ChatGPT

Agent silniejszy od modelu i pricing jako dźwignia topologii

Streszczenie wykonawcze

W analizowanym tygodniu (okno dat: ósmy–piętnasty lutego) wzorzec „agent > model" stał się wyraźniejszy, bo trzy strumienie zdarzeń nałożyły się na siebie: (a) eskalacja sporu o **granice użycia** modeli w zastosowaniach militarnych (permissions i „hard limits" jako oś konfliktu), (b) eskalacja sporu o **własność wartości modelowej** poprzez distillation, oraz (c) dojrzewanie **wielolicznikowego pricingu** (tokeny + cache/batch + narzędzia + storage + fine-tuning + metryki czasu/priorytetu) jako realnego regulatora architektury. [1]

Kluczowy wniosek: model (wagi) coraz częściej jest **wymiennym silnikiem** (commodity compute), podczas gdy agent jest **warstwą własności i kontroli**: utrzymuje pamięć, polityki (RBAC/limity), ślady pochodzenia (provenance), budżety działań i logikę narzędzi — czyli to, co użytkownik/firma chce „mieć w kieszeni". W efekcie zmiana pricingu z „$/token" w stronę „$/zadanie" (mierzone liczbą kroków, narzędzi i retencją) przeorganizowuje topologię systemów: cache-first, RAG jako amortyzator kosztu kontekstu, hybryda/on-device dla wrażliwych przepływów, a także rosnące znaczenie audytu i rozliczalności. [2]

Sygnały z analizowanego tygodnia

Depesze Reuters [3] z tego okna dat pokazują, że napięcie wokół „autonomii vs ograniczeń" przestało być abstrakcją, a stało się elementem negocjacji kontraktów i dostępu do infrastruktury: Departament Obrony USA [4] rozważał zakończenie relacji z Anthropic [5], bo firma utrzymywała ograniczenia użycia (m.in. twarde limity dotyczące autonomicznej broni i masowej inwigilacji). To jest sygnał, że „permissions i boundaries" nie są już detalem UX, tylko rdzeniem produktu agentowego. [6]

Równolegle, OpenAI [7] w memo do ustawodawców oskarżył DeepSeek [8] o próby odtwarzania zdolności modeli przez distillation oraz obchodzenie ograniczeń dostępu (m.in. maskowanie źródła przez routery stron trzecich). To ustawia IP modeli jako pole walki o „wyciek wartości" i wzmacnia tezę, że trwały moat przesuwa się do warstwy agenta (kontekst, pamięć, integracje), trudniejszej do skopiowania samą destylacją. [9]

W tym samym oknie czasowym pojawiły się jednocześnie bodźce rynkowe: Blackstone [10] zwiększył ekspozycję kapitałową na Anthropic do ok. 1 mld USD; Reuters cytuje źródło mówiące o wycenie rzędu ~350 mld USD (wartość jest niezwykle wysoka i należy ją traktować jako deklarację „source familiar", nie twarde dane audytowane). [11] Z tym skorelował globalny niepokój o „AI-disruption" w software, widoczny nawet w komentarzach o ryzyku kredytowym sektora (Morgan Stanley). [12]

Dodatkowy sygnał „hybryda/on-device jako kierunek topologiczny" pokazał Business Insider [13]: z akt sądowych wynikało przesunięcie terminu wysyłek tajnego urządzenia AI współtworzonego z Jony Ive [14]; niezależnie od harmonogramu, sam fakt, że oś sporu biegnie przez branding i dystrybucję sprzętu, wzmacnia hipotezę o dążeniu do „wartości w kieszeni" (lokalność, kontrola kontekstu, prywatność, koszt). [15]

Dlaczego agent wygrywa technicznie

Technicznie przewaga agenta wynika z tego, że rozdziela on system na dwie warstwy o innych własnościach ekonomicznych i prawnych: **silnik generacji** (model) oraz **warstwę kontroli i własności** (agent). W analizowanym tygodniu widać to w komunikatach i produktach: OpenAI opublikował informacje o modelu „agentic coding" (GPT-5.3-Codex) oraz aktualizację GPT-5.2 Instant, a komunikaty produktowe wprost wskazują przejście od „code generation" do „coding agent you can actively steer while it works". To jest definicyjnie agentowość: dłuższy horyzont, sterowalność w trakcie, iteracje i sprzężenie z narzędziami. [16]

U Anthropic [5] komunikat o Claude Opus 4.6 podkreślał zdolność do dłuższego utrzymywania „agentic tasks", niezawodniejszej pracy na większych codebase'ach i bardzo dużego okna kontekstu (w beta). To wzmacnia wartość „pamięci operacyjnej" — ale jednocześnie powiększa rachunek (więcej tokenów, większa powierzchnia błędów i nadużyć, większa potrzeba audytu). [17]

Agent jest „silniejszy" także dlatego, że w praktyce to on realizuje: - **pamięć i kompresję kontekstu** (np. compaction/auto-compact jako mechanizmy zarządzania kosztem i ciągłością rozmowy), [18]
- **polityki i uprawnienia** (RBAC/role, limity, zgodność, logowanie), co wprost pojawia się w release notes dot. Codex w środowiskach enterprise (kontrola dostępu i logowanie użycia), [19]
- **provenance i rozliczalność**: standardy pochodzenia treści są projektowane jako kryptograficznie weryfikowalne metadane zmian i źródeł, co lokuje „prawa do prawdy i autorstwa" w warstwie narzędziowej/produktowej, nie w wagach modelu. [20]

W tym sensie teza, którą sygnalizujesz w swoich wcześniejszych badaniach (pricing → topologia → kaskady), zyskuje tu techniczny rdzeń: **agent jest interfejsem własności** (danych, działań, pamięci, polityk), a model jest interfejsem wydajności.

Mechanizmy pricingu i kaskady topologii

W tygodniu analizowanym istotność pricingu wzrosła, bo koszty nie są już jednowymiarowe („$/token"), tylko składają się na **cenę zadania** (cost-per-task), szczególnie w systemach agentowych. Ujęcie formalne:

Koszt zadania ≈
*(tokeny wejścia × stawka) + (tokeny wyjścia × stawka) + (tokeny cachowane × stawka) + (wywołania narzędzi × stawka) + (storage × GB-dzień) + (fine-tuning: trening × stawka) + (fine-tuning: inferencja × stawka) + (czas treningu RFT × stawka) + (overhead bezpieczeństwa/audytu). [21]

To rozbicie na liczniki jest udokumentowane w źródłach pierwotnych: - OpenAI podaje osobne ceny za fine-tuning (wejście/wyjście/trening) oraz dla RFT rozliczenia czasowe (wall-clock w rdzeniu pętli treningu) plus ewentualne tokeny „graderów". [22]
- OpenAI w pricingu narzędzi dolicza osobno m.in. storage file search w modelu „GB-day" i opłaty per „tool call"; dodatkowo widać przesunięcie w kierunku metryk czasu dla container usage (planowane rozliczanie per 20 minut). [23]
- Anthropic jawnie wprowadza ekonomię cache: cache-write i cache-read mają multiplikatory cenowe względem bazowej stawki, co tworzy natychmiastową zachętę do projektowania agentów pod wysoki cache-hit. [24]
- Anthropic oferuje Batch API z dyskontem na tokeny (asynchroniczność za niższą cenę), a dla bardzo długiego kontekstu wskazuje „premium" stawki po przekroczeniu progu. [25]

Mechanizm kaskady (wprost zgodny z Twoją wcześniejszą hipotezą „repricing → zmiana topologii → kaskady") działa typowo tak: 1) wzrost kosztu marginalnego (tokeny/kontekst/narzędzia) →
2) architektury ograniczają liczniki (kompresja, cache, batch, RAG) →
3) rośnie rola agenta jako optymalizatora i strażnika budżetu →
4) pojawia się presja na hybrydy/on-device (koszt + prywatność + compliance) →
5) w kolejnej iteracji zmienia się oferta cenowa (np. premium za szybkość, zniżki za asynchroniczność) i cykl się zamyka. [26]

To jest „pętla pricingowa" w czystej postaci: cena przestaje być tabelą, a staje się **sterowaniem topologią systemu**.

Tabela porównawcza przed/po (istota przesunięcia wartości; szczegóły implementacyjne zależą od dostawcy i są zmienne): | Wymiar | Przed: model-centric | Po: agent-centric | |---|---|---| | Kontrola | wagi modelu jako „centrum" | permissions, RBAC, budżety, audyt | | Prywatność | domyślnie cloud | hybryda, selektywna lokalność danych | | Monetyzacja | głównie $/token | $/task: tokeny+tools+storage+SLA | | Ryzyko IP | łatwiejszy „wyciek zdolności" | większy moat w workflow/pamięci/provenance | | Koszty operacyjne | głównie GPU inference | orkiestracja, FinOps, bezpieczeństwo narzędzi | | Szybkość innowacji | skoki wersji modeli | iteracje agentów, integracji i polityk |

Ważne doprecyzowanie: spadek ryzyka IP w modelu agent-centric nie jest automatyczny; rośnie rola zabezpieczeń i audytu, bo agent zwiększa „powierzchnię wykonawczą".

```
flowchart LR
  Model[Model LLM: silnik inferencji] --> Agent[Agent: pamięć + polityki + narzędzia]
  Agent --> Billing[Pricing: tokeny + cache/batch + tool-calls + storage + FT/RFT]
  Billing --> Behavior[Zachowania: cache-first, RAG, batch, compaction]
  Behavior --> Topology[Topologia: cloud -> hybrid -> edge/on-device]
  Topology --> Risk[Nowe ryzyka: agency, IP, compliance]
  Risk --> Agent
```

Skutki rynkowe, polityczne i energetyczne

Rynkowo tydzień był istotny, bo „agentowość" zaczęła wpływać na wyceny nie tylko przez obietnicę lepszych modeli, ale przez obietnicę „wdrażalnych pracowników AI" w enterprise. Reuters opisał zarówno dopływ kapitału do Anthropic, jak i równoległy niepokój o trwałość tradycyjnych modeli biznesowych SaaS wobec automatyzacji agentowej — do poziomu analizy ryzyka na rynku kredytowym. [27]

Politycznie/regulacyjnie tydzień skumulował dwa konflikty: (a) państwo chce rozszerzać użycie (w tym na sieciach niejawnych i w obszarach „lawful purposes"), (b) dostawca chce utrzymać ograniczenia i zasady odpowiedzialnego wdrożenia. To napędza rozwój agentów jako warstwy „governance-by-design". [28]

Energetycznie, w tym samym tygodniu Reuters relacjonował prognozę Energy Information Administration [29] o rekordowym zapotrzebowaniu na energię w USA w kolejnych latach z AI i centrami danych jako jedną z przyczyn. To wzmacnia znaczenie KPI „energy per inference" i sprzyja strategiom hybrydowym, gdzie część zadań amortyzuje się lokalnie lub asynchronicznie (batch). [30]

Ryzyka, KPI i rekomendacje

Ryzyka, które w tym tygodniu stały się bardziej „namacalne”, układają się w trzy klasy: (a) **ryzyka agency** (niepożądane działania, eskalacja uprawnień, błędy narzędzi), (b) **ryzyka IP** (distillation jako forma przenoszenia wartości modelowej, potencjalnie „model theft”), (c) **ryzyka compliance i prywatności** (dane osobowe w pamięci/retencji, terytorialny zakres GDPR). [31]

KPI do monitorowania (minimalny zestaw „systemowy”): - value-per-token i cost-per-task (łącznie ze składowymi tool-calls i storage), [32]
- cache-hit rate oraz oszczędność z batch (wpływ na jednostkowy koszt SLA), [33]
- energy per inference (lub energy per task jako proxy), w kontekście napięć sieciowych i wzrostu konsumpcji data center, [34]
- incydenty IP leakage (sygnały distillation/scraping/obejść ograniczeń), [9]
- metryki bezpieczeństwa agentów: „excessive agency”, prompt injection, supply chain narzędzi. [35]

Rekomendacje techniczne i organizacyjne (skondensowane do działań o największej dźwigni): 1) Projektuj agentów jako „budżetowane procesy”: limit kroków, limit narzędzi, limit storage, policy-as-code i logowanie decyzji, bo opłaty per tool-call/GB-day i ryzyka agency rosną nieliniowo wraz z autonomią. [36]
2) Wprowadź warstwę provenance: podpisywanie i audyt (np. standardy content credentials) dla krytycznych artefaktów, by ograniczać spory o pochodzenie i odpowiedzialność. [20]
3) Przygotuj strategię „repricing resilience”: cache-first + RAG + batch jako domyślne mechanizmy amortyzacji; compaction jako kontrola kosztu długich rozmów; to jest bezpośrednia odpowiedź na wielolicznikowy pricing. [37]
4) Włącz ramy zarządzania ryzykiem AI (governance, audyt, testy) oraz mapowanie danych osobowych i podstaw prawnych przetwarzania zgodnie z GDPR, bo „agentowość” przenosi ryzyka z modelu na integracje i pamięć. [38]
5) Traktuj distillation jako ryzyko ekonomiczne, nie wyłącznie techniczne: to mechanizm, który (w normalnych zastosowaniach) służy kompresji, ale w sporze rynkowym staje się wektorem erozji moat w wagach — co zwiększa wagę moat w agentach (workflow, dane, integracje). [39]

Źródła priorytetowe do dalszego śledzenia

Najbardziej „nośne” źródła pierwotne w tym wątku to: dokumentacje pricingu i narzędzi OpenAI oraz Anthropic (w tym tool-calls, storage, caching, batch, fine-tuning/RFT), oficjalne release notes modeli (agentic features), depesze Reuters i artykuły Axios/Business Insider z analizowanego tygodnia, literatura bazowa o distillation i federated learning, tekst GDPR oraz standard C2PA dla provenance. [40]

---

[1] [6] [8] [28] [29] Pentagon threatens to cut off Anthropic in AI safeguards dispute, Axios reports | Reuters
https://www.reuters.com/technology/pentagon-threatens-cut-off-anthropic-ai-safeguards-dispute-axios-reports-2026-02-15/

[2] [3] [23] [26] [32] [36] Pricing | OpenAI API
https://platform.openai.com/docs/pricing/

[4] [19] ChatGPT Enterprise & Edu - Release Notes | OpenAI Help Center
https://help.openai.com/en/articles/10128477-chatgpt-enterprise-edu-release-notes%23.svgz?utm_source=chatgpt.com

[5] [30] US power use to beat record highs in 2026 and 2027, EIA says | Reuters
https://www.reuters.com/business/energy/us-power-use-beat-record-highs-2026-2027-eia-says-2026-02-10/

7 24 33 37 Prompt caching - Anthropic
https://docs.anthropic.com/en/docs/build-with-claude/prompt-caching?utm_source=chatgpt.com

9 OpenAI says China's DeepSeek trained its AI by distilling US models, memo shows | Reuters
https://www.reuters.com/world/china/openai-accuses-deepseek-distilling-us-models-gain-advantage-bloomberg-news-2026-02-12/

10 17 Claude Opus 4.6
https://www.anthropic.com/news/claude-opus-4-6?utm_source=chatgpt.com

11 27 Blackstone boosts stake in AI startup Anthropic to about $1 billion, source says | Reuters
https://www.reuters.com/technology/blackstone-boosts-stake-ai-startup-anthropic-about-1-billion-source-says-2026-02-10/

12 AI-led software selloff may pose risk for $1.5 trillion U.S. credit market, says Morgan Stanley | Reuters
https://www.reuters.com/business/finance/ailed-software-selloff-may-pose-risk-15-trillion-us-credit-market-says-morgan-2026-02-10/

13 14 21 22 40 Pricing | OpenAI
https://openai.com/api/pricing/?utm_source=chatgpt.com

15 OpenAI Reveals Timeline for Mystery AI Hardware Device With Jony Ive - Business Insider
https://www.businessinsider.com/openai-timeline-hardware-ai-device-launch-jony-ive-iyo-2026-2

16 Model Release Notes | OpenAI Help Center
https://help.openai.com/es-es/articles/9624314-model-release-notes

18 Gérer les coûts efficacement - Anthropic
https://docs.anthropic.com/fr/docs/claude-code/costs?utm_source=chatgpt.com

20 Content Credentials : C2PA Technical Specification :: C2PA Specifications
https://spec.c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html?utm_source=chatgpt.com

25 Pricing - Claude API Docs
https://docs.anthropic.com/en/docs/about-claude/pricing

31 35 OWASP Top 10 for Large Language Model Applications | OWASP Foundation
https://owasp.org/www-project-top-10-for-large-language-model-applications/?utm_source=chatgpt.com

34 Energy demand from AI – Energy and AI – Analysis - IEA
https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai?utm_source=chatgpt.com

38 Artificial Intelligence Risk Management Framework (AI RMF 1.0) | NIST
https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10?utm_source=chatgpt.com

39 Neural Network Knowledge Distillation
https://www.emergentmind.com/papers/1503.02531?utm_source=chatgpt.com