`), A6 ma wartość „n/d”, bez etykiety hallucinated, a pozostałe kryteria 0. Bieg powtarza się raz; dwie blokady z rzędu są wynikiem. W zbiorze jest jeden taki przypadek: Gemini 4 Argon w W1 na P07, dwie próby, obie zablokowane.
- **Koszt** (18 października 2026 r.). Dla każdej odpowiedzi liczymy koszt równoważny API według publicznych cenników z 18 października 2026 r. Dla Claude Code i Codex z subskrypcji to równoważnik, dla Vertex koszt faktyczny. Główną metryką jest koszt na odpowiedź poprawnie ugruntowaną. Koszt nie wpływa na punkty. Znak „≥” oznacza dolną granicę: Claude Code nie zapisuje liczby wyszukiwań, a skrypt Vertex w biegu r3 nie zapisywał tokenów treści z URL context.

4.5. Reguły graniczne i kalendarz decyzji

Reguły graniczne rozstrzygają przypadki, w których definicja klasy nie wystarcza. Każdą z nich ustalił prawnik na konkretnej karcie, a karta ta służy dalej sędziom jako wzorzec. Tabela 3 zestawia decyzje z okresu 17–25 października 2026 r., które wpłynęły na karty zbioru. Decyzje z 9–11 października (A6 zero-jedynkowe, klasy F/T/W/S, zniesienie klasy O, etykieta „bez weryfikacji źródła”) poprzedzają ten okres i opisałyśmy je w rozdziałach 4.2 i 4.3.

Tabela 3. Decyzje z 17–26 października 2026 r. (pełne brzmienie decyzji w specyfikacji oceny, która trafi do repozytorium LexAlpha Lab na GitHub; aneks B, „Materiał źródłowy”)

DATA	DECYZJA	SKUTEK DLA KART
17.10	Przykłady graniczne A8 nr 1–3 i B1 nr 1 (dygresja wynikająca z błędnej tezy; błędna konstrukcja w odpowiedzi na zadane pytanie; pomyłony rodzaj orzeczenia; krok kontrolny przy tezie grożącej utratą terminu)	zmiany punktów w 5 kartach, etykiety bez zmian
17.10	Systemy kancelarii poza benchmarkiem; W1 dla wszystkich modeli	124 pliki w archiwum; 80 nowych biegów

DATA	DECYZJA	SKUTEK DLA KART
18.10	A6_zakres (rdzeń / poboczne); blokada dostawcy; soft law w kluczach; metryka kosztu; wariant „narzędzia nieużyte”	nowe pola w kartach rozszerzenia
18.10	B8: błędna historia zmian przepisu = W, rdzeń, gdy historia jest elementem rubryki	P01
18.10	P09 v4 i P10 v3: rdzeń warunkowy, linia SN jako nagroda, uchylenie B6	20 kart przeliczonych
18.10	Decyzje nr 1–34 na kartach rozszerzenia; B10 (stara redakcja bez zmiany normy = tylko A3), B10a (stare brzmienie ze zmianą merytoryczną = W); nr 19 (zmiana zwrotu definiowanego w cytacie = A3 0); B11 (B2 przy kalibracji warunkowej P09/P10); B12 (linia orzecznicza — wystarczy wykładnia z art. 65 k.c.); bez podwójnej kary w bloku B; A8 = 1 za trafne uwagi poboczne	59 z 80 kart zmienionych
19.10	Zalecenie redakcyjne podane jako wymóg ustawowy = A8 ½, bez T; zawężenie definicji wbrew wytycznym = błąd wykładni w A-R, bez T	5 kart P08, 1 karta P09
22.10	Sąsiednia jednostka tego samego artykułu, gdy niesie istotę przypisaną treści = nieściśłość bez T; jednostka uchylona z jej dawną prawdziwą treścią = W, nie F	P10 (Gemini 4 Argon W1, GPT-5.6 Terra W1)
23.10	B9 w wersji wąskiej: zdanie o orzecznictwie bez sygnatury = T tylko wtedy, gdy wyrok z klucza wprost odrzuca przypisaną tezę; twierdzeń o „jednolitości” linii nie karzemy	Claude Opus 5.5 W1 P10: A6 0 → 1
24.10	Zawężony lub rozszerzony krąg adresatów albo zakres normy: W, gdy pokrywa się z brzmieniem z innej daty; T, gdy nie odpowiada żadnemu brzmieniu; samo niewspomnienie to pominięcie	reguła także dla warstwy weryfikacji
24.10	Warstwa weryfikacji: S na orzeczeniach z oceną każdego kandydata przy sygnaturze niejednoznacznej; u.n.n.c. do legal-cite; raport dla sędziego może dodać błąd, nie może go odjąć	rozdział 7
25.10	Karta meta v3: „✓” tylko przy brzmieniu na datę potwierdzonym; rejestrowanie każdego wywołania narzędzia sędziego	rozdział 7.5
25.10	A6_zakres mówi o miejscu błędu, nie o skutku dla punktów: błąd w treści elementu rubryki = rdzeń także przy zaliczonym elemencie; S przy zgodnej tezie = poboczne, chyba że orzeczenie jest w podstawach wymaganych	pole uzupełnione w 49 kartach pilotażu
26.10	Publikacja: pierwszych 10 pytań z kluczami, specyfikacją i kartami wzorcowymi ukaże się razem z metodologią po przeglądzie kart; reszta puli 50 pytań zamknięta. Karty wersjonowane datą ostatniej kontroli zbioru (ta wersja: 25.10.2026)	oświadczenie o danych w artykule v1.2; aneks E

Reguły zmienialiśmy po lekturze odpowiedzi, w tym trzykrotnie regułę A6 w dniach 9–11 października 2026 r. Z tego powodu zbiór ma status kalibracyjny, a wersję specyfikacji zapisujemy w każdej karcie

(obecnie 1.0.1 z doprecyzowaniami z 17–24 października).

Wyniki modeli

160 odpowiedzi: wyniki, wpływ wyszukiwania, koszt, klasy błędów i przykłady z kart.

5.1. Wyniki zbiorcze

Tabela 4 zestawia wyniki szesnastu biegów (osiem modeli w dwóch warunkach) na dziesięciu pytaniach. Kolumny A %, B % i „Halucynacje” pochodzą z gene-

ratora zestawień czytającego karty ocen, liczba odpowiedzi poprawnie ugruntowanych z tych samych kart, a koszt z zestawienia kosztów na odpowiedź (cennik z 18 października 2026 r.).

Tabela 4. Wyniki ośmiu modeli w warunkach W0 i W1 (10 odpowiedzi w każdym wierszu; poprawnie ugruntowana: A6 = 1 i A-R = 2; „≥” = dolna granica kosztu; „sprawdzone powołania” = suma z wierszy A6 kart)

MODEL	WARUNEK	A %	B %	HALUCYNACJE (Z 10)	POPRAWNIE UGRUNTOWANE (Z 10)	KOSZT ŁĄCZNY USD	USD NA POPRAWNIE UGRUNTOWANĄ	SPRAWDZONE POWOŁANIA
GPT-6.1 Sol	W0	72,8	86,7	1	6	0,250	0,042	51 z 51
GPT-6.1 Sol	W1	84,0	96,7	1	7	≥ 1,544	≥ 0,221	72 z 73
GPT-6 Astra	W0	75,2	83,3	1	6	1,415	0,236	59 z 59
GPT-6 Astra	W1	82,5	93,3	0	7	3,816	0,545	63 z 63
GPT-5.6 Terra	W0	41,7	66,7	6	3	0,206	0,069	41 z 41
GPT-5.6 Terra	W1	72,7	86,7	4	4	1,321	0,330	70 z 70
Claude Opus 5.5	W0	55,0	73,3	9	0	0,961	n/d	123 z 123
Claude Opus 5.5	W1	86,0	96,7	5	4	≥ 1,467	≥ 0,367	147 z 147
Claude Sonnet 5.5	W0	43,5	60,0	9	0	0,613	n/d	93 z 93
Claude Sonnet 5.5	W1	67,8	90,0	9	0	≥ 0,842	n/d	104 z 111
Claude Haiku 5.5	W0	30,1	56,7	9	0	0,029	n/d	62 z 62
Claude Haiku 5.5	W1	49,5	66,7	8	0	≥ 0,225	n/d	75 z 77
Gemini 3.8 Flash	W0	49,6	70,0	8	2	0,116	0,058	75 z 77
Gemini 3.8 Flash	W1	78,1	86,7	5	4	≥ 0,514	≥ 0,128	97 z 97
Gemini 4 Argon	W0	61,0	80,0	8	2	0,372	0,186	115 z 121
Gemini 4 Argon	W1	75,3	86,7	8	1	≥ 2,085	≥ 2,085	189 z 189

Ze 160 odpowiedzi 91 zawierało co najmniej jedną halucynację: 51 z 80 w W0 i 40 z 80 w W1. Poprawnie ugruntowanych było 46: 19 w W0 i 27 w W1. Rozpiętość między modelami jest duża: od 0 do 7 odpowiedzi poprawnie ugruntowanych na 10. GPT-6.1 Sol i GPT-6 Astra miały w obu warunkach

najwyżej po jednej odpowiedzi z halucynacją i po 6–7 odpowiedzi poprawnie ugruntowanych. Claude Sonnet 5.5 i Claude Haiku 5.5 nie dały ani jednej odpowiedzi poprawnie ugruntowanej, ani bez narzędzi, ani z wyszukiwaniem.

Wysoki wynik w bloku A nie wyklucza halucynacji. Claude Opus 5.5 w W1 ma najwyższą średnią A % w zestawieniu (86,0 %) i jednocześnie pięć odpowiedzi z halucynacją. Claude Sonnet 5.5 w W1 ma B % 90,0 i dziewięć odpowiedzi z halucynacją na die-sięć. Profil „trafny rdzeń, wadliwe powołanie” jest w

tym zbiorze częsty: z 93 odpowiedzi z pełnym rdze-niem (A-R = 2) 47 zawierało co najmniej jedno powo-łanie niezgodne ze źródłem urzędowym, więc po-prawnie ugruntowanych zostało 46 (aneks B).

5.2. Wyszukiwanie: co zmienia, a czego nie

Tabela 5. Ten sam model bez narzędzi i z wyszukiwaniem

MODEL	A % W0 → W1	B % W0 → W1	HALUCYNACJE W0 → W1	POPRAWNIE UGRUNTOWANE W0 → W1	USD NA POPRAWNIE UGRUNTOWANĄ W0 → W1
GPT-6.1 Sol	72,8 → 84,0	86,7 → 96,7	1 → 1	6 → 7	0,042 → ≥ 0,221
GPT-6 Astra	75,2 → 82,5	83,3 → 93,3	1 → 0	6 → 7	0,236 → 0,545
GPT-5.6 Terra	41,7 → 72,7	66,7 → 86,7	6 → 4	3 → 4	0,069 → 0,330
Claude Opus 5.5	55,0 → 86,0	73,3 → 96,7	9 → 5	0 → 4	n/d → ≥ 0,367
Claude Sonnet 5.5	43,5 → 67,8	60,0 → 90,0	9 → 9	0 → 0	n/d → n/d
Claude Haiku 5.5	30,1 → 49,5	56,7 → 66,7	9 → 8	0 → 0	n/d → n/d
Gemini 3.8 Flash	49,6 → 78,1	70,0 → 86,7	8 → 5	2 → 4	0,058 → ≥ 0,128
Gemini 4 Argon	61,0 → 75,3	80,0 → 86,7	8 → 8	2 → 1	0,186 → ≥ 2,085

Średni wynik bloku A wzrósł z wyszukiwaniem u wszystkich ośmiu modeli, o 7,3 pp (GPT-6 Astra) do 31,0 pp (GPT-5.6 Terra i Claude Opus 5.5). Liczba odpowiedzi z halucynacją spadła wyraźnie u trzech modeli: Claude Opus 5.5 (z 9 do 5), Gemini 3.8 Flash (z 8 do 5) i GPT-5.6 Terra (z 6 do 4). U Claude Sonnet 5.5 i Gemini 4 Argon nie zmieniła się, u Clau-de Haiku 5.5 spadła o jedną. Liczba odpowiedzi po-prawnie ugruntowanych wzrosła o 4 u Claude Opus 5.5, o 2 u Gemini 3.8 Flash, o 1 u modeli OpenAI, nie zmieniła się u Claude Sonnet 5.5 i Claude Haiku 5.5, a u Gemini 4 Argon spadła z 2 do 1. W badanej serii średni wynik A był więc wyższy w W1 u wszystkich modeli, a liczba odpowiedzi z halucynacją zmniejszy-

ła się u pięciu; blok A obejmuje także A6, więc obie miary nie są niezależne.

Dostęp do wyszukiwarki nie oznacza, że model z niej korzysta. W W1 w 50 z 80 odpowiedzi ślad narzędzi albo adresy w odpowiedzi pokazują co najmniej jed-no źródło urzędowe właściwe dla pytania; 18 odpo-wiedzi opierało się wyłącznie na źródłach wtórnych (serwisy prawnicze, blogi, portale), a w 12 model nie użył narzędzi, choć je miał (12 kart ma etykietę „na-rzędzia nieużyte”; historia poprawki etykiety P12 Gemini 3.8 Flash w rozdziale 8.2). Tabela 6 pokazu-je, że zależy to od modelu: GPT-6.1 Sol i GPT-6 Astra sięgnęły po źródło urzędowe w każdej odpo-wiedzi, modele Claude w 3–4 odpowiedziach na 10.

Tabela 6. Źródła w odpowiedziach W1 i pochodzenie halucynacji (etykieta „bez weryfikacji źródła” i pole A6_poc
hodzenie z kart)

MODEL (W1)	ŹRÓDŁO URZĘDOWE UŻYTE	TYLKO ŹRÓDŁA WTÓRNE	NARZĘDZIA NIEUŻYTE	„BRAK NARZĘDZI” W KARCIE W1	HALUCYNACJE: PAMIĘĆ / ODCZYT ŹRÓDŁA / ŹRÓDŁO WTÓRNE / NIEUSTALONE
GPT-6.1 Sol	10	0	0	0	0 / 1 / 0 / 0
GPT-6 Astra	10	0	0	0	0 / 0 / 0 / 0
GPT-5.6 Terra	9	1	0	0	3 / 1 / 0 / 0
Claude Opus 5.5	4	4	2	0	4 / 0 / 0 / 1
Claude Sonnet 5.5	4	2	4	0	6 / 0 / 0 / 3
Claude Haiku 5.5	3	4	3	0	3 / 1 / 3 / 1
Gemini 3.8 Flash	3	4	3	0	1 / 1 / 1 / 2
Gemini 4 Argon	7	3	0	0	4 / 3 / 0 / 1

Z 40 odpowiedzi W1 z halucynacją 21 ma błąd z pa-
mięci: model miał wyszukiwarkę, ale błędnego źródła
nie otworzył. W 7 model otworzył właściwe źródło
urzędowe i mimo to oddał je błędnie (pochodzenie
„odczyt źródła”), w 4 powtórzył błąd ze źródła wtórne-
go, w 8 pochodzenia nie dało się ustalić: w 5 odpo-
wiedziach modeli Claude, bo Claude Code nie za-
chowuje śladu narzędzi, i w 3 odpowiedziach modeli
Gemini (2 Gemini 3.8 Flash, 1 Gemini 4 Argon), w
których ślad nie wskazywał źródła błędu. W W0
wszystkie 51 halucynacji ma z definicji pochodzenie
„pamięć”.

**5.3. Koszt na odpowiedź poprawnie
ugruntowaną**

Łączny koszt równoważny API 160 odpowiedzi wy-
niósł co najmniej 15,78 USD: 3,96 USD w W0 i co
najmniej 11,81 USD w W1. Najniższy koszt serii dzie-
sięciu odpowiedzi w przeliczeniu na jedną odpowiedź
poprawnie ugruntowaną wyniósł 0,042 USD
(GPT-6.1 Sol w W0), kolejne 0,058 USD (Gemini 3.8
Flash w W0) i 0,069 USD (GPT-5.6 Terra w W0); naj-
wyższy co najmniej 2,085 USD (Gemini 4 Argon w
W1, jedna taka odpowiedź na dziesięć). Wskaźnik
dzieli koszt całej serii przez liczbę odpowiedzi po-
prawnie ugruntowanych, nie jest ceną pojedynczej
odpowiedzi. Bieg z wyszukiwaniem kosztował od 1,4

raza (Claude Sonnet 5.5) do 7,7 raza (Claude Haiku
5.5) więcej niż bieg bez narzędzi, przy zmianie liczby
odpowiedzi poprawnie ugruntowanych od -1 do +4.
Dla modeli bez odpowiedzi poprawnie ugruntowanej
metryka ma wartość „n/d”. Mediana czasu odpowie-
dzi wynosiła od 21,0 s (Claude Haiku 5.5 w W0) do
169 s (Gemini 4 Argon w W1).

Koszt jest równoważnikiem, a nie wydatkiem kance-
larii w przypadku Claude Code i Codex, które działa-
ły z subskrypcji; dla Vertex AI jest kosztem faktycz-
nym, w cenach promocyjnych obowiązujących 18
października 2026 r. Wartości W1 Claude i Gemini są
dolną granicą (rozdział 4.4). Pełny koszt badania,
razem z sędziami i warstwą weryfikacji, podaje
rozdział 7.7.

5.4. Zasięg weryfikacji

Karty podają, ile powołań sprawdzono w źródle urzę-
dowym. Na 160 kartach sprawdzono 1 436 z 1 454
powołań (98,8 %); jedna karta (blokada dostawcy)
nie ma powołań. Pozostałe 18 powołań w 11 kartach
ma status „nie znaleziono”, „nieobsługiwany” albo
„do sprawdzenia ręcznie” i nie wchodzi do błędów
modelu. Wśród 46 odpowiedzi poprawnie ugruntowa-
nych dwie mają po jednym takim powołaniu (P10
GPT-6.1 Sol W1, 7 z 8 sprawdzonych; P13 Gemini 4

Argon W0, 6 z 8); ich etykieta opiera się na powołaniach sprawdzonych. Liczba powołań na odpowiedź różni się między modelami ponad czterokrotnie: od 41 w dziesięciu odpowiedziach GPT-5.6 Terra w W0 do 189 w odpowiedziach Gemini 4 Argon w W1. Model, który powołuje więcej, ma więcej okazji do błędu

przy regule „jedna zła jednostka = 0”; to argument za raportowaniem obok etykiety także udziału poprawnych powołań w odpowiedzi.

5.5. Klasy błędów

Tabela 7. Klasy błędów w odpowiedziach z halucynacją (liczba odpowiedzi zawierających daną klasę; jedna odpowiedź może mieć kilka klas; wartość bieżąca pola A6_typ)

KLASA	W0 (80 ODPOWIEDZI)	W1 (80 ODPOWIEDZI)	RAZEM
T — treść, której w źródle nie ma	42	24	66
W — brzmienie z innej daty	18	17	35
F — źródło lub jednostka nie istnieje	4	4	8
S — zniekształcony identyfikator	4	5	9
odpowiedzi z halucynacją	51	40	91

Najczęstszą klasą jest T: prawdziwy przepis lub wyrok z treścią, której w nim nie ma (66 z 91 odpowiedzi z halucynacją). Drugą jest W, brzmienie z innej daty podane jako obowiązujące (35). Zmyślane źródło (F, 8 odpowiedzi) i zniekształcony identyfikator (S, 9) są rzadkie. Wyszukiwanie zmniejszyło liczbę odpowiedzi z klasą T z 42 do 24, a liczby odpowiedzi z klasą W prawie nie zmieniło (18 i 17). Liczba odpowiedzi z błędem W była więc podobna w obu warunkach. W kartach W1 z błędem W model podawał treść z pamięci albo ze starszego tekstu znalezione go w sieci; udziału starszych tekstów w wynikach wyszukiwania nie mierzyliśmy.

Pole A6_zakres wypełniono w kartach pięciu pytań rozszerzenia. Z 42 odpowiedzi z halucynacją na tych pytaniach 22 mają błąd w rdzeniu, a 20 tylko w uwadze pobocznej przy trafnej konkluzji i jej podstawach. W 49 kartach z halucynacją na pytaniach pilotażowych pole uzupełniono 25 października 2026 r. (decyzja z tego dnia; wpis do przeglądu prawnika w każdej karcie): 35 ma błąd w rdzeniu, a 14 tylko w uwadze pobocznej. Łącznie na dziesięciu pytaniach 57 z 91 odpowiedzi z halucynacją ma błąd w rdzeniu, a 34 tylko w uwadze pobocznej. Dwie reguły przyjęte

tego dnia rozstrzygnęły siedem przypadków granicznych: błąd w treści elementu rubryki to rdzeń także wtedy, gdy element zaliczono, a zniekształcony identyfikator orzeczenia przy zgodnej tezie to uwaga poboczna, chyba że orzeczenie należy do podstaw wymaganych. W 4 kartach pilotażu rozstrzygnięcie pozostaje oznaczone jako niepewne.

5.6. Typowe błędy: przykłady z kart

Przykłady wybraliśmy tak, aby każdy ilustrował inny mechanizm. Odesłanie w nawiasie wskazuje kartę o ceny/<pytanie>_<system>.md.

- **Stary stan prawny w wagach i w wyszukiwarce (W).** W P01 reguła „ $\frac{1}{4}$ minimalnego wynagrodzenia”, obowiązująca do 14 listopada 2018 r., pojawiła się jako stan obecny w trzech niezależnych odpowiedziach: Claude Haiku 5.5 i Claude Sonnet 5.5 bez narzędzi oraz GPT-5.6 Terra z wyszukiwaniem. GPT-5.6 Terra bez narzędzi podał poprawnie 800 zł, a z wyszukiwaniem przypisał „ $\frac{1}{4}$ płacy minimalnej” art. 119 § 1, 120 § 1, 122 § 1 i 124 § 1 k.w.; jest to brzmienie z tekstu jednolitego z 2015 r. Wynik tej odpowiedzi w bloku A spadł z 78 % do 0 % (P01_codex_terra56_w0_r3, P01_codex_terra56_web_r3). Odpowiedź W1 zawierała regułę odpowiadającą dawnemu brzmieniu przepisu, podczas gdy odpowiedź W0 podawała aktualną kwotę; możliwym wyjaśnieniem jest wykorzystanie nieaktualnego źródła.
- **Przepis uchylony jako podstawa (W).** W P01 art. 278a k.k. powołano jako obowiązującą podstawę w czterech odpowiedziach (Claude Opus 5.5 w W0 i W1, Gemini 3.8 Flash i Gemini 4 Argon w W1); przepis w tej treści obowiązywał do 30 września 2023 r. (P01_opus_cli_w0_r3).
- **Brzmienie sprzed noweli ze zmianą merytoryczną (W, reguła B10a).** Art. 44 pr.aut. opisany jako „rażąca dysproporcja” wynagrodzenia — tak brzmiał przed nowelą z 26 lipca 2024 r. (Dz.U. 2024 poz. 1254); obecnie przepis mówi o wynagrodzeniu „niewspółmiernie niskim” (P10_opus_cli_w0_r3, błąd poboczny).

- **Krąg adresatów sprzed zmiany (W, reguła z 24 października 2026 r.).** W P05 opis art. 131^{1a} § 1 k.p.c. ograniczający doręczenia przez portal do profesjonalnych pełnomocników odpowiada brzmieniu sprzed 1 czerwca 2026 r. (P05_codex_terra56_w0_r3). W P07 art. 25b u.k.s.c. z opłatą 100 zł od każdego postanowienia odpowiada tekstowi jednolitemu z 2020 r.; obecnie 100 zł dotyczy tylko wyroku i postanowienia co do istoty (m.in. P07_codex_terra56_web_r3).
- **Prawdziwa sygnatura, cudza sprawa (T).** Wyrok SN II CSKP 526/22 powołany na tezę o polach eksploatacji, z datą 14 września 2022 r.; według bazy orzeczeń to wyrok z 12 stycznia 2023 r. w sprawie o zadośćuczynienie (P10_argon4_high_vertex_w0_r3, błąd poboczny). Wyrok SN III CK 400/03 dotyczy pól eksploatacji jako elementu przedmiotowo istotnego umowy, ale nie zawiera przypisanej mu tezy o bezskuteczności klauzuli ogólnej (P09_gemini38flash_vertex_w1_r3, błąd w rdzeniu).
- **Właściwe źródło otwarte, treść zmieniona (T, odczyt źródła).** GPT-6.1 Sol z wyszukiwaniem podał adres pełnego tekstu postanowienia SN I CZ 83/09 w SAOS i przypisał mu tezę, że SN „potwierdził” stanowisko w kwestii, do której wcześniejsza ocena według tekstu postanowienia się nie odnosiła (P13_chatgpt_codex_web_r3).
- **Właściwe źródło otwarte, zły numer (S, odczyt źródła).** Gemini 4 Argon z wyszukiwaniem otworzył wyrok SA w Łodzi I ACa 1157/12 z 27 lutego 2013 r. i trafnie oddał jego tezę, ale w tekście wpisał sygnaturę I ACa 1188/12, czyli innego wyroku tego sądu (P10_argon4_high_vertex_w1_r3). GPT-5.6 Terra z wyszukiwaniem podał właściwe rekordy SAOS z błędnym sądem: I ACz 390/13 (SA w Rzeszowie, nie w Szczecinie) i VIII Gz 43/16 (SO w Szczecinie, nie SA w Gdańsku) (P13_codex_terra56_web_r3).

- **Ogólnik o orzecznictwie sprzeczny z wyrokiem z klucza (T, reguła B9).** Claude Opus 5.5 w W1 (P09) i Gemini 3.8 Flash w W0 (P09 i P10) przypisały „orzecznictwu” tezę, że ogólna formuła nie przenosi praw do żadnego pola; linia SN (III CK 124/05, I CSK 596/17, IV CSKP 44/21) mówi odwrotnie. Po zawężeniu reguły B9 23 października 2026 r. jedna podobna teza (P10 Claude Opus 5.5 w W1) przestała być halucynacją, bo żaden wyrok z klucza jej wprost nie odrzuca.
- **Sąsiednia jednostka z cudzą treścią (T).** W P09 Claude Opus 5.5 w W1 powołał „art. 64” pr.aut. z treścią art. 52 ust. 1 (P09_opus_cli_w1_r3).
- **Blokada dostawcy.** Gemini 4 Argon z wyszukiwaniem na P07 dwukrotnie oddał pustą odpowiedź z sygnałem RECITATION: Vertex blokuje tekst, gdy model zbyt dosłownie przytacza znaleziony tekst. System z wyszukiwaniem może więc zamilknąć wtedy, gdy znalazł i cytuje przepis (P07_argon4_high_vertex_w1_r3).

Dwa pytania pokazują granice rdzenia jako miary. W P04 żaden z 16 systemów nie dał pełnego rdzenia (średnio 43,0 % w bloku A): prawidłowa odpowiedź wymaga zestawienia k.p.a., p.p.s.a. i ustawy o nabywaniu nieruchomości przez cudzoziemców oraz znajomości instytucji zaświadczenia, a w pierwszym przeglądzie kart tego pytania (25 kart, w tym karty systemów spoza tego raportu) dominowała odpowiedź „ostateczność” albo „zasadniczo ostateczność, ewentualnie prawomocność” — także z wyszukiwaniem. W P07 Claude Haiku 5.5 odpowiedział trafnie „tak”, ale w całości o wyroku i apelacji zamiast o postanowieniu i zażaleniu. Sonda z 24 zapytaniemi (18 z nich opisujemy) pokazała powtarzalny wzorec za-

leżny od sformułowania pytania: na oryginalne pytanie model 10 razy na 10 odpowiedział „tak” i 10 razy na 10 o wyroku; gdy pytanie wprost nazywało postanowienie, odpowiedział poprawnie 5 razy na 5; zapytany wprost o art. 357 § 5 k.p.c. trzy razy na trzy odpowiedział, że nie wie. Odpowiedź była trafna, lecz dotyczyła innej instytucji, czego ocena samego rdzenia nie wykrywa (zapis sondy trafi do repozytorium, aneks B).

5.7. Co wynika z rozdziału 5

- Na tym zbiorze 46 ze 160 odpowiedzi (29 %) jest poprawnie ugruntowanych. Halucynacja w rozumieniu przyjętym w raporcie dotyczy 91 odpowiedzi (57 %), przy czym na pytaniach rozszerzenia blisko połowa takich odpowiedzi (20 z 42) ma błąd tylko w uwadze pobocznej.
- Wyszukiwanie podnosi wynik bloku A u wszystkich modeli, ale liczba odpowiedzi poprawnie ugruntowanych rośnie o 4 tylko u jednego (Claude Opus 5.5), a u jednego spada (Gemini 4 Argon). Raportowanie wyniku W1 bez informacji, czy model użył narzędzi i jakiego źródła, miesza dwa różne zjawiska.
- Liczba odpowiedzi z błędem W była podobna w obu warunkach (18 w W0, 17 w W1), podczas gdy liczba odpowiedzi z klasą T spadła z 42 do 24. Kryterium aktualności brzmienia na datę chroni także przed „ugruntowaniem” w nieaktualnym źródle.
- Część błędów powstaje po otwarciu właściwego źródła (7 z 40 halucynacji w W1). Dostęp do źródeł na wejściu nie zastępuje kontroli powołań na wyjściu.

Sędziowie-modele

Pięciu sędziów-modeli wobec kart prawnika: κ , wyszukiwania, czego nie widzą, preferencja rodziny, panel.

6.1. Procedura

Kartę każdej odpowiedzi wypełniał najpierw model Claude ze sprawdzeniem powołań w źródłach urzędowych, a sprawdzał i zatwierdzał ją prawnik (rozdział 4.1). Sprawdzone karty są w tym rozdziale punktem odniesienia. Te same odpowiedzi oceniało niezależnie pięciu sędziów-modeli: Gemini 4 Argon (rozumowanie high), Gemini 3.1 Pro i Gemini 3.8 Flash na Vertex AI (region global, temperatura 0, z Google Search) oraz GPT-6.1 Sol i GPT-6 Astra przez Codex CLI (wyszukiwanie na żywo, rozumowanie medium, tryb tylko do odczytu). Czterech sędziów (Gemini 4 Argon, Gemini 3.8 Flash, GPT-6.1 Sol, GPT-6 Astra) jest zarazem ocenianymi modelami; w części kart sędzia ocenia więc własną odpowiedź albo odpowiedź modelu własnej rodziny.

Każdy sędzia dostawał w jednym poleceniu trzy elementy: reguły oceny (bloki A i B wraz z przykładami granicznymi), pełny klucz pytania oraz tekst odpowiedzi bez nazwy systemu. Nie widział karty ani ocen innych sędziów. Polecenie wymagało sprawdzenia każdego powołanego przepisu i orzeczenia w źródle urzędowym przez wyszukiwanie (ISAP i ELI, EUR-Lex, SAOS, sn.pl), porównania numeru jednostki i

treści dosłownie, nieoceniania z pamięci i wpisania do osobnego pola „niesprawdzone” każdego powołania, którego nie udało się sprawdzić. Sędzia oddawał kartę w tym samym układzie co karta wzorcowa. Zgodność mierzymy κ Cohena dla etykiety hallucinated; etykietę karty czytamy z ostatniej linii wyniku karty („** Δ = ...”), aby uwzględnić zmiany naniesione po decyzjach.

Zbiór dzieli się na dwie części o różnej wartości dowodowej. Karty pytań pilotażowych (P05, P06, P07, P12, P13) prawnik sprawdzał po obejrzeniu ocen sędziów, a sporne punkty rozstrzygał często na korzyść argumentu sędziego; zgodność na tej części jest zawyżona. Karty pytań rozszerzenia (P01, P04, P08, P09, P10) sprawdzono przed uruchomieniem sędziów, z wyjątkiem siedmiu rozbieżności GPT-6.1 Sol rozstrzygniętych 19 października 2026 r.; ta część jest właściwym testem. Sędziowie na pytaniach rozszerzenia zaczynali przed dopisaniem do specyfikacji reguł B9–B12 i znali je tylko z kluczy. GPT-6 Astra ocenił z powodu limitu konta 43 karty pilotażu i 6 kart rozszerzenia, dlatego jego wyników nie interpretujemy.

6.2. Zgodność z kartą prawnika

Tabela 8. Zgodność etykiety hallucinated sędziów-modeli z kartami (16 systemów z tabeli 4; fałszywy alarm: sędzia „tak”, karta „nie”; pominięcie: karta „tak”, sędzia „nie”)

SĘDZIA	ZBIÓR	N	ZGODNOŚĆ	κ	FAŁSZYWE ALARMY	POMINIĘCIA	WYSZUKAŃ NA KARTĘ	KARTY BEZ WYSZUKANIA	„NIESPRAWDZONE” NA KARTĘ
Gemini 4 Argon	rozszerzenie (P01, P04, P08–P10)	79	95 %	0,90	1	3	12,2	0	0,01
Gemini 4 Argon	pilotaż (P05–P07, P12, P13)	68	99 %	0,97	0	1	16,2	1	0,01

SĘDZIA	ZBIÓR	N	ZGODNOŚĆ	K	FAŁSZYWE ALARMY	POMINIĘCIA	WYSZUKAŃ NA KARTĘ	KARTY BEZ WYSZUKANIA	„NIESPRAWDZONE” NA KARTĘ	
GPT-6.1 Sol	rozszerzenie (P01, P04, P08–P10)	80	90 %	0,80	7	1	11,2	0	0,36	
GPT-6.1 Sol	pilotaż (P05–P07, P12, P13)	80	92 %	0,84	6	0	13,5	0	0,38	
Gemini 3.1 Pro	rozszerzenie (P01, P04, P08–P10)	80	90 %	0,80	0	8	4,3	25	0,00	
Gemini 3.1 Pro	pilotaż (P05–P07, P12, P13)	79	86 %	0,71	5	6	7,6	9	0,08	
Gemini 3.8 Flash	rozszerzenie (P01, P04, P08–P10)	80	79 %	0,58	1	16	4,7	9	0,01	
Gemini 3.8 Flash	pilotaż (P05–P07, P12, P13)	80	91 %	0,82	2	5	6,8	4	0,00	
GPT-6 Astra	rozszerzenie (P01, P04, P08–P10)	6	100 %	1,00	0	0	10,3	0	0,17	
GPT-6 Astra	pilotaż (P05–P07, P12, P13)	43	98 %	0,95	1	0	10,2	0	0,37	

Na pytaniach rozszerzenia najwyższą zgodność osiągnął Gemini 4 Argon (κ 0,90, n = 79; jedna karta bez oceny). GPT-6.1 Sol i Gemini 3.1 Pro mają tę samą κ 0,80, ale błędy w przeciwnych kierunkach: GPT-6.1 Sol zgłosił 7 halucynacji, których karty nie przyjął, i pominął jedną; Gemini 3.1 Pro nie zgłosił żadnego fałszywego alarmu i pominął 8 halucynacji. Gemini 3.8 Flash pominął 16 z 42 halucynacji zapisanych w kartach (κ 0,58). Sześć z siedmiu fałszywych alarmów GPT-6.1 Sol wynikało z definicji surowszej niż przyjęta w specyfikacji: pięć kart P08 (zawężenie definicji deepfake wbrew wytycznym Komisji, które specyfikacja od 19 października traktuje jako błąd wykładni w rdzeniu, nie jako T) i karta P09 Gemini 4 Argon W0 („musi” w zaleceniu redakcyjnym); w żadnym z tych przypadków sędzia nie podał błędnie treści prawa. Siódmy to karta P10 Claude Opus 5.5 W1, której etykietę zmieniono na „nie” po zawężeniu reguły B9 23 października. Na pytaniach pilotażowych κ jest wyższa niż na rozszerzeniu u Ge-

mini 4 Argon (0,97), GPT-6.1 Sol (0,84) i Gemini 3.8 Flash (0,82), a niższa u Gemini 3.1 Pro (0,71: 5 fałszywych alarmów i 6 pominieć); kierunek zawyżenia nie jest więc jednakowy dla wszystkich sędziów. Macierze pomyłek i przedziały 95 % dla każdej κ podaje aneks E; na rozszerzeniu tylko przedziały Gemini 4 Argon (0,80–0,97) i Gemini 3.8 Flash (0,40–0,70) się nie pokrywają.

Wartości te różnią się od podanych w zestawieniu z 19 października 2026 r. (Gemini 4 Argon 0,92, GPT-6.1 Sol 0,82, Gemini 3.1 Pro 0,78, Gemini 3.8 Flash 0,56), bo po tej dacie zmieniły się etykiety części kart, między innymi po zawężeniu reguły B9 23 października. Raport podaje wartości na kartach w stanie z dnia składania; generator κ jest częścią skryptu składającego raport (trafi do repozytorium, aneks B).

6.3. Wyszukiwania i deklaracje o weryfikacji

Wszyscy sędziowie mieli wyszukiwarkę i polecenie sprawdzenia każdego powołania. Na pytaniach rozszerzenia Gemini 4 Argon wykonał średnio 12,2 wyszukiwania na kartę, a GPT-6.1 Sol 11,2; obaj wyszukiwali przy każdej karcie. Gemini 3.1 Pro wykonał 4,3 wyszukiwania na kartę i przy 25 z 80 kart nie wyszukał niczego, Gemini 3.8 Flash 4,7 i przy 9 kartach żadnego. Najwięcej wyszukiwał sędzia najzgodniejszy z kartami. Gemini 3.1 Pro, który wyszukiwał najmniej, ma jednak tę samą k co GPT-6.1 Sol, bo jego błędy to wyłącznie pominięcia, a błędy GPT-6.1 Sol to głównie fałszywe alarmy. Wśród czterech sędziów mniejsza liczba pominięć współwystępowała z wyszukiwaniem przy każdej karcie: dwaj sędziowie wyszukujący przy każdej karcie pominieli łącznie 4 halucynacje zapisane w kartach, dwaj pozostali 24. Badanie nie pozwala oddzielić wpływu wyszukiwania od innych różnic między modelami.

Drugą miarą jest rzetelność deklaracji o weryfikacji: czy model wskazuje powołania, których nie sprawdził. Pole „niesprawdzone” wypełniał praktycznie tylko GPT-6.1 Sol (0,36 powołania na kartę na rozszerzeniu); u Gemini 4 Argon, Gemini 3.1 Pro i Gemini 3.8 Flash średnia wynosi 0,00–0,01. Sędzia, który po zerze wyszukań raportuje wszystkie powołania jako sprawdzone, oddaje werdykt nieodróżnialny od zweryfikowanego. Wcześniejszy przykład z pilotażu: przy uchwale III CZP 50/22 GPT-6.1 Sol i GPT-6 Astra wpisały „niesprawdzone”, bo nie dotarły do uzasadnienia, a oba Gemini — „w pełni zgodne”, bez śladu lektury uzasadnienia. Statystyki wyszukań ist-

nieją tylko dlatego, że wywołania narzędzi były rejestrowane: dla sędziów Gemini zapisywaliśmy zapytania i źródła, dla sędziów Codex tylko liczbę wyszukań (od 25 października 2026 r. także zapytania i źródła).

6.4. Czego sędziowie nie widzą

Halucynacje przeoczone przez co najmniej dwóch sędziów na pytaniach rozszerzenia należą prawie wyłącznie do trzech grup: brzmienie albo zakres przepisu inne niż w dniu stanu prawnego (błędna historia progu w P01, stan prawa „na 2024 r.” zamiast 2026 r., art. 44 pr.aut. w starym brzmieniu, zawężające „jedynie” w art. 50 ust. 2 AI Act, art. 127a k.p.a.), prawdziwa sygnatura z tezą z innej sprawy (klasa T) oraz ogólnik „SN przyjmuje...” bez sygnatury (reguła B9). Wykrycie większości z nich wymaga informacji spoza odpowiedzi, klucza i reguł; wyjątkiem jest ogólnik sprzeczny z wyrokiem zapisanym w kluczu, który da się wychwycić z samego klucza. Potrzebny jest tekst urzędowy na datę stanu prawnego, pełny tekst orzeczenia albo znajomość linii orzeczniczej. Wyszukiwarka zwracała je zawodnie: tekst jednolity bieżący stawiała przed wcześniejszymi, zamiast orzeczeń podawała omówienia, a nie sygnalizowała, że wyroku nie ma. Sędzia, który nie dotrze do źródła, korzysta z tej samej pamięci co model, który napisał odpowiedź; ocenia wtedy wiarygodność twierdzenia, nie jego prawdziwość.

6.5. Pominięcia według dostawcy ocenianego modelu

Tabela 9. Pominięte halucynacje według dostawcy ocenianego modelu (wszystkie 10 pytań; mianownik: karty z halucynacją ocenione przez sędziego)

SĘDZIA	ANTHROPIC: POMINIĘTE / HALUCYNACJE W KARTACH	GOOGLE: POMINIĘTE / HALUCYNACJE	OPENAI: POMINIĘTE / HALUCYNACJE
Gemini 4 Argon	1 / 44 (2 %)	2 / 24 (8 %)	1 / 12 (8 %)
GPT-6.1 Sol	0 / 49 (0 %)	1 / 29 (3 %)	0 / 13 (0 %)
Gemini 3.1 Pro	6 / 48 (12 %)	7 / 29 (24 %)	1 / 13 (8 %)

SĘDZIA	ANTHROPIC: POMINIĘTE / HALUCYNACJE W KARTACH	GOOGLE: POMINIĘTE / HALUCYNACJE	OPENAI: POMINIĘTE / HALUCYNACJE
Gemini 3.8 Flash	6 / 49 (12 %)	13 / 29 (45 %)	2 / 13 (15 %)

Gemini 3.8 Flash pominął 45 % halucynacji w odpowiedziach modeli Google, wobec 12 % w odpowiedziach modeli Anthropic i 15 % OpenAI; Gemini 3.1 Pro odpowiednio 24 %, 12 % i 8 %. U Gemini 4 Argon różnica jest mała w liczbach bezwzględnych (2 z 24 wobec 1 z 44); jedna z dwóch pominiętych halucynacji w odpowiedziach Google jest w odpowiedzi samego Gemini 4 Argon (W0 P08: „jedynie” w art. 50 ust. 2 AI Act), druga w odpowiedzi Gemini 3.8 Flash (W0 P04: art. 127a k.p.a.). Projekt meta-sędziego z 19 października wymienia jeszcze przeoczenie w karcie P09 Gemini 4 Argon W1, ale ocena Gemini 4 Argon dla tej karty nie ma pliku i nie weszła do tabeli (rozdział 8). Wynik ma znane zastrzeżenie: odpowiedzi Google mają w tym zbiorze dużo błędów klasy W, które wszyscy sędziowie wykrywają gorzej, więc część różnicy może wynikać z rodzaju błędu, nie z rodziny modelu. Liczba kart na rodzinę (12–49) nie pozwala tych przyczyn rozdzielić.

6.6. Panel sędziów

Na 79 kartach rozszerzenia ocenionych przez czterech sędziów Gemini 4 Argon sam ma κ 0,90. Reguła „halucynacja, gdy zgłosi ją Gemini 4 Argon albo Gemini 3.1 Pro” daje κ 0,92 (1 fałszywy alarm, 2 pominięcia), większość z trzech (Gemini 4 Argon, GPT-6.1 Sol, Gemini 3.1 Pro) 0,90, a reguła „którykolwiek z czterech” 0,82 (7 fałszywych alarmów, 0 pominięć). Reguły dobieraliśmy na tych samych danych, więc nawet przyrost o 0,02 jest dopasowaniem, nie wynikiem. Łączenie ocen zmniejszało liczbę pominięć kosztem fałszywych alarmów; każdą halucynację z kart wykrył co najmniej jeden sędzia. Błędy pominię-

te przez dwóch lub więcej sędziów (rozdział 6.4) dotyczą faktów, których ci sędziowie nie sprawdzili w źródle. To samo zjawisko opisuje Kohli [3]: panel dziewięciu sędziów z siedmiu rodzin modeli dał informację odpowiadającą około dwóm niezależnym głosom, a najlepszy pojedynczy sędzia dorównał panelowi. Nasze dane wskazują prawdopodobną przyczynę skorelowanych błędów: niesprawdzone fakty.

6.7. Co wynika z rozdziału 6

- Na kartach sprawdzonych przed sędziami najlepszy sędzia-model osiąga κ 0,90 w etykiecie halucynacji; najslabszy 0,58. Wybór sędziego zmienia wynik benchmarku bardziej niż wybór reguły panelu.
- Pominięcia sędziów skupiają się na błędach wymagających źródła na datę: dawnego brzmienia, tezy z innej sprawy, linii orzeczniczej; przy błędach widocznych z odpowiedzi, klucza i reguł pominięć było niewiele (analiza przypadków, rozdział 6.4).
- W zbadanych konfiguracjach zwiększenie liczby sędziów nie dawało wyższej κ niż najlepszy pojedynczy sędzia; różnice między sędziami szły w parze z tym, jak weryfikowali powołania. Karta sędziego powinna podawać liczbę wyszukikań, listę niesprawdzonych powołań oraz zapytania i zwrócone źródła z wersją i datą; bez takiego dziennika czytelnik nie odróżni werdyktu sprawdzonego od niesprawdzonego.
- Powtórzeń biegu tego samego sędziego nie badaliśmy, więc o stabilności werdyktów nic nie twierdzimy.

07 / WARSTWA WERYFIKACJI

Warstwa weryfikacji

Kod sprawdzający powołania w źródłach urzędowych, weryfikator poparcia i raport dla sędziego w trzech wariantach.

7.1. Dwa kroki przed sędzią

Ponieważ sędziom brakowało źródła, zbudowaliśmy sprawdzenie, które wykonuje się przed oceną i niezależnie od sędziego. Nazywamy je **warstwą weryfikacji** (w artykule angielskim *verification layer*). Ma dwa kroki.

- **Krok faktów** (sam kod, bez modelu). Kod wyodrębnia z odpowiedzi powołania. Dla przepisu pobiera z legal-cite brzmienie jednostki na datę stanu prawnego klucza oraz brzmienia wcześniejsze (co 6, 12, 24, 60, 120 i 180 miesięcy), aby wykryć klasę W. Dla orzeczenia szuka sygnatury w SAOS, a gdy wyroku tam nie ma — w bazie orzeczeń LexAlpha, wyłącznie do odczytu przez bramkę. Cytaty porównuje z tekstem dosłownie.
- **Weryfikator poparcia** (model z ograniczonym wejściem). Model dostaje tylko twierdzenie, które odpowiedź przypisuje jednemu źródłu, i tekst tego źródła. Rozstrzyga, czy źródło popiera twierdzenie. Etykietę „popiera” może dać tylko z dosłownym cytatem ze źródła, którego obecność sprawdza potem kod.

Wynikiem jest **raport weryfikacji**: dla każdego powołania klasa (✓, F, T, W, S) albo status weryfikacji („tekst niedostępny”, „brzmienie na datę nieustalone”, „sygnatura niejednoznaczna”, „akt nieustalony” i inne) wraz z brzmieniem urzędowym, jego wersją i datą. Status nie jest błędem modelu: nieudane sprawdzenie nie dowodzi, że model się pomylił, i obu wyników nie wolno oznaczać tą samą etykietą. Podział pracy odpowiada ramie „legal warrant” Taranu-

khina i Shwartz [2]: istnienie, jurysdykcję i aktualność źródła da się sprawdzić kodem, poparcie twierdzenia przez źródło wymaga czytania.

7.2. Wyniki na pytaniach strojenia i na pytaniach testowych

Na dwóch pytaniach, na których warstwę budowaliśmy (P09 i P10, 32 odpowiedzi), krok faktów wykrył wszystkie błędy F (3 z 3), W (5 z 5) i S (1 z 1) zapisane w kartach oraz 7 z 7 cytatów niedosłownych, przy jednym fałszywym alarmie; nie wykrył żadnego z 15 błędów T merytorycznych, bo ich wykrycie wymaga czytania. Weryfikator poparcia w drugiej wersji wykrył 9 z 13 przepisów z cudzą treścią, 3 z 3 orzeczeń nie na temat i 1 z 4 ogólników B9 (razem 13 z 20, 65 %), przy 4 fałszywych alarmach na 355 ocen (1 %). Te liczby mogą być zawyżone, bo mechanizm rozpoznawania powołań i instrukcję dla modelu dostosowywaliśmy na tych samych danych.

Właściwy test objął osiem pytań, na których warstwy wcześniej nie strojono (P01, P04, P05, P06, P07, P08, P12, P13; 128 odpowiedzi; dalej: pytania testowe), z zamrożonymi instrukcjami. Pierwsza kolumna tabeli 10 to ten test; dwie następne pokazują wyniki po poprawkach wprowadzonych na podstawie analizy błędów z tych samych pytań, więc są wynikami eksploracyjnymi, nie testem na nietkniętym zbiorze. Punktem odniesienia jest lista 147 błędów z powołaniem zapisanych w kartach i 665 powołań, które karty uznały za poprawne. Tabela 10 pokazuje trzy kolejne biegi.

Tabela 10. Warstwa weryfikacji na pytaniach testowych (czułość wobec listy błędów z kart; „dowolna klasa” — błąd wykryty, choćby z inną klasą; „ta sama klasa” — z klasą z karty; źródła: raporty biegów warstwy z 23.10 i 24.10 w repozytorium)

MIARA	23.10 (PROMPTY ZAMROŻONE)	24.10 (NAPRAWY WEJŚCIA, PLIK „B”)	24.10 (DECYZJE NR 2–4, PLIK „C”)
czułość, dowolna klasa	84 / 147 (57 %)	87 / 147 (59 %)	94 / 147 (64 %)
czułość, ta sama klasa	n/d	66 / 147 (45 %)	74 / 147 (50 %)
T, dowolna klasa	39 / 83 (47 %)	49 / 83 (59 %)	50 / 83 (60 %)
W: ta sama / dowolna klasa	n/d / 15 z 44 (sam krok faktów)	17 / 28 z 44 (39 / 64 %)	24 / 32 z 44 (55 / 73 %)
S: ta sama / dowolna klasa	n/d / 7 z 14 (sam krok faktów)	0 / 4 z 14	2 / 6 z 14
F: ta sama / dowolna klasa	n/d / 3 z 6 (sam krok faktów)	0 / 6 z 6	0 / 6 z 6
fałszywe alarmy na 665 ✓	8,0 % weryfikator + 4,1 % krok faktów	1,8 %	1,8 %
koszt biegu (Vertex)	8,09 USD	7,87 USD	7,09 USD

Uwaga do kolumny 23.10: łączna czułość 84 ze 147 to suma wykryć kroku faktów i weryfikatora poparcia liczona na całym zbiorze; raport z tego biegu nie podaje tej sumy w podziale na klasy W, S i F, dlatego dla tych klas podajemy wyłącznie wykrycia kroku faktów (łącznie z T: 64), które nie sumują się do 84.

7.3. Większość strat to wejście, nie ocena

Różnica między 65 % na pytaniach strojenia a 57 % na pytaniach testowych łączy dwie różne miary (czułość weryfikatora poparcia na 20 błędach treściowych wobec czułości całej warstwy na 147 błędach), więc nie jest ścisłym spadkiem; wzrost fałszywych alarmów z 1 % do 8 % wynikał głównie z przetwarzania wejścia, nie z oceny modelu. Cztery przyczyny dało się usunąć bez zmiany promptów:

- parser powołań nie rozpoznawał litery w indeksie górnym (art. 131^{1a}, 394^{1a} k.p.c.) i szukał sąsiedniego artykułu — 22 fałszywe alarmy przed naprawą, 1 po niej;

- akt domyślny z klucza podstawiano za inny akt nazwany w odpowiedzi (k.p.a., AI Act, k.w. zamiast ustawy o nabywaniu nieruchomości przez cudzoziemców, RODO, prawa autorskiego) — 37 alarmów przed naprawą, po niej 0 i 38 statusów „akt nieustalony”;
- ta sama sygnatura w różnych sądach: kod brał cudzy wyrok (np. VIII Gz 43/16) — 14 alarmów przed naprawą, 6 po niej;
- łańcuch „art. X w zw. z art. Y” liczony jako kilka twierdzeń — 21 alarmów przed naprawą, po niej 0 i 120 statusów „w związku”.

Po naprawach fałszywe alarmy spadły do 1,8 %, a czułość wykrywania błędów klasy T wzrosła z 47 % do 59 %. Część pozostałych pominieć wynikała z różnic w stosowaniu definicji, nie z kodu. Najważniejsza dotyczyła zawężenia kręgu adresatów albo zakresu normy: weryfikator czytał je jako pominiecie, karta jako W albo T. Decyzja z 24 października 2026 r. (rozdział 4.5) i jej wprowadzenie do promptu weryfikatora podniosły czułość „ta sama klasa” na klasie W z 39 % do 55 %, kosztem jednego nowego fałszywego alarmu i jednej zmiany klasy z W na T.

Dwie obserwacje dotyczą samych źródeł. SAOS był niedostępny przez cały bieg 23 października (błędy odpowiedzi i przekroczenia czasu), a kod nie odróżniał awarii od braku wyroku; po poprawce awaria jest osobnym statusem. W bazie orzeczeń LexAlpha tekst wyroku SN III CK 124/05 ma 10 238 znaków wobec 18 748 w SAOS i nie zawiera zdania o relacji art. 41 ust. 2 pr.aut. do art. 65 k.c., dla którego wyrok się cytuje; jest to inna, krótsza wersja redakcyjna. SAOS nie ma z kolei 5 z 8 badanych wyroków SN (III CK 400/03, II CSKP 526/22, IV CSKP 44/21, I CSK 596/17, I CSK 104/18). Ta sama sygnatura może więc mieć w dwóch bazach różne teksty, a weryfikator musi wiedzieć, którą wersję czyta. Od 21 października 2026 r. pierwszeństwo ma tekst z SAOS.

7.4. Granica źródła dla klasy W

Klasę W da się zmierzyć tylko tam, gdzie źródło urzędowe pozwala odtworzyć brzmienie na wskazaną datę. Dla Kodeksu postępowania cywilnego, który między tekstami jednolitymi ma kilkadziesiąt nowelizacji, legal-cite zwrócił „brzmienie na datę nieustalone” dla 931 z 941 powołań; podobnie dla u.k.s.c., k.p.k., p.p.s.a. i k.k.s. Klasa F jako ta sama klasa ma 0 % czułości, bo legal-cite odnajduje jednostkę pod sąsiednim numerem i błąd zostaje zaklasyfikowany

jako T. Klasa S na orzeczeniach zamienia się w status, gdy sygnatura jest niejednoznaczna albo wyroku nie znaleziono; od 24 października weryfikator ocenia teżę wobec każdego z co najwyżej czterech kandydatów, co dało 2 trafne S z 14. Kryterium, którego nie da się zastosować, bo źródła nie można pobrać w wymaganej wersji, nie mierzy niczego. Dlatego obok wyników benchmarku należy podawać zasięg weryfikacji: ile powołań sprawdzono i w jakiej wersji źródła.

7.5. Raport dla sędziego: wiążący, pomocniczy, v3

Raport weryfikacji podaliśmy dwóm sędziom, Gemini 4 Argon i GPT-6.1 Sol, w trzech wariantach instrukcji (karta meta). W wariacie pierwszym raport był **wiązący**: klasy A6 sędziego miał opierać na raporcie. W drugim był **pomocniczy** (decyzja z 24 października 2026 r.: „raport może dodać błąd, nie może go odjąć”): klasa F, T, W albo S z raportu wiąże, a „✓” nie zwalnia sędziego z porównania treści przypisanej źródłu z brzmieniem z raportu. W trzecim (v3, 25 października 2026 r.) raport pokazuje „✓” tylko wtedy, gdy brzmienie na datę stanu prawnego zostało potwierdzone; w pozostałych przypadkach podaje status, który odsyła sędziego do własnej oceny.

Tabela 11. Sędzia bez raportu i z raportem weryfikacji (κ etykiety hallucinated wobec kart na kartach wspólnych z sędzią bez raportu; wiersze różnią się zbiorem kart, porównanie wariantów na jednym zbiorze w tekście niżej)

SĘDZIA	WARIANT	PYTANIA	N (KARTY WSPÓLNE Z SĘDZIĄ BEZ RAPORTU)	K BEZ RAPORTU	K Z RAPORTEM	FAŁSZYWE ALARMY BEZ → Z	POMINIĘCIA BEZ → Z
Gemini 4 Argon	raport wiążący (v1)	P05, P06, P07, P13	54	0,96	0,85	0 → 0	1 → 4
Gemini 4 Argon	raport pomocniczy (v2)	P01, P04, P05, P06, P07, P08, P12, P13	115	0,95	0,88	0 → 1	3 → 6
Gemini 4 Argon	v3: ✓ tylko przy brzmieniu potwierdzonym	P01, P04, P05, P06, P07, P08, P12, P13	116	0,95	0,88	0 → 2	3 → 5
GPT-6.1 Sol	raport wiążący (v1)	P05, P06, P07, P13	61	0,88	0,83	3 → 0	0 → 5

SĘDZIA	WARIANT	PYTANIA	N (KARTY WSPÓLNE Z SĘDZIĄ BEZ RAPORTU)	K BEZ RAPORTU	K Z RAPORTEM	FALSZYWE ALARMY BEZ → Z	POMINIĘCIA BEZ → Z
GPT-6.1 Sol	raport pomocniczy (v2)	P01, P04, P05, P06, P07, P08, P12, P13	119	0,84	0,97	8 → 1	1 → 1
GPT-6.1 Sol	v3: ✓ tylko przy brzmieniu potwierdzonym	P01, P04, P05, P06, P07, P08, P12, P13	128	0,80	0,89	11 → 4	1 → 3

Raport wiążący obniżył zgodność obu sędziów (Gemini 4 Argon z 0,96 do 0,85, GPT-6.1 Sol z 0,88 do 0,83). Mechanizm widać na kartach: „✓” w raporcie oznaczał tylko, że przepis istnieje, ma podane brzmienie, a cytat jest dosłowny, a sędziowie traktowali go jak potwierdzenie całego powołania i przestawali sprawdzać, czy odpowiedź opisuje na przykład krąg adresatów przepisu sprzed nowelizacji. Raport zmienił etykietę zawsze z „tak” na „nie”; błędnie w kartach P05 GPT-5.6 Terra W0 (art. 131^{1a} § 1 k.p.c.: „tylko profesjonalni pełnomocnicy”), P05 Claude Sonnet 5.5 W1 (reguła doręczenia z uchwały III CZP 24/23 podana jako aktualna), P07 Gemini 3.8 Flash W1 (art. 25b ust. 1 u.k.s.c.) i P13 GPT-5.6 Terra W0 (art. 492 § 3 k.p.c.).

Raport pomocniczy na ośmiu pytaniach testowych podniósł zgodność GPT-6.1 Sol z 0,84 do 0,97 (n = 119): zniknęło 7 z 8 fałszywych alarmów, pominięć nie przybyło. Zgodność Gemini 4 Argon spadła z 0,95 do 0,88 (n = 115). Trzy dodatkowe pominięcia Argona to błędy klasy W, które raport pokazał jako „✓”: w jednym brzmienia na datę nie dało się ustalić (P07 Gemini 3.8 Flash W1), w jednym raport powstał przed wprowadzeniem reguły o kręgu adresatów (P05 GPT-5.6 Terra W0), w jednym odpowiedź przyjęła błędny rok dla progu ustawowego, czego krok faktów nie sprawdza (P01 Gemini 4 Argon W0). Karty v2 dla P05, P06, P07 i P13 pochodzą z raportu „b”, a dla P01, P04, P08 i P12 z raportu „c”; czyste porównanie na jednym raporcie wymaga powtórzenia 64 kart Argona.

Wariant v3 ukończono 25 października 2026 r. na 128 kartach dla GPT-6.1 Sol i 127 dla Gemini 4 Argon (jedna karta bez odpowiedzi Vertex AI po dwóch próbach). Wariant miał pokazać, jak sędzia zachowa się, gdy zamiast „✓” zobaczy status; nie zakładaliśmy, że podniesie zgodność. Na 119 kartach wspólnych dla ocen bez raportu oraz wariantów v2 i v3 GPT-6.1 Sol ma k 0,84 bez raportu, 0,97 z raportem v2 i 0,91 z raportem v3 (fałszywe alarmy 8 → 1 → 2, pominięcia 1 → 1 → 3); Gemini 4 Argon na 115 kartach wspólnych 0,95 bez raportu i 0,88 zarówno z v2, jak i z v3. Dwa nowe pominięcia GPT-6.1 Sol w v3 to karty klasy W (P05 Claude Sonnet 5.5 W1, P07 Gemini 3.8 Flash W1), które w v2 ten sam sędzia wykrył. W v3 raport pokazywał przy tych powołaniach status „brzmienie na datę nieustalone” zamiast „✓”; sędzia ocenił brzmienie sam i przyjął je za obowiązujące, a w P07 potraktował opłatę 100 zł z dawnego brzmienia jako pominięcie, nie jako W. W tych dwóch kartach sędzia w v3 pominął błąd wykryty w v2; jedną z możliwych przyczyn jest zmiana informacji w raporcie (status zamiast „✓”), po której sędzia oparł się na własnym odczycie odpowiedzi. Pełny bilans zmian na 119 kartach to jeden fałszywy alarm i dwa pominięcia więcej w v3 niż w v2. Najwyższą zgodność z kartami daje nadal v2 u sędziego surowszego. Wyjaśnienie mechanizmu traktujemy jako hipotezę roboczą: różnica między wariantami opiera się na kilku kartach jednego sędziego, bez powtórzeń biegu, i wymaga sprawdzenia na większej próbie oraz na większej liczbie sędziów. Przedziały 95 % z bootstrapu po pytaniach (aneks E) się pokrywają: u

GPT-6.1 Sol 0,92–1,00 z raportem v2 i 0,85–0,98 z v3, a bez raportu 0,70–0,96; w badanej próbie k była wyższa w v2 niż w v3, ale niepewności tej różnicy nie oszacowano (wymaga porównania sparowanego).

Wynik zgadza się z badaniem agentów, które przejmują podrobione wyniki narzędzi, dla wyszukiwania w 68 % przypadków [5]. Wymaga też zawężenia: gdy podana referencja przeczy wiedzy sędziego, sędziowie zwykle ją ignorują [4]. Efekt, który obserwujemy, dotyczy raportu zgodnego z tym, co sędzia i tak by przyjął: „✓” potwierdza powołanie, którego sędzia nie zamierzał badać. Reguła, którą z tego wyprowadzamy, brzmi: automatyczna weryfikacja może dodać błąd do oceny sędziego, nie może go usunąć, a powołanie może być pokazane jako poprawne tylko po zakończeniu każdego elementu sprawdzenia.

7.6. Koszt warstwy weryfikacji

Krok faktów nie ma kosztu poza zapytaniami do źródeł publicznych. Weryfikator poparcia kosztował 3,05 USD za 383 ocenione powołania na pytaniach strojenia (około 0,8 centa za powołanie) i 8,09 USD za 1 452 powołania na pytaniach testowych w biegu z 23 października (około 0,6 centa). Karta meta Gemini 4 Argon kosztuje około 0,10 USD (6,57 USD za 64 kar-

ty w wariancie v1), dwa do trzech razy więcej niż karta sędziego bez raportu, bo polecenie rośnie do około 19 tys. tokenów, a Gemini 4 Argon nie obsługuje jawnie zarządzanej pamięci podręcznej kontekstu na Vertex AI. Pełny cykl na ośmiu pytaniach testowych (krok faktów, weryfikator poparcia i karta meta Gemini 4 Argon) kosztował 14,70 USD. GPT-6.1 Sol działał z subskrypcji, a jego wąskim gardłem był limit konta, nie koszt.

7.7. Pełny koszt badania

Tabela 12 zbiera wycenę użycia modeli w tym badaniu według cenników z 18 października 2026 r. (dla Codex i Claude Code równoważnik API, nie wydatek): odpowiedzi ocenianych modeli, sędziów-modeli na kartach w specyfikacji v1.0.1 oraz warstwę weryfikacji ze wszystkimi biegami testowymi i strojeniem. Koszt sędziów i kart meta liczymy z tokenów zapisanych w plikach ocen i z cennika z 18 października 2026 r. Wyszukiwania Gemini (Grounding with Google Search) podajemy osobno w cenie katalogowej 0,014 USD za zapytanie. Cennik Vertex AI zwalnia z opłaty 5000 takich zapytań miesięcznie, dlatego zapisy kosztu w raportach biegów ich nie obejmują, a faktura była niższa niż suma w tabeli.

Tabela 12. Pełny koszt badania (USD, cennik z 18.10.2026; „≥” = dolna granica; „n/d” = Codex z subskrypcji bez zapisu tokenów)

CZĘŚĆ	POZYCJA	KARTY	TOKENY USD	WYSZUKANIA	WYSZUKANIA USD	RAZEM USD
Odpowiedzi	160 odpowiedzi ośmiu modeli (W0 i W1)	160	≥ 15,78	—	0,00	≥ 15,78
Sędziowie-modele	Gemini 4 Argon, specyfikacja v1.0.1	147	≥ 6,18	2072	29,01	≥ 35,19
Sędziowie-modele	Gemini 3.1 Pro, specyfikacja v1.0.1	159	≥ 5,73	945	13,23	≥ 18,96
Sędziowie-modele	Gemini 3.8 Flash, specyfikacja v1.0.1	160	≥ 1,91	913	12,78	≥ 14,69
Sędziowie-modele	GPT-6.1 Sol, specyfikacja v1.0.1	160	n/d	n/d	n/d	n/d (Codex z subskrypcji; tokeny niezapisane)

CZĘŚĆ	POZYCJA	KARTY	TOKENY USD	WYSZUKANIA	WYSZUKANIA USD	RAZEM USD
Sędziowie-modele	GPT-6 Astra, specyfikacja v1.0.1	49	n/d	n/d	n/d	n/d (Codex z subskrypcji; tokeny niezapisane)
Warstwa weryfikacji	krok faktów (kod, źródła publiczne)	—	0,00	—	0,00	0,00
Warstwa weryfikacji	weryfikator poparcia (6 biegów: testy, strojenie P09–P10, test na 8 pytaniach, poprawki)	—	26,38	—	0,00	26,38
Warstwa weryfikacji	karta meta Gemini 4 Argon v1	64	6,57	193	2,70	9,28
Warstwa weryfikacji	karta meta Gemini 4 Argon v2	125	15,97	571	7,99	23,96
Warstwa weryfikacji	karta meta Gemini 4 Argon v3	127	17,07	652	9,13	26,20
Warstwa weryfikacji	karta meta GPT-6.1 Sol v1	61	n/d	n/d	n/d	n/d (Codex z subskrypcji)
Warstwa weryfikacji	karta meta GPT-6.1 Sol v2	119	n/d	n/d	n/d	n/d (Codex z subskrypcji)
Warstwa weryfikacji	karta meta GPT-6.1 Sol v3	128	n/d	n/d	n/d	n/d (Codex z subskrypcji)
Warstwa weryfikacji	testy wstępne poprawności działania	—	0,26	—	0,00	0,26
Razem			≥ 95,85		74,84	≥ 170,70

Pełny koszt wyniósł co najmniej 170,70 USD, z czego 74,84 USD przypada na wyszukiwania w cenie katalogowej. Odpowiedzi ocenianych modeli to co najmniej 15,78 USD, sędziowie-modele Vertex co najmniej 68,85 USD, a warstwa weryfikacji 86,07 USD. Ocena kosztowała więc kilka razy więcej niż wygenerowanie ocenianych odpowiedzi. U sędziów Gemini większą część kosztu stanowią wyszukiwania, a nie tokeny. Kwota jest dolną granicą z trzech powodów: sędziowie w specyfikacji v1.0.1 nie zapisywali tokenów rozumowania, W1 modeli Claude i Gemini nie ma pełnego zapisu narzędzi, a biegi Codex z subskrypcji nie zapisywały tokenów. Tabela nie obejmuje pracy prawnika ani modelu Claude, który wstępnie wypełniał karty.

7.8. Co wynika z rozdziału 7

- Kod umożliwia automatyczne sprawdzenie istnienia jednostki i dosłowności cytatu, ale wynik zależy od poprawnego rozpoznania powołania i dostępu do właściwego tekstu. Wykrywanie dawnego brzmienia, nieistniejących jednostek i błędnych oznaczeń orzeczeń ogranicza jakość źródeł urzędowych dostępnych w wymaganej wersji i jakość parsowania powołań.
- Cztery usterki przetwarzania wejścia (indeksy górne z literą, akty domyślne, te same sygnatury w różnych sądach, łańcuchy „w zw.”) odpowiadały za 94 fałszywe alarmy w pierwszym biegu i 7 po poprawkach; osobnym ograniczeniem są krótsze wersje tekstów orzeczeń. Te elementy techniczne wymagają tyle samo pracy projektowej co kryteria oceny.

- Zgodność sędziego z kartami różniła się między wariantami instrukcji i raportu; badanie nie oddziaływało wpływu samej instrukcji od treści raportu. Raport wiążący obniżył zgodność obu sędziów; raport pomocniczy podniósł zgodność sędziego surowszego (GPT-6.1 Sol), a u Gemini 4 Argon zwiększył liczbę pominieć tam, gdzie „✓” oznaczał sprawdzenie niepełne. Sędzia musi mieć polecenie samodzielnego sprawdzenia każdego powołania, także oznaczonego w raporcie jako poprawne.

08 / OGRANICZENIA

Ograniczenia

Granice interpretacji wyników i niespójności danych wykryte przy składaniu raportu.

8.1. Zakres interpretacji wyników

- 1. Jeden prawnik.** Wszystkie karty sprawdził jeden prawnik, Adam Piotrowski. Karty wypełniał wstępnie model Claude, a więc model z rodziny trzech ocenianych systemów. Nie mamy drugiego oceniającego, więc nie znamy zgodności człowiek–człowiek, do której należałoby odnieść k sędziów-modeli. κ w rozdziałach 6 i 7 to zgodność modelu z kartą zamkniętą przez człowieka.
- 2. Kolejność kart i sędziów w pilotażu.** Karty pytań pilotażowych sprawdzano po obejrzeniu ocen sędziów, a sporne punkty rozstrzygano często na korzyść argumentu sędziego. Zgodność na tej części, w tym κ „bez raportu” w porównaniu z kartą meta (rozdział 7.5), jest zawyżona. Kierunek zmian między wariantami raportu jest mimo to jednoznaczny.
- 3. Zbiór kalibracyjny.** Reguły oceny zmienialiśmy po lekturze odpowiedzi (tabela 3), a klucze P09 i P10 zmieniono po zebraniu odpowiedzi (rozdział 2.3). Wyniki nie są pomiarem według zamrożonej specyfikacji.
- 4. Mała próba.** Dziesięć pytań, w tym pięć z postępowania cywilnego, jedno uruchomienie każdego systemu; dla wyników ocenianych modeli z rozdziału 5 nie oszacowano przedziałów niepewności ani nie przeprowadzono formalnych porównań sparowanych (przedziały dla κ sędziów są w aneksie E). Różnice o jedną–dwie odpowiedzi między modelami mieszczą się w zmienności, której nie mierzyliśmy.
- 5. Warunki biegu.** W W1 modele Claude czytały strony przez streszczenie wykonane przez Claude Haiku 5.5, a Claude Code nie zachowuje śladu narzędzi, więc 5 z 40 halucynacji w W1 ma pochodzenie nieustalone z tego powodu (3 dalsze dotyczą modeli Gemini, u których ślad nie wskazywał źródła błędu). Claude Code CLI dokleja do polecenia sekcję środowiska z danymi konta, której nie da się wyłączyć. Koszt W1 dla Claude i Gemini jest dolną granicą.
- 6. Sędziowie.** Sędziowie na pytaniach rozszerzenia startowali przed dopisaniem do specyfikacji reguł B9–B12. GPT-6 Astra ocenił tylko część kart. Dla sędziów Codex do 25 października logowaliśmy tylko liczbę wyszukikań, nie zapytania i źródła. Powtórzeń biegu sędziego nie badaliśmy.
- 7. Warstwa weryfikacji.** Parser i prompty strojono na P09 i P10; test na ośmiu pytaniach jest pierwszym pomiarem poza strojeniem. Karty meta v2 Gemini 4 Argon pochodzą z dwóch wersji raportu („b” i „c”). Wariant v3 ma komplet kart GPT-6.1 Sol i 127 z 128 kart Gemini 4 Argon.
- 8. Literatura.** Większość prac o sędziach-modelach, na które się powołujemy [3]–[6], to preprinty z 2026 r. bez recenzji; w przeglądzie literatury czytaliśmy ich streszczenia, nie pełne teksty.

8.2. Niespójności w danych wykryte przy składaniu raportu

Raport liczy wszystkie wartości z kart w stanie z 25 października 2026 r. Przy składaniu wykryliśmy rozbieżności między kartami a wcześniejszymi zesta-

wieniami i zapisami; liczby w raporcie ich nie powielają, ale wymagają one poprawienia u źródła.

- Zestawienie W0/W1 z 19.10 (aktualizacja z 23.10) podaje dla Claude Opus 5.5 W1 4 odpowiedzi poprawnie ugruntowane, ale koszt na odpowiedź $\geq 0,489$ USD, co odpowiada trzem (1,467 / 3). Na obecnych kartach: 4 odpowiedzi i $\geq 0,367$ USD.
- Wcześniejsza notatka robocza (A18) podaje dla Claude Opus 5.5 halucynacje $9 \rightarrow 6$ i odpowiedzi poprawnie ugruntowane $0 \rightarrow 3$; po decyzji z 23.10 (B9) karty dają $9 \rightarrow 5$ i $0 \rightarrow 4$. Wpis A14 („wyszukiwanie dokłada najwyżej 1–2 poprawnie ugruntowane odpowiedzi”) jest nieaktualny: u Claude Opus 5.5 przyrost wynosi 4.
- Zestawienie z 19.10 podaje κ sędziów na rozszerzeniu $0,92 / 0,82 / 0,78 / 0,56$; na obecnych kartach $0,90 / 0,80 / 0,80 / 0,58$ (rozdział 6.2). Artykuł v1.2 (opublikowany 26.10) podaje obie serie: na kartach w stanie pierwotnym i na obecnych. Zdanie z notatki roboczej (B11) i z artykułu, że kolejność sędziów według wyszukiń jest taka sama jak według κ , nie jest ścisłe: Gemini 3.1 Pro wyszukiwał mniej niż Gemini 3.8 Flash (4,3 wobec 4,7), a ma wyższą κ .
- Karta P12_gemini38flash_vertex_w1_r3 miała etykietę „bez weryfikacji źródła: tak (brak narzędzi)”, choć bieg W1 miał narzędzia, a model ich nie użył. Decyzją z 25.10 etykietę poprawiono na „narzędzia nieużyte” według reguły z 18.10. Dziennik z 18.10 podaje 9 kart „narzędzia nieużyte”, po tej poprawce jest ich 12.
- Pole A6_zakres było tylko w kartach pytań rozszerzenia. W 49 kartach z halucynacją na pytaniach pilotażowych go brakowało, w tym w kartach wskazanych w specyfikacji jako wzory zakresu (P05_sonnet_cli_w1_r3, P12_argon4_high_vertex_w1_r3, P07_sonnet_cli_w1_r3). Uzupełniono je 25.10 (35 rdzeń, 14 poboczne); 7 przypadków granicznych rozstrzygnęły dwie reguły z tego dnia, 4 pozostają do przeglądu prawnika.
- Karta P07_opus_cli_w0_r3 zawiera nadal klasę O (zniesioną 11.10) w adnotacji historycznej; punkt i etykieta wynikają z klasy T.
- Klucze P05, P06 i P07 nie mają pola status, wymaganego zasadą 7 specyfikacji.
- P01 ma w kluczu typ obliczenie, jak wyłączone P02, którego powodem wyłączenia był charakter obliczeniowy. Decyzją z 25.10 uznaliśmy to za spójne: odpowiedź P01 to kwota zapisana wprost w ustawie (800 zł w art. 119 § 1 k.w.), więc pytanie sprawdza treść prawa, a nie rachunek.
- Brak pliku oceny Gemini 4 Argon dla karty P09_argon4_high_vertex_w1_r3, choć projekt meta-sędziego i notatka robocza (B8) liczą tę kartę jako przeoczenie Argona.
- Skrypty porównujące sędziów z kartami czytały etykietę karty z pierwszego wystąpienia wzorca i nie obsługiwały zapisu „nie $\rightarrow$ tak”; poprawione 25.10 (odczyt z ostatniej linii „**Δ”). Na 327 kartach w katalogu ocen (160 raportu i karty systemów spoza niego) stary i nowy odczyt dają ten sam wynik (0 rozbieżności), ale przy kolejnych zmianach kart może dać inny.
- Karta P06 Claude Opus 5.5 W1 nie ma oceny Gemini 4 Argon w wariancie v3: Vertex AI dwa razy nie odpowiedział w 20 minut; karta nie została zapisana i nie weszła do tabeli 11.
- Karty P05 GPT-6.1 Sol W0 i P05 GPT-6 Astra W0 miały na końcu wiersza A6 nieaktualny dopisek „A6_typ: [W]” po decyzji z 13.10 (A6 = 1, etykieta „nie”); usunięty 26.10. Na liczby raportu nie wpływał, bo klasy liczymy tylko w odpowiedziach z halucynacją.

09 / WNIOSKI I DALSZE KROKI

Wnioski i dalsze kroki

Pięć wniosków i kroki przed zbiorem testowym.

9.1. Wnioski

Pierwszy wniosek dotyczy miary. Na dziesięciu krótkich pytaniach z polskiego prawa 29 % odpowiedzi ośmiu modeli było poprawnie ugruntowanych, a 57 % zawierało co najmniej jedno twierdzenie o źródle prawa niezgodne z tekstem urzędowym. Z 93 odpowiedzi z pełnym rdzeniem połowa miała wadliwe powołanie. Ocena samego rdzenia uznałaby te odpowiedzi za prawidłowe; prawnik, który przeniósłby powołanie do pisma bez sprawdzenia, popełniłby błąd.

Drugi wniosek dotyczy rodzaju błędu. W badanym zbiorze najczęstszym błędem nie było zmyślane źródło, lecz przypisanie istniejącemu przepisowi lub orzeczeniu treści, której ono nie zawiera (klasa T, 66 odpowiedzi), i brzmienie z innej daty (klasa W, 35). Wyszukiwanie ogranicza pierwszy, prawie nie zmienia drugiego. Benchmark, który nie sprawdza brzmienia na datę stanu prawnego, pomija błąd najtrwalszy.

Trzeci wniosek dotyczy warunku W1. Dostęp do wyszukiwarki podnosi wynik bloku A u wszystkich modeli, ale liczba odpowiedzi poprawnie ugruntowanych rośnie o 4 u jednego modelu, o 1–2 u trzech, nie zmienia się u dwóch i spada o 1 u jednego; ilorazy wykazanych kosztów W1 i W0 wynoszą około 1,4–7,7, dla części modeli jako dolne oszacowania. Wynik W1 trzeba raportować razem z tym, czy model użył narzędzi i do jakiego źródła dotarł.

Czwarty wniosek dotyczy oceny. Sędzia-model z wyszukiwarką odtwarza kartę prawnika z k do 0,90 na kartach sprawdzonych przed sędziami, ale pomija między innymi błędy aktualności brzmienia przepisu i treści orzeczenia: dawne brzmienie, tezę z innej sprawy, nieistniejącą linię orzeczniczą. Reguła łączenia sędziów w panel podniosła k najwyżej o 0,02; w poszczególnych porównaniach raport weryfikacji

zmieniał ją od –0,11 do +0,13, w zależności od sędziego i wariantu; porównania obejmowały różne podzbiory kart.

Piąty wniosek dotyczy warstwy weryfikacji. Kod sprawdzający powołania w źródłach urzędowych jest potrzebny, ale jego zasięg wyznaczają źródła: dla k.p.c. brzmienia na datę nie dało się ustalić w 931 z 941 powołań. Niepełne sprawdzenie nie powinno być pokazywane jako potwierdzenie poprawności; w badaniu wpływ raportu zależał od modelu oceniającego. Podobny wniosek dla całych systemów prawnych z wyszukiwaniem formułują Butler i Butler: jakość wyszukiwania wyznacza górną granicę jakości odpowiedzi [6].

9.2. Dalsze kroki

- Nowy zbiór pytań testowych z kluczami zatwierdzonymi przed biegiem, ocena według zamrożonej wersji specyfikacji, bez zmian reguł w trakcie.
- Co najmniej trzy biegi na system, przedziały ufności i porównania parami na tych samych pytaniach.
- Drugi prawnik na próbce około 30 kart jako punkt odniesienia dla k sędziów-modeli.
- Miara pomocnicza obok etykiety: udział poprawnych powołań w odpowiedzi; tabela krzyżowa dwóch osi (halucynacja × ugruntowanie) dla każdego systemu.
- Przegląd pola A6_zakres uzupełnionego 25.10 w kartach pilotażu (zwłaszcza 4 rozstrzygnięć niepewnych w P06 i P07) i pozostałe poprawki z rozdziału 8.2.

- Powtórzenie kart v2 Gemini 4 Argon na jednym raporcie (dziś z dwóch wersji „b” i „c”) i próba wariantu, w którym status „brzmienie nieustalone” wymusza na sędzim cytaty ze źródła przed uznaniem powołania.
- Logowanie każdego wywołania narzędzia sędziego (zapytanie, zwrócone źródło, wersja i data źródła) — od 25 października 2026 r. także dla sędziów Codex.
- Bieg kontrolny modeli Claude przez API z narzędziem pobierającym pełną treść strony, aby oddzielić błąd modelu od błędu streszczenia.
- Zapis liczby wyszukikań w biegach Claude Code i tokenów treści URL context w biegach Vertex, aby koszt W1 nie był dolną granicą.
- Rozstrzygnięcie, czy B1 ma obejmować orzeczenie z tezą, której w nim nie ma, przytoczone jako argument do pisma (plan wersji 1.1, pkt 10).

ANEKS A

Pytania i klucze w skrócie

Pytania w brzmieniu z biegu i rdzenie kluczy w skrócie.

Tabela zawiera pytanie w brzmieniu przekazanym systemom (z zachowaniem literówek) i pierwsze dwa zdania rdzenia klucza; dla P09 i P10, których rdzeń zmieniono po biegu na warunkowy (rozdział 2.3), podajemy streszczenie z warunkiem. Pełne klucze, z rubryką, karami, wzorcami cytatów i źródłami, trafią do repozytorium LexAlpha Lab na GitHub (aneks B, „Materiał źródłowy”).

Tabela A.1. Klucze dziesięciu pytań (wersja i data klucza; data stanu prawnego we wszystkich: 1.10.2026)

PYTANIE	WERSJA KLUCZA (DATA)	PYTANIE W BRZMIENIU Z BIEGU	RDZEŃ KLUCZA (SKRÓT)	PODSTAWY WYMAGANE
P01	v3 (2026-10-18)	Ile wynosi w tym roku wartość mienia przy wykroczeniu max ? Czy zmieniła się w zw. ze wzrostem płacy minimalnej ?	800 zł. Od 1.10.2023 próg jest kwotą stałą wpisaną w przepis (art. 119 § 1 k.w. przy kradzieży i przywłaszczeniu; tak samo art. 120 § 1, 122 § 1 i 124 § 1 k.w.), więc wzrost płacy minimalnej go nie zmienia.	art. 119 § 1 k.w.
P04	v3 (2026-10-18)	Postanowienie (zgoda) na nabycie nieruchomości przez obywatela Ukrainy. Takiemu orzeczeniu nadaje się klauzulę prawomocności czy ostateczności?	Nie postanowienie, lecz decyzja ministra właściwego do spraw wewnętrznych (art. 1 ust. 1 ustawy o nabywaniu nieruchomości przez cudzoziemców), i nie klauzula, lecz zaświadczenie organu (art. 217 § 2 pkt 2 w zw. z art. 218 § 1 k.p.a.). Decyzja jest ostateczna, gdy nie służy już wniosek o ponowne rozpatrzenie sprawy (art. 16 § 1, art. 127 § 3 k.p.a.; 14 dni), a prawomocna, gdy nie można jej zaskarżyć do sądu administracyjnego (art. 16 § 3 k.p.a.) — po upływie 30 dni od doręczenia bez skargi do WSA (art. 53 § 1 p.p.s.a.; skargę można wnieść także bez wniosku o ponowne rozpatrzenie, art. 52 § 3 p.p.s.a.).	art. 1 ust. 1 ustawy z 24.03.1920 o nabywaniu nieruchomości przez cudzoziemców; art. 16 § 1 i § 3 k.p.a.; art. 217 § 2 pkt 2 k.p.a.
P05	v2 (2026-10-10)	Jak wygląda doręczenia pism sądowych osobie fizycznej przez portal ?	Od 1.06.2026 sąd doręcza przez Portal Informacyjny także stronie, która wybrała ten sposób doręczeń (art. 131 ^{1a} § 1 k.p.c.). Strona składa oświadczenie o wyborze przez portal, z podpisem kwalifikowanym, zaufanym albo osobistym (§ 1 ¹).	art. 131 ^{1a} § 1 k.p.c. (brzmienie od 1.06.2026); art. 131 ^{1a} § 1 ¹ k.p.c.; art. 131 ^{1a} § 2 k.p.c.
P06	v3 (2026-10-10)	Postanowienie o odrzuceniu zażalenia (SA) - brak opłaty, czy przysługuje zażalenie ?	Zależy od tego, jakie zażalenie SA odrzucił. Jeżeli było to zażalenie do innego składu sądu drugiej instancji (art. 394 ² § 1 lub § 1 ¹ k.p.c.), odrzuca je jeden sędzia, a na to postanowienie przysługuje zażalenie do innego składu SA (art. 394 ² § 1 ² zd. 2).	art. 394 ² § 1 ² k.p.c.; art. 397 § 1 ¹ k.p.c.
P07	v2 (2026-10-10)	Czy jak Sąd wyda zasadnicze motywy rozstrzygnięcia, to jest możliwe złożenie wniosku o uzasadnienie Zasadnicze powody rozstrzygnięcia (art. 357 § 5 k.p.c.)	Tak. Wskazanie zasadniczych powodów rozstrzygnięcia nie jest uzasadnieniem — art. 357 § 5 zd. 2 k.p.c. każe je odróżnić od uzasadnienia.	art. 357 § 5 k.p.c.; art. 357 § 1 albo § 2 ¹ k.p.c.

PYTANIE	WERSJA KLUCZA (DATA)	PYTANIE W BRZMIENIU Z BIEGU	RDZEŃ KLUCZA (SKRÓT)	PODSTAWY WYMAGANE
P08	v4 (2026-10-18)	Treścią typu deepfake nie będzie głos wygenerowany przez AI, który niczego nie udaje i w żaden sposób nie wprowadza w błąd. czy to prawda na gruncie AI Act?	W zasadzie tak, co do samej definicji: deepfake to treść, która przypomina istniejące osoby, przedmioty, miejsca, podmioty lub zdarzenia i którą odbiorca mógłby niesłusznie uznać za autentyczną lub prawdziwą (art. 3 pkt 60). Decyduje obiektywna możliwość wzięcia za autentyczne, nie zamiar ani to, czy głos „udaje” konkretną osobę — według wytycznych KE (C(2026) 5054, soft law) kryterium „istniejące” obejmuje też osoby, które mogą realnie istnieć, więc realistyczny syntetyczny głos może być deepfake, jeśli odbiorca mógłby go wziąć za nagranie prawdziwego człowieka.	art. 3 pkt 60 AI Act; art. 50 ust. 2 AI Act; art. 50 ust. 4 AI Act
P09	v4 (2026-10-18)	Przeniesienie praw autorskich do materiałów wytworzonych w ramach umowy zlecenia (social media) — rozumiem, że ogólne stwierdzenie „przenosi na wszelkie pola eksploatacji”, czyli ogólny zapis, nie wystarcza?	Co do zasady nie wystarcza (art. 41 ust. 2 i 4, art. 53 pr.aut.). Według SN (I CSK 104/18, I CSK 596/17, IV CSKP 44/21) pola mogą być dookreślone w drodze wykładni z art. 65 k.c. z treści i celu umowy, w razie wątpliwości na korzyść twórcy — kalibracja warunkowa.	art. 41 ust. 2 pr.aut.; art. 53 pr.aut.
P10	v3 (2026-10-18)	Przeniesienie praw autorskich do materiałów wytworzonych w ramach umowy zlecenia (social media) — rozumiem, że ogólne stwierdzenie „przenosi na wszelkie pola eksploatacji” wystarcza i nie trzeba ich wyliczać?	Nie — co do zasady nie wystarcza; wykładnia z art. 65 k.c. może dookreślić pola, ale nie zwalnia z wyliczenia. Odpowiedź „wystarcza i nie trzeba wyliczać” pozostaje błędem (pułapka); sprzeczność z własną odpowiedzią na P09 jest karana.	art. 41 ust. 2 pr.aut.; art. 53 pr.aut.
P12	v1 (2026-10-10)	Klauzula wykonalności na małżonka, pozew był związany z działalnością gospodarczą, ale w NZ nie ma nic na temat "prowadzący działalność" w klauzuli jest PESEL i nie ma żadnej wzmianki nt. działalności, czy można wnosić o klauzulę na małżonka w "normalnym trybie"? tj. czy można wnieść o egzekucję "bez ograniczeń", które wynikają z 781[1] k.p.c. ?	Nie w tym sensie, o jaki pyta: klauzuli przeciwko małżonkowi „bez ograniczeń” nie ma. Każda klauzula przeciwko małżonkowi dłużnika jest ograniczona: do majątku objętego wspólnością (art. 787 k.p.c., gdy wierzytelność powstała z czynności prawnej dokonanej za zgodą małżonka) albo do przedsiębiorstwa wchodzącego w skład majątku wspólnego (art. 787 ¹ k.p.c., gdy wierzytelność powstała w związku z prowadzeniem przedsiębiorstwa).	art. 787 k.p.c.; art. 787 ¹ k.p.c.
P13	v1 (2026-10-10)	Sąd wydał NZ w nakazowym z weksla i wstrzymał jego wykonanie. Czy obejmuje to też możliwość wszczęcia w oparciu o NZ post zabezpieczającego ?	Nie. Wstrzymanie wykonania nakazu (art. 492 § 3 zd. 2 k.p.c.) dotyczy natychmiastowej wykonalności nakazu z weksla jako tytułu egzekucyjnego (§ 3 zd. 1), a nie jego funkcji tytułu zabezpieczenia, którą nakaz ma od chwili wydania, niezależnie od wykonalności i bez klauzuli (§ 1).	art. 492 § 1 k.p.c.; art. 492 § 3 k.p.c.

ANEKS B

Wyniki: pytanie × system

Wynik każdej ze 160 odpowiedzi i ścieżki do materiału źródłowego.

Tabela B.1. Wynik bloku A (A %) każdej odpowiedzi

SYSTEM	P01	P04	P05	P06	P07	P08	P09	P10	P12	P13
Sol W0	100	44	22	22	89	82	90	100	90	89
Sol W1	89	44	100	78	89	73	100	100	90	78
Astra W0	100	44	22	44	89	73	100	100	90	89
Astra W1	89	44	100	50	100	73	90	100	90	89
Terra W0	78	11	0	11	0	46	80	90	90	11
Terra W1	0	44	100	78	89	73	90	80	95	78
Opus W0	78	56	56	6	0	75	75	75	80	50
Opus W1	89	89	89	61	100	92	73	90	100	78
Sonnet W0	0	22	56	17	39	75	82	80	65	0
Sonnet W1	72	22	78	67	67	83	65	75	60	89
Haiku W0	0	22	22	17	22	82	65	35	25	11
Haiku W1	78	39	0	17	28	71	55	75	55	78
Flash W0	78	60	6	6	0	42	64	73	80	89
Flash W1	89	44	100	75	90	58	64	82	90	89
Argon W0	83	33	0	17	90	58	77	82	80	89
Argon W1	89	67	83	75	0	92	82	91	85	90

Tabela B.2. Halucynacje i poprawne ugruntowanie (litery: klasy błędu z pola A6_typ; ✓: odpowiedź poprawnie ugruntowana; –: bez halucynacji, rdzeń niepełny; blokada: blokada dostawcy)

SYSTEM	P01	P04	P05	P06	P07	P08	P09	P10	P12	P13
Sol W0	✓	–	–	T	✓	–	✓	✓	✓	✓
Sol W1	✓	–	✓	✓	✓	–	✓	✓	✓	T
Astra W0	✓	–	–	W	✓	–	✓	✓	✓	✓
Astra W1	✓	–	✓	–	✓	–	✓	✓	✓	✓
Terra W0	W	T	W	F	T	–	✓	✓	✓	T
Terra W1	W	–	✓	✓	W	–	✓	W	✓	S
Opus W0	T,W	T	S,T	T,W	T	–	W	W	T	T,W
Opus W1	W	–	W	W	✓	✓	T	✓	✓	T,W
Sonnet W0	T,W	T	T	T,W	T,W	–	T	F,T	T	T
Sonnet W1	T	T	W	W	S	–	F,W	T,W	F,T	F,T

SYSTEM	P01	P04	P05	P06	P07	P08	P09	P10	P12	P13
Haiku W0	W	T	T	T	F,T	–	T	F,T	T	T
Haiku W1	T	T	F,T	T	T	–	T	T	–	W
Flash W0	T,W	T	T,W	T	S,T	T	T	T	✓	✓
Flash W1	T,W	–	✓	✓	W	T	T	✓	✓	S
Argon W0	W	T	T,W	S,W	S,T,W	T	✓	T	T	✓
Argon W1	T,W	T	T,W	T	blokada	✓	T	S	T	S,W

Materiał źródłowy. Standardy benchmarku (specyfikacja oceny z regułami i przykładami granicznymi), dziesięć pytań testowych wraz z odpowiedziami szesnastu systemów i pełnymi kluczami, karty ocen, oceny sędziów-modeli, raporty warstwy weryfikacji, zestawienia kosztów i skrypty generujące tabele tego raportu opublikujemy wkrótce po tym raporcie w repozytorium LexAlpha Lab na GitHub (<https://github.com/apiotrowski-afk>), po przeglądzie kart; do tego czasu materiały udostępniamy na prośbę. Karty są wersjonowane datą ostatniej kontroli zbioru; ten raport liczy wszystko na kartach w stanie z 25 października 2026 r.

ANEKS C

Sędziowie: zgodność i ustawienia

Zgodność sędziów według pytania i ustawienia sędziów.

Tabela C.1. Zgodność etykiety hallucinated z kartą według pytania (zgodne / ocenione karty; 16 systemów)

SĘDZIA	P01	P04	P05	P06	P07	P08	P09	P10	P12	P13
Gemini 4 Argon	16/16	15/16	15/15	12/12	13/13	15/16	15/15	14/16	14/14	13/14
GPT-6.1 Sol	15/16	16/16	16/16	15/16	15/16	11/16	15/16	15/16	13/16	15/16
Gemini 3.1 Pro	14/16	14/16	13/16	14/15	15/16	15/16	14/16	15/16	11/16	15/16
Gemini 3.8 Flash	9/16	14/16	13/16	15/16	15/16	14/16	13/16	13/16	15/16	15/16
GPT-6 Astra	2/2	1/1	9/9	9/9	9/9	1/1	1/1	1/1	8/8	7/8

Tabela C.2. Ustawienia sędziów

SĘDZIA (IDENTYFIKATOR)	MODEL	DOSTĘP	USTAWIENIA	ZAPIS WYSZUKAŃ
Gemini 4 Argon (argon4v101)	gemini-4-argon	Vertex AI, region global	rozumowanie high, temperatura 0, Google Search	zapytania i źródła (do 20 zapytań i 25 źródeł na kartę)
Gemini 3.1 Pro (gemini31v101)	gemini-3.1-pro-preview	Vertex AI, region global	temperatura 0, Google Search	jak wyżej
Gemini 3.8 Flash (gemini38flashv101)	gemini-3.8-flash	Vertex AI, region global	temperatura 0, Google Search	jak wyżej
GPT-6.1 Sol (sol61v101)	gpt-6.1-sol	Codex CLI	rozumowanie medium, wyszukiwanie na żywo, tryb tylko do odczytu	liczba wyszukiwań; od 25.10 także zapytania i źródła
GPT-6 Astra (astra6v101)	gpt-6-astra	Codex CLI	jak wyżej	jak wyżej
Karta meta: argon4meta, argon4meta2, argon4meta3	gemini-4-argon	Vertex AI	jak argon4v101 + raport weryfikacji (v1 wiążący, v2 pomocniczy, v3)	jak wyżej
Karta meta: sol61meta, sol61meta2, sol61meta3	gpt-6.1-sol	Codex CLI	jak sol61v101 + raport weryfikacji	jak wyżej

Plik oceny sędziego zawiera punkty każdego kryterium, klasy błędu, etykiety hallucinated, listę sprawdzonych powołań ze źródłem i werdyktem, listę „niesprawdzone”, uzasadnienie oraz metadane: identyfikator sędziego, model, pytanie, system, liczbę wyszukiwań, tokeny, czas, datę, wersję specyfikacji i informację, że ocena była ślepa (bez nazwy systemu). Biegi sędziów v101 odbyły się w dniach 13–20 października 2026 r.

ANEKS D

Słownik pojęć

Pojęcia, nazwy robocze i ich odpowiedniki w artykule angielskim, identyfikatory systemów.

Tabela D.1. Warunki, kryteria i etykiety

POJĘCIE	ZNACZENIE W RAPORCIE
W0	bieg bez narzędzi: bez wyszukiwania i bez tekstów aktów; odpowiedź z wag modelu
W1	bieg z wyszukiwaniem w internecie i poleceniem podania adresów źródeł
A-R	rdzeń odpowiedzi: 2 pełny, 1 kierunek trafny bez kluczowego warunku, 0 brak
A2, A3, A5, A8	podstawy wymagane; dosłowność cytatów; sporność; ramy odpowiedzi
A6	halucynacje: 1 albo 0; 0 przy co najmniej jednym twierdzeniu o źródle prawa niezgodnym ze źródłem urzędowym na datę stanu prawnego
R, K	elementy rubryki (+1) i kary (-1) z klucza
B1, B2	przydatność dla prawnika (0–2) i kalibracja pewności (0–1)
Δ	A % – B %, różnica opisowa
F	źródło albo jednostka nie istnieje
T	źródło istnieje, ale nie zawiera przypisanej treści
W	przepis w brzmieniu z innej daty podany jako obowiązujący
S	prawdziwe źródło z przekreślonym identyfikatorem
O	klasa zniesiona 11.10.2026 (odczyt sprzeczny z brzmieniem)
A6_pochodzenie	pamięć / odczyt źródła / źródło wtórne / nieustalone
A6_zakres	rdzeń (konkluzja, podstawy wymagane, element rubryki) / poboczne
bez weryfikacji źródła	tak (brak narzędzi) / tak (narzędzia nieużyte) / tak (tylko źródła wtórne) / nie
poprawnie ugruntowana	A6 = 1 i A-R = 2
blokada dostawcy	pusta odpowiedź zablokowana przez usługę; A6 = n/d, osobna kategoria
falszywy alarm (nadmiar)	sędzia: halucynacja; karta: brak
pominięcie (brak)	karta: halucynacja; sędzia: brak
zbiór pilotażowy	P05, P06, P07, P12, P13; karty sprawdzane po obejrzeniu sędziów
rozszerzenie	P01, P04, P08, P09, P10; karty sprawdzone przed sędziami
pytania testowe	osiem pytań, na których nie strojono warstwy weryfikacji: P01, P04–P08, P12, P13

Tabela D.2. Nazwy robocze i nazwy w artykule angielskim

NAZWA W RAPORCIE	NAZWA ROBOCZA W REPOZYTORIUM	NAZWA W ARTYKULE V1.2
poprawnie ugruntowana	poprawnie ugruntowana	correctly grounded
karta oceny (karta wzorcowa)	karta, oceny/<P>_<system>.md	scoring card (reference card)

NAZWA W RAPORCIE	NAZWA ROBOCZA W REPOZYTORIUM	NAZWA W ARTYKULE V1.2
sędzia-model	sędzia, _sedziowie/*.json	model judge
data stanu prawnego	data_stanu_prawnego	legal-state date
warstwa weryfikacji	meta-sędzia (krok faktów + poparcie)	verification layer
krok faktów	fakty.py , raport faktów	citation check
weryfikator poparcia	poparcie.py	support check
raport weryfikacji	raport faktów	verification report
karta meta	karta_meta.py, *.argon4meta*.json , *.sol61meta*.json	judge with the verification report
raport wiążący / pomocniczy	meta / meta2	binding / asymmetric report
status weryfikacji	status_weryfikacji	verification status
tekst niedostępny	tekst_niedostepny	source text unavailable
brzmienie na datę nieustalone	na_dzien nieustalone	wording on the legal-state date undetermined
sygnatura niejednoznaczna	sygnatura_niejednoznaczna	ambiguous case number
akt nieustalony	akt_nieustalony	act not identified
zasięg weryfikacji	„sprawdzone X z Y”	verification coverage

Tabela D.3. Identyfikatory systemów

MODEL	W0	W1
GPT-6.1 Sol	chatgpt_codex_w0_r3	chatgpt_codex_web_r3
GPT-6 Astra	codex_astra6_w0_r3	codex_astra6_web_r3
GPT-5.6 Terra	codex_terra56_w0_r3	codex_terra56_web_r3
Claude Opus 5.5	opus_cli_w0_r3	opus_cli_w1_r3
Claude Sonnet 5.5	sonnet_cli_w0_r3	sonnet_cli_w1_r3
Claude Haiku 5.5	haiku_cli_w0_r3	haiku_cli_w1_r3
Gemini 3.8 Flash	gemini38flash_vertex_w0_r3	gemini38flash_vertex_w1_r3
Gemini 4 Argon	argon4_high_vertex_w0_r3	argon4_high_vertex_w1_r3

ANEKS E

Macierze pomyłek i przedziały niepewności dla każdej k

Każda k w rozdziałach 6 i 7 streszcza tabelę 2 × 2: zarówno karta wzorcowa, jak i ocena modelu wskazują halucynację (TP), tylko ocena modelu (FP, fałszywy alarm), tylko karta (FN, pominięcie), żadna (TN). Kartę z blokadą dostawcy (P07 Gemini 4 Argon W1) liczymy jak kartę bez halucynacji; wszyscy sędziowie, którzy ją ocenili, także nie przypisali jej halucynacji. Tabela E.1 podaje te cztery liczby dla każdego porównania wraz z liczbą halucynacji w kartach, zgodnością surową, κ i przedziałem 95 %. Przedział to bootstrap percentylowy po pytaniach: losujemy całe pytania ze wszystkimi ich kartami, ze zwracaniem, 4000 razy; uwzględnia to zależność między odpowiedziami na to samo pytanie. Przy czterech do ośmiu pytaniach przedziały są szerokie; ich szerokość odzwierciedla niepewność oszacowania przy małej liczbie pytań; przedział 1,00–1,00 znaczy tylko, że w żadnym pytaniu nie było niezgodności. Osobne przedziały nie testują różnicy między sędziami wprost, bo sędziowie oceniali te same odpowiedzi; porównanie sparowane wymagałoby większej liczby pytań. Liczbę halucynacji w kartach podajemy, bo κ od niej zależy: ta sama zgodność surowa daje niższą κ , gdy udział kart z halucynacją jest mały. Wszystkie wiersze liczone na kartach w stanie z 25.10.2026; zbiór pilotażowy obejmuje tu 16 systemów raportu (80 kart), inaczej niż 65 kart 13 systemów w artykule.

Tabela E.1. Macierze pomyłek, κ i przedziały 95 % (etykieta hallucinated; sędziowie w specyfikacji v1.0.1; karty meta na kartach wspólnych z sędzią bez raportu; $TN = n - TP - FP - FN$; $\text{zgodność} = (TP + TN) / n$; „karty v1/v2/v3” = karty ocenione w danym wariancie, „wspólne v2 i v3” = karty ocenione w obu)

SĘDZIA / REGUŁA	ZBIÓR	N	HAL. W KARTACH	TP	FP	FN	κ	PRZEDZIAŁ 95 %
Gemini 4 Argon	rozszerzenie	79	41	38	1	3	0,90	0,80–0,97
Gemini 4 Argon	pilotaż	68	39	38	0	1	0,97	0,90–1,00
GPT-6.1 Sol	rozszerzenie	80	42	41	7	1	0,80	0,59–0,95
GPT-6.1 Sol	pilotaż	80	49	49	6	0	0,84	0,73–0,94
Gemini 3.1 Pro	rozszerzenie	80	42	34	0	8	0,80	0,75–0,85
Gemini 3.1 Pro	pilotaż	79	48	42	5	6	0,71	0,52–0,86
Gemini 3.8 Flash	rozszerzenie	80	42	26	1	16	0,58	0,40–0,70
Gemini 3.8 Flash	pilotaż	80	49	44	2	5	0,82	0,71–0,87
GPT-6 Astra	rozszerzenie	6	0	0	0	0	1,00	1,00–1,00
GPT-6 Astra	pilotaż	43	29	29	1	0	0,95	0,82–1,00
Gemini 4 Argon sam	rozszerzenie, 4 sędziów	79	41	38	1	3	0,90	0,80–0,97
Argon albo Gemini 3.1 Pro	rozszerzenie, 4 sędziów	79	41	39	1	2	0,92	0,87–0,97
większość z: Argon, Sol, Pro	rozszerzenie, 4 sędziów	79	41	38	1	3	0,90	0,86–0,95
którykolwiek z czterech	rozszerzenie, 4 sędziów	79	41	41	7	0	0,82	0,61–0,97
Gemini 4 Argon, bez raportu	karty v1	54	35	34	0	1	0,96	0,88–1,00
Gemini 4 Argon, z raportem v1	karty v1	54	35	31	0	4	0,85	0,76–0,95
Gemini 4 Argon, bez raportu	karty v2	115	62	59	0	3	0,95	0,90–0,98

SĘDZIA / REGUŁA	ZBIÓR	N	HAL. W KARTACH	TP	FP	FN	K	PRZEDZIAŁ 95 %
Gemini 4 Argon, z raportem v2	karty v2	115	62	56	1	6	0,88	0,86–0,91
Gemini 4 Argon, bez raportu	karty v3	116	63	60	0	3	0,95	0,90–0,98
Gemini 4 Argon, z raportem v3	karty v3	116	63	58	2	5	0,88	0,86–0,91
Gemini 4 Argon, bez raportu	wspólne v2 i v3	115	62	59	0	3	0,95	0,90–0,98
Gemini 4 Argon, z raportem v2	wspólne v2 i v3	115	62	56	1	6	0,88	0,86–0,91
Gemini 4 Argon, z raportem v3	wspólne v2 i v3	115	62	57	2	5	0,88	0,85–0,91
GPT-6.1 Sol, bez raportu	karty v1	61	41	41	3	0	0,88	0,81–0,96
GPT-6.1 Sol, z raportem v1	karty v1	61	41	36	0	5	0,83	0,75–0,93
GPT-6.1 Sol, bez raportu	karty v2	119	68	67	8	1	0,84	0,70–0,96
GPT-6.1 Sol, z raportem v2	karty v2	119	68	67	1	1	0,97	0,92–1,00
GPT-6.1 Sol, bez raportu	karty v3	128	73	72	11	1	0,80	0,68–0,93
GPT-6.1 Sol, z raportem v3	karty v3	128	73	70	4	3	0,89	0,82–0,95
GPT-6.1 Sol, bez raportu	wspólne v2 i v3	119	68	67	8	1	0,84	0,70–0,96
GPT-6.1 Sol, z raportem v2	wspólne v2 i v3	119	68	67	1	1	0,97	0,92–1,00
GPT-6.1 Sol, z raportem v3	wspólne v2 i v3	119	68	65	2	3	0,91	0,85–0,98

Trzy odczyty. Po pierwsze, na rozszerzeniu przedział Gemini 3.8 Flash nie pokrywa się z przedziałem Gemini 4 Argon; pozostałe różnice między sędziami mieszczą się w przedziałach. Po drugie, reguły paneli różnią się od samego Gemini 4 Argon od +0,02 („Argon albo Pro”) do –0,08 („którykolwiek z czterech”), a ich przedziały się pokrywają. Po trzecie, u GPT-6.1 Sol przedziały z raportem v2 i v3 się pokrywają, podobnie jak przedziały bez raportu i z raportem v2 (na krawędzi); kierunek zmiany jest spójny, ale jej wielkość przy ośmiu pytaniach pozostaje nierozstrzygnięta. Wniosek z rozdziału 7.5 o wariancie v3 jest z tego powodu hipotezą roboczą.

LITERATURA

Literatura

- [1] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., Ho, D. E. (2025). *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*. *Journal of Empirical Legal Studies*, 22(2), s. 216–242, doi:10.1111/jels.12413; preprint arXiv: 2405.20362 (2024). W tekście powołujemy: definicję halucynacji jako odpowiedzi niepoprawnej albo błędnie osadzonej w źródłach i pojęcie ugruntowania.
- [2] Taranukhin, M., Shwartz, V. (2026). *Legal LLM Hallucination Should Be Evaluated as Failure of Legal Warrant*. AI4Law @ ICML 2026, arXiv:2609.17546. W tekście powołujemy: warunki „legal warrant” — źródło istnieje, dotyczy właściwej jurysdykcji, jest aktualne na datę analizy, ma przypisany status i popiera twierdzenie.
- [3] Kohli, G. (2026). *Nine Judges, Two Effective Votes: Correlated Errors Undermine LLM Evaluation Panels*. Preprint, arXiv: 2605.29800. W tekście powołujemy: panel dziewięciu sędziów z siedmiu rodzin modeli odpowiada około dwóm niezależnym głosom; najlepszy pojedynczy sędzia dorównuje panelowi.
- [4] Lee, D., Hwang, Y., Kang, T., Lee, M., Chae, Y., Jung, K. (2026). *Judging Against the Reference: Uncovering Knowledge-Driven Failures in LLM-Judges on QA Evaluation*. Preprint, arXiv:2601.07506. W tekście powołujemy: sędzia ignoruje referencję sprzeczną z własną wiedzą.
- [5] Yang, H., Song, W., Kim, T., Song, J., Park, S., Jo, Y. (2026). *Agents' Overreliance on Unreliable Tools*. Preprint, arXiv: 2609.05587. W tekście powołujemy: modele przejmują podrobione wyniki narzędzi, dla wyszukiwania w 68,0 % przypadków.
- [6] Butler, A.-R., Butler, U. (2026). *Legal RAG Bench: an end-to-end benchmark for legal RAG*. Preprint, arXiv:2603.01710. W tekście powołujemy: jakość wyszukiwania wyznacza górną granicę jakości odpowiedzi systemu prawniczego z wyszukiwaniem.
- [7] Piotrowski, A. (2026). „Hallucination” Is Not Enough: Why Legal LLM Benchmarks Should Measure „Correctly Grounded”. LexAlpha Lab, KTZR, wersja 1.2 z 26.10.2026, doi:10.5281/zenodo.23674794 (wszystkie wersje: doi:10.5281/zenodo.23438195; wersja 1.1 z 17.10.2026: doi:10.5281/zenodo.23438196); SSRN 7672779; <https://www.ktzt.pl/legal-llm-benchmarks-correctly-grounded/>.
- [8] Piotrowski, A., Koman, A. (2026). *Benchmark systemów AI pracujących z polskim prawem. Propozycja formatu i pilotaż na kazusie terminu zawitego sankcji kredytu darmowego*. LexAlpha Lab, wersja z 1.10.2026 (Benchmark 01).

ŹRÓDŁA PRAWNE I ORZECZENIA POWOŁANE W RAPORCIE

Źródła prawne i orzeczenia powołane w raporcie

- Ustawa z dnia 20 maja 1971 r. — Kodeks wykroczeń, art. 119 § 1, 120 § 1, 122 § 1, 124 § 1 (t.j. Dz.U. 2025 poz. 734; brzmienie z t.j. Dz.U. 2015 poz. 1094 jako dawne).
- Ustawa z dnia 6 czerwca 1997 r. — Kodeks karny, art. 278a (uchylony z dniem 1.10.2023, Dz.U. 2022 poz. 2600).
- Ustawa z dnia 24 marca 1920 r. o nabywaniu nieruchomości przez cudzoziemców, art. 1 ust. 1 (t.j. Dz.U. 2017 poz. 2278).
- Ustawa z dnia 14 czerwca 1960 r. — Kodeks postępowania administracyjnego, art. 16 § 1 i § 3, 127 § 3, 217 § 2 pkt 2, 218 § 1; ustawa z dnia 30 sierpnia 2002 r. — Prawo o postępowaniu przed sądami administracyjnymi, art. 52 § 3, 53 § 1.
- Ustawa z dnia 17 listopada 1964 r. — Kodeks postępowania cywilnego, art. 131^{1a} (w brzmieniu od 1.06.2026), 357 § 5, 394² § 1², 397 § 1¹, 492 § 1 i § 3, 787, 787¹.

- Ustawa z dnia 28 lipca 2005 r. o kosztach sądowych w sprawach cywilnych, art. 25b (t.j. Dz.U. 2025 poz. 1228).
- Ustawa z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych, art. 41 ust. 2 i 4, 44, 52, 53 (t.j. Dz.U. 2025 poz. 24; nowela Dz.U. 2024 poz. 1254).
- Rozporządzenie (UE) 2024/1689 (AI Act), art. 3 pkt 60, art. 50 ust. 2 i 4; Komisja Europejska, wytyczne C(2026) 5054 do art. 50 (soft law).
- Sąd Najwyższy: III CK 124/05 (14.09.2005), I CSK 596/17 (5.10.2018), I CSK 104/18 (22.03.2019), IV CSKP 44/21 (14.10.2021), III CK 400/03 (23.09.2004), II CSKP 526/22 (12.01.2023), I CZ 83/09, III CZP 24/23, III CZP 50/22.
- Sądy powszechne: SA w Łodzi I ACa 1157/12 (27.02.2013) i I ACa 1188/12; SA w Rzeszowie I ACz 390/13; SO w Szczecinie VIII Gz 43/16.