

Jak zbudować własny model AI

Kompletna ścieżka od surowego tekstu do działającego asystenta — dziewięć etapów, cztery scenariusze budżetowe i uczciwa odpowiedź na pytanie, co da się zrobić samodzielnie, a czego nie da się zrobić za żadne pieniądze dostępne poza kilkoma korporacjami świata.

NA LAPTOPIE

150 zł

tydzień pracy
wagi w pełni własne

ADAPTACJA

0,7 mln zł

dziewięć miesięcy
realne dla polskiej spółki

8 MLD OD ZERA

15–38 mln zł

rok do półtora
zespół jedenastu osób

KLASA CHATGPT

1,2–5,5 mld \$

poza zasięgiem
kogośkolwiek w Polsce

Rachunek na jednej stronie

Dla czytelnika, który potrzebuje wyłącznie wniosków. Wszystko poniżej jest rozwinięte i udokumentowane w dalszej części opracowania.

6 · N · D

JEDYNY WZÓR, KTÓRY TRZEBA ZNAĆ

Nakład obliczeniowy treningu to sześciokrotność iloczynu liczby parametrów i liczby tokenów. Przy 1,76 USD za eksaoperację daje to koszt w dolarach — bez cudzych szacunków.

200–300 mld

TYŁE JEST POLSKIEGO TEKSTU

Cała dostępna, wysokiej jakości polszczyzna nasyci model o około 780 mln parametrów. To ograniczenie fizyczne, którego nie da się pokonać budżetem.

2 %

UDZIAŁ MOCY OBLICZENIOWEJ

W realnym projekcie dla polskiej firmy karty graficzne to ułamek budżetu. Ponad połowę pochłania zespół, resztę dane i ich anotacja.

Siedem wniosków

Po pierwsze, „własny model” oznacza trzy różne rzeczy. Własną warstwę na cudzych wagach, własne wagi wyprowadzone z cudzego punktu startowego albo własne wagi od zera. Kosztują odpowiednio setki złotych miesięcznie, dziesiątki tysięcy złotych i od tysiąca złotych do miliardów dolarów — zależnie od skali. Bez rozstrzygnięcia, o którym z tych trzech mówimy, żadna rozmowa o budżecie nie ma sensu.

Po drugie, trening od zera nie jest trudny technicznie. Kod referencyjny modelu ma pięćset pięćdziesiąt pięć linii i jest publicznie dostępny. Trudność jest wyłącznie ilościowa — buduje się nie algorytm, tylko elektrownię. Dlatego cała ta dziedzina jest przedsięwzięciem kapitałowym, nie badawczym.

Po trzecie, pretrening pochłania osiemdziesiąt do dziewięćdziesięciu procent budżetu, a jego produkt można pobrać za darmo. Wagi modeli otwartych są udostępniane publicznie przez podmioty, które opłaciły ten etap. Odtwarzanie go samodzielnie jest płaceniem za wiedzę już zakupioną przez kogoś innego.

Po czwarte, polszczyzna ma twardy sufit. Polski to dwa procent zasobów internetu. Dlatego wszystkie poważne polskie projekty — bez wyjątku — importują kompetencje ogólne z modelu wytrenowanego na angielszczyźnie i douczają polszczyzny w kolejnym etapie. Nie jest to niedbałość, tylko jedyne rozwiązanie zgodne z arytmetyką.

Po piąte, o jakości decyduje post-trening, a nie skala. Model, który zdobył rozgłos jako pierwszy publicznie dostępny model rozumujący, kosztował na etapie uczenia ze wzmocnieniem 294 tysiące dolarów — pięć procent kosztu swojej bazy. Zdolności robiące największe wrażenie powstają w najtańszym etapie, a decyduje o nich jakość danych, nie wielkość klastra.

Po szóste, w Polsce jest dostępna bezpłatna moc obliczeniowa. Przydział do pięćdziesięciu tysięcy godzin pracy kart, rozpatrywany w cztery dni robocze i bezpłatny dla małych i średnich przedsiębiorstw, odpowiada wartości rynkowej około 430 tysięcy złotych. Wystarcza na pełną adaptację językową modelu ośmiomiliardowego. Bardzo niewiele firm o tym wie.

Po siódme, twórca modelu jest w Unii adresatem konkretnych obowiązków. Dokumentacja techniczna, polityka przestrzegania prawa autorskiego i publiczne streszczenie danych treningowych to wymogi, a

nie dobre praktyki. Dokumentację prowadzi się od pierwszego dnia zbierania danych — odtworzenie jej po fakcie bywa niewykonalne.

Rekomendacja w jednym zdaniu

CO ZROBIĆ

Zbudować własny tokenizator dla polszczyzny, przeprowadzić kontynuowany pretrening otwartego modelu ośmiomiliardowego na przydziale publicznym i włożyć cały wysiłek w dane, których nikt inny nie ma. Koszt: poniżej siedmiuset tysięcy złotych w pierwszym roku. Efekt: własne wagi, działające bez połączenia z internetem, w dziedzinie, w której model ogólny jest słaby.

Trening od zera zarezerwować na sytuację, w której pojawi się powód architektoniczny albo licencyjny — a nie ambicjonalny. Pełne uzasadnienie w rozdziale siedemnastym.

Jak czytać dalej: część I ustala pojęcia, część II opisuje dziewięć etapów budowy z kosztem każdego, część III zawiera pełne kosztorysy, część IV obowiązki prawne, część V rekomendację. Załączniki A i B to gotowe przepisy wykonawcze — pierwszy na laptopa, drugi na wynajęty klaster.

Spis treści

PUNKT WYJŚCIA

- | | | |
|---|---|---|
| 1 | Z czego naprawdę składa się ChatGPT | 4 |
| | siedem warstw, z których model to tylko pięć | |
| 2 | Trzy poziomy własności modelu | 6 |
| | co znaczy „nasz model” i ile każde znaczenie kosztuje | |

BUDOWA KROK PO KROKU

- | | | |
|----|--|----|
| 3 | Krok 1. Zespół i organizacja | 8 |
| | kogo trzeba mieć; 1,6–2,7 mln zł rocznie | |
| 4 | Krok 2. Dane treningowe | 10 |
| | skąd wziąć tekst i gdzie kończy się polszczyzna | |
| 5 | Krok 3. Tokenizer | 13 |
| | dlaczego polski kosztuje o jedną czwartą więcej | |
| 6 | Krok 4. Architektura modelu | 16 |
| | 555 linii kodu, reszta to inżynieria uruchomienia | |
| 7 | Krok 5. Infrastruktura obliczeniowa | 17 |
| | ceny kart, rachunek klastra i moc dostępna bezpłatnie | |
| 8 | Krok 6. Pretrening i jego matematyka | 21 |
| | wzór 6ND — policz koszt dowolnego modelu sam | |
| 9 | Krok 7. Post-trening, czyli jak powstaje asystent | 25 |
| | tu powstaje asystent; decydują dane, nie sprzęt | |
| 10 | Krok 8. Ewaluacja i bezpieczeństwo | 28 |
| | jak mierzyć jakość i czemu cudze wyniki nic nie znaczą | |
| 11 | Krok 9. Wdrożenie i ekonomika działania | 31 |
| | stukrotna różnica kosztu na tym samym sprzęcie | |

RACHUNEK

- | | | |
|----|--|----|
| 12 | Kosztorys w czterech scenariuszach | 34 |
| | 150 zł · 0,7 mln zł · 15–38 mln zł · miliardy | |
| 13 | Harmonogram, ryzyka i najczęstsze przyczyny porażki | 37 |
| | ile to trwa i na czym takie projekty się wykładają | |

PRAWO

- | | | |
|----|---|----|
| 14 | Obowiązki prawne twórcy modelu | 39 |
| | akt o AI po nowelizacji z lipca 2026 i ustawa krajowa | |

DECYZJA

- | | | |
|----|--|----|
| 15 | Drabina opcji: pięć dróg i ich ceny | 46 |
| | od promptu do treningu od zera, z cenami każdej drogi | |
| 16 | Polskie punkty odniesienia: Bielik i PLLuM | 48 |
| | jeden za dwa tysiące złotych, drugi za czterdzieści milionów | |

17 Rekomendacja	50
co zrobić naprawdę i pierwsze trzy ruchy	
• Załącznik A. Przepis: własny model od zera na MacBooku wykonany naprawdę — z wykresem i próbką tekstu	4
• Załącznik B. Przepis: model miliardowy na wynajętym klastrze ten sam łańcuch w skali, za około 1200 zł	4
• Załącznik C. Słownik pojęć pojęcia po polsku, z angielskimi oryginałami	4
• Załącznik D. Źródła i dalsza lektura źródła i uczciwa nota o tym, co w nich najsłabsze	4

CZĘŚĆ I

Punkt wyjścia

Zanim policzymy pieniądze, trzeba dokładnie ustalić, co się właściwie buduje. Słowo „model” oznacza w praktyce trzy zupełnie różne przedsięwzięcia, różniące się kosztem o sześć rzędów wielkości.

ROZDZIAŁ 1 · PUNKT WYJŚCIA

Z czego naprawdę składa się ChatGPT

Pytanie „jak zbudować własne ChatGPT” zawiera ukryte założenie, że ChatGPT jest jedną rzeczą. Nie jest. To siedem warstw, z których każda powstaje inaczej, kosztuje inaczej i wymaga innych kompetencji. Zanim policzymy pieniądze, trzeba wiedzieć, za co się płaci.

Rozróżnienie, od którego wszystko zależy, przebiega między **modelem** a **systemem**. Model to plik — a właściwie zbiór plików — zawierający macierze liczb zmiennoprzecinkowych, zwanych wagami. Dla modelu o siedmiu miliardach parametrów jest to około czternastu gigabajtów w precyzji szesnastobitowej. Ten plik sam z siebie nie robi nic; potrzebuje programu, który wczyta go do pamięci karty graficznej i przepuści przez niego tekst. System to wszystko pozostałe: interfejs, wyszukiwanie w dokumentach, narzędzia, pamięć rozmów, moderacja treści, rozliczanie użycia, monitoring.

Znaczna część tego, co użytkownik przypisuje jakości modelu, pochodzi w rzeczywistości z warstwy systemowej. Gdy asystent podaje aktualny kurs waluty, nie „wie” go — wywołał wyszukiwarkę. Gdy poprawnie liczy, zwykle uruchomił interpreter kodu. Gdy pamięta, o czym była mowa tydzień temu, ktoś zbudował mechanizm zapisu i przywracania kontekstu. To jest istotne dla rachunku, który przedstawia niniejsze opracowanie: *warstwa systemowa bywa tańsza od warstwy modelowej o cztery rzędy wielkości, a odpowiada za znaczną część odczuwanej jakości produktu.*

Siedem warstw

Poniższy podział odpowiada realnej kolejności prac i realnemu podziałowi zespołów w firmach, które takie modele budują.

WARSTWA 1 Korpus Kilkanaście bilionów słów tekstu, zebranych, oczyszczonych i odsianych. tygodnie pracy, tysiące rdzeni	WARSTWA 2 Tokenizer Słownik zamieniający tekst na liczby. Decyduje o koszcie obsługi polszczyzny. godziny, jeden komputer	WARSTWA 3 Architektura Kod sieci neuronowej. Kilkaset linii, publicznie znany schemat. dni pracy, zero sprzętu	WARSTWA 4 Pretrening Miesiące liczenia na tysiącach kart. Tu powstają wagi i tu znika budżet. 80–90 % kosztu całości
WARSTWA 5 Post-trening Zamiana automatu dopisującego tekst w asystenta odpowiadającego na polecenia. praca ludzi, nie sprzętu	WARSTWA 6 Serwowanie Uruchomienie modelu tak, by odpowiadał szybko i wielu osobom naraz. koszt stały, miesięcznie	WARSTWA 7 Produkt Interfejs, narzędzia, wyszukiwanie, pamięć, moderacja, rozliczenia. klasyczna praca programistyczna	WNIOSEK Rozkład kosztu Warstwy 1–5 to model. Warstwy 6–7 to produkt. Większość firm potrzebuje wyłącznie 6–7. tu leży decyzja

Dlaczego warstwa czwarta pochłania prawie wszystko

Pretrening polega na powtarzaniu jednej banalnej czynności: model dostaje fragment tekstu i ma przewidzieć, jaki znak wystąpi jako następny. Za każdym razem, gdy się myli, wszystkie jego wagi przesuwają się o mikroskopijny ułamek w stronę odpowiedzi poprawnej. Nie ma w tym żadnej dodatkowej logiki, żadnej wiedzy wprowadzanej ręcznie, żadnych reguł. Jest wyłącznie ta jedna operacja, powtórzona odpowiednią liczbę razy.

Cała trudność polega na tym, jaka jest ta liczba. Współczesny model trenuje się na kilkunastu bilionach tokenów, a każdy token wymaga przeliczenia wszystkich parametrów sieci — kilkakrotnie, bo osobno w przód i wstecz. Iloczyn tych dwóch wielkości daje liczbę operacji zmiennoprzecinkowych o rzędzie wielkości, który dla modeli czołowych sięga 10^{25} i wyżej. Dla porównania: gdyby jedna operacja trwała sekundę, wykonanie ich po kolei zajęłoby więcej czasu, niż istnieje wszechświat, i to o wiele rzędów wielkości. Dlatego wykonuje się je równolegle, na dziesiątkach tysięcy procesorów graficznych, przez wiele tygodni.

SEDNO SPRAWY

Trening modelu językowego nie jest trudny koncepcyjnie ani programistycznie. Kod referencyjny mieści się w kilkuset liniach i jest publicznie dostępny. Trudność jest wyłącznie ilościowa — i właśnie dlatego przekłada się bezpośrednio na pieniądze. Buduje się nie algorytm, tylko elektrownię.

Co odróżnia model surowy od asystenta

Po zakończeniu pretreningu powstaje obiekt, który potrafi jedynie kontynuować tekst. Zapytany „Jak ugotować rosół?” równie chętnie dopisze kolejne pytania z tej samej listy, co przepis — z jego perspektywy jedno i drugie jest prawdopodobną kontynuacją. Zamiana takiego obiektu w narzędzie, które odpowiada na polecenia, odmawia wykonania poleceń szkodliwych i utrzymuje spójny sposób wypowiedzi, to osobny etap zwany post-treningiem. Wymaga on danych tworzonych przez ludzi: przykładowych rozmów, ocen porównawczych, poprawek. Sprzętowo jest tani, kadrowo bywa kosztowniejszy niż sam trening, ponieważ wymaga anotacji wykonywanej przez specjalistów.

Ta różnica ma znaczenie praktyczne dla każdego, kto rozważa własny model. Etap pierwszy — pretrening — jest zaporowo drogi, ale jego produkt można dziś pobrać za darmo w postaci otwartych wag. Etap drugi

— post-trening — jest w zasięgu małego zespołu i to on decyduje o tym, czy model zachowuje się jak przydatne narzędzie w konkretnej dziedzinie. Znakomita większość projektów określanych jako „własny model AI” to w rzeczywistości własny post-trening na cudzym pretreningu — i nie ma w tym nic wstydliwego, bo tak samo postępują firmy dysponujące budżetami dziewięciocyfrowymi.

ROZDZIAŁ 2 · PUNKT WYJŚCIA

Trzy poziomy własności modelu

Zdanie „to nasz model” bywa prawdziwe na trzy różne sposoby, różniące się kosztem od kilkuset złotych do kilkuset milionów dolarów. Rozdział porządkuje te znaczenia, żeby dalsze rozdziały miały do czego się odnosić.

Poziom I — własna warstwa na cudzym modelu

Wagi należą do dostawcy, cała reszta do nas. Budujemy polecenie systemowe, bazę wiedzy przeszukiwaną przed każdą odpowiedzią, zestaw narzędzi, które model może wywołać, oraz logikę produktową. Model jest wymienialny — można go podmienić w jeden dzień, gdy pojawi się lepszy lub tańszy.

To rozwiązanie bywa lekceważone jako „tylko nakładka”, co jest nieporozumieniem. Wartość takiego systemu nie leży w wagach, tylko w danych, do których ma dostęp, i w regułach, według których działa. Kancelaria, która zbudowała uporządkowaną bazę orzecznictwa i sprawdzony sposób zadawania pytań, ma aktyw, którego konkurent nie odtworzy przez podmianę dostawcy — bo brakuje mu bazy, nie modelu.

Poziom II — własne wagi z cudzego punktu startowego

Pobieramy publicznie dostępne wagi modelu otwartego, dokładamy własne dane i uruchamiamy dalszy trening. Powstaje plik z wagami, który fizycznie należy do nas, można go trzymać na własnym serwerze, uruchamiać bez połączenia z internetem i wykorzystywać komercyjnie w granicach licencji modelu bazowego.

Wewnątrz tego poziomu mieszczą się rozwiązania różniące się kosztem stukrotnie. Dostrojenie metodą niskiego rzędu, w której modyfikuje się ułamek procenta wag, wykonuje się na jednej wynajętej karcie w kilka godzin. Adaptacja językowa, polegająca na dołożeniu kilkudziesięciu miliardów tokenów polskiego tekstu do modelu wytrenowanego głównie po angielsku, wymaga już klastra i budżetu rzędu setek tysięcy złotych. Wspólne dla obu jest to, że nie płaci się za odtworzenie wiedzy ogólnej — ona jest już w wagach.

TU LEŻY REALNA GRANICA OPŁACALNOŚCI

Praktycznie wszystkie modele branżowe wdrażane dziś w przedsiębiorstwach — także te reklamowane jako autorskie — powstają na tym poziomie. Powód jest ekonomiczny, nie ambicjonalny: pretrening odtwarza wiedzę, którą ktoś już opłacił i udostępnił za darmo.

Poziom III — własne wagi od zera

Zaczynamy od losowych liczb i budujemy wszystko: korpus, tokenizer, architekturę, pełny pretrening, post-trening. Nie zaciągamy niczyich wag, nie podlegamy niczyjej licencji, a pochodzenie każdego bitu w modelu

jest nam znane — co ma znaczenie przy obowiązkach dokumentacyjnych i przy rozstrzyganiu, czyje utwory znalazły się w danych treningowych.

Ten poziom rozciąga się jednak przez sześć rzędów wielkości kosztu i to jest najważniejsze spostrzeżenie całego opracowania. Model o dwudziestu milionach parametrów, wytrenowany na polskiej Wikipedii, jest w pełni „od zera” i kosztuje mniej niż kolacja we dwoje. Model klasy tych, które napędzają dzisiejsze usługi komercyjne, też jest „od zera” — i kosztuje tyle, ile budowa fabryki. Oba zdania są prawdziwe; różnica polega wyłącznie na skali, a skala jest tu wszystkim.

TABELA 2.1 — TRZY POZIOMY WŁASNOŚCI W ZESTAWIENIU

KRYTERIUM	POZIOM I WARSTWA WŁASNA	POZIOM II DOSTROJENIE WAG	POZIOM III TRENING OD ZERA
Czyje są wagi	dostawcy	nasze, pochodne	nasze w całości
Działanie bez internetu	nie	tak	tak
Kontrola nad danymi treningowymi	żadna	częściowa	pełna
Czas do pierwszego wyniku	godziny	dni	miesiące
Kompetencje zespołu	programista	inżynier uczenia maszynowego	zespół badawczy
Ryzyko techniczne	niskie	umiarkowane	wysokie
Rząd kosztu	setki zł miesięcznie	tysiące–setki tys. zł	od tysiąca zł do setek mln USD

Zestawienie własne. Widelki kosztowe poziomu III wynikają ze skali modelu i zostały rozpisane w rozdziale 12.

Czwarte znaczenie, którym warto się nie posługiwać

W obiegu funkcjonuje jeszcze jedno rozumienie własnego modelu: pobranie gotowych otwartych wag, uruchomienie ich na własnym serwerze i nadanie im własnej nazwy. Technicznie nie jest to nic więcej niż zmiana pliku konfiguracyjnego, a przedstawianie takiego rozwiązania jako autorskiego modelu bywa kwestionowane — zarówno przez warunki licencji, które często wymagają zachowania oznaczenia pochodzenia, jak i przez przepisy o nieuczciwych praktykach rynkowych. Warto o tym pamiętać, ponieważ w komunikacji marketingowej granica bywa zacierana, a odbiorca — zwłaszcza instytucjonalny — coraz częściej pyta wprost o pochodzenie wag.

Budowa krok po kroku

Dziewięć etapów, przez które przechodzi każdy model językowy — od zebrania tekstu po uruchomienie usługi. Przy każdym: co się robi, jakimi narzędziami, jak długo i za ile.

ROZDZIAŁ 3 · KROK 1

Zespół, który to buduje

Trening modelu jest przedsięwzięciem inżynierskim, nie badawczym — co oznacza, że o powodzeniu decyduje nie geniusz, tylko obsada kilku bardzo konkretnych stanowisk. Rozdział opisuje, kogo trzeba mieć, ilu ludzi i za ile.

Powszechne wyobrażenie o twórcach modeli językowych — kilkoro badaczy odkrywających nowe algorytmy — jest mylące. Architektura jest publicznie znana i od kilku lat zmienia się w szczegółach, nie w istocie. Codzienna praca polega na czymś innym: na utrzymaniu przy życiu klastra tysięcy kart, na diagnozowaniu, dlaczego przepustowość spadła o piętnaście procent, na czyszczeniu danych i na tym, żeby trwający sześć tygodni proces nie przerwał się w czwartym z powodu wadliwej kości pamięci w jednym węźle.

Role, bez których się nie da

Rdzeń zespołu, który jest w stanie doprowadzić do końca trening modelu średniej wielkości, liczy od sześciu do dwunastu osób. Poniżej role wraz z tym, co realnie robią, oraz orientacyjnym rocznym kosztem zatrudnienia w Polsce — wynagrodzenie brutto powiększone o narzuty pracodawcy albo równoważna stawka współpracy.

TABELA 3.1 — OBSADA RDZENIA ZESPOŁU I ROCZNY KOSZT

ROLA	ZA CO ODPOWIADA	OSÓB	KOSZT ROCZNY (PLN)
Inżynier danych	Pozyskanie korpusu, ekstrakcja tekstu, filtrowanie, odsiewanie duplikatów, dokumentacja pochodzenia danych.	2	340–520 tys.
Inżynier treningu	Konfiguracja równoległości, wydajność, punkty kontrolne, diagnostyka rozbieżności funkcji straty.	2–3	420–780 tys.
Inżynier infrastruktury	Klaster, sieć, magazyn danych, kolejkovanie zadań, monitoring, obsługa awarii sprzętu.	1–2	220–440 tys.
Specjalista post-treningu	Zbiory instrukcji, uczenie z informacji zwrotnej, dobór sposobu wypowiedzi modelu.	1–2	210–420 tys.
Specjalista ewaluacji	Zestawy testowe, odkażanie zbiorów, pomiary regresji, testy w języku polskim.	1	180–260 tys.
Inżynier wdrożeniowy	Serwowanie modelu, kwantyzacja, wydajność odpowiedzi, koszt jednostkowy.	1	190–290 tys.
Razem rdzeń	Minimalna obsada zdolna dowieźć model bez podwykonawców.	8–11	1,56–2,71 mln

Widelki własne, oparte na stawkach rynkowych w Polsce w 2026 r. dla stanowisk uczenia maszynowego, powiększone o koszty pracodawcy. Zatrudnienie za granicą lub w ośrodkach o wysokiej konkurencji o specjalistów podnosi te kwoty dwu- do czterokrotnie.

NAJCZĘSTSZY BŁĄD W BUDŻETOWANIU

Zespół pracuje przez cały rok, a klaster wynajmuje się na dwa miesiące. W projektach mniejszych niż największe koszt osobowy zwykle przewyższa koszt obliczeń — czasem kilkukrotnie. Rachunek sprowadzany wyłącznie do ceny kart obliczeniowych jest z reguły zaniżony o połowę.

Czego nie trzeba mieć

Nie trzeba badaczy publikujących na czołowych konferencjach, o ile celem nie jest przesunięcie granicy dziedziny, tylko zbudowanie działającego modelu. Nie trzeba autorskiej architektury — odstępstwa od sprawdzonych rozwiązań są w tym przedsięwzięciu głównym źródłem ryzyka, nie przewagi. Nie trzeba własnej biblioteki treningowej; istnieją dojrzałe rozwiązania otwarte, a pisanie własnej to typowy sposób na stratę pół roku.

Trzeba natomiast czegoś, co bywa niedoceniane: kogoś, kto już raz przeprowadził trening w podobnej skali. Różnica między zespołem, który to zrobił, a zespołem, który tylko czytał, mierzy się nie w jakości modelu, lecz w liczbie miesięcy potrzebnych na dojście do pierwszego udanego przebiegu. Rekrutacja takiej osoby jest zwykle najtrudniejszym elementem całego przedsięwzięcia i najbardziej opłacalnym.

Struktura pracy w czasie

Praca nie rozkłada się równomiernie. Pierwsze miesiące to niemal wyłącznie dane i infrastruktura, przy niemal zerowym zużyciu mocy obliczeniowej. Faza treningu właściwego jest dla zespołu okresem stosunkowo spokojnym — dyżury, obserwacja wykresów, reagowanie na awarie — za to najdroższym rachunkowo. Po niej następuje najbardziej pracochłonny etap post-treningu i ewaluacji, gdzie znów decyduje praca ludzi.

Dane i infrastruktura		3–4 mies.
Próby w małej skali		1–2 mies.
Pretrening właściwy		1,5–6 mies.
Post-trening		2–3 mies.
Ewaluacja i bezpieczeństwo		1,5–2 mies.
Wdrożenie		1 mies.

Etapy nakładają się na siebie, więc suma czasu kalendarzowego jest krótsza niż suma powyższych okresów — realnie od dziesięciu do osiemnastu miesięcy dla pierwszego modelu zespołu, który wcześniej tego nie robił, i od sześciu do dziewięciu dla kolejnego.

ROZDZIAŁ 4 · KROK 2

Dane treningowe

Model jest funkcją swoich danych — i to jedyny etap, na którym można zbudować przewagę nieodtwarzalną przez konkurenta z większym budżetem. Rozdział mówi, ile tekstu trzeba, skąd się go bierze, ile kosztuje jego przygotowanie i gdzie kończy się polszczyzna.

Ile tekstu potrzeba

Modele wydane w latach 2024–2026 trenowano na zbiorach od kilkunastu do kilkudziesięciu bilionów tokenów. Llama 3.1 — 15,6 biliona. DeepSeek-V3 — 14,8 biliona. Qwen3 — około trzydziestu sześciu bilionów, obejmujących sto dziewiętnaście języków. Rodzina Llama 4 — ponad trzydzieści bilionów przy pokryciu dwustu języków. Zbiory rosną 3,7-krotnie rocznie, czyli podwajają się co pół roku.

Relacja między wielkością modelu a ilością danych przesunęła się radykalnie. W 2022 roku na parametr przypadało około dziesięciu tokenów, w 2025 — około trzystu. Ten stosunek jest dziś użytecznym punktem odniesienia: *trzysta tokenów na każdy parametr*. Model o miliardzie parametrów wymaga zatem trzystu miliardów tokenów, żeby jego potencjał został wykorzystany.

Skąd bierze się ten tekst

Podstawą pozostaje publiczne archiwum internetu gromadzone przez organizację Common Crawl. Pojedyncza miesięczna migawka obejmuje dwa do dwóch i pół miliarda stron i waży od 250 do 400 terabajtów przed kompresją. Dane leżą w otwartym zasobie chmurowym i są bezpłatne — płaci się wyłącznie za ich przetworzenie.

Z tego surowca zbudowano zbiory pochodne, po które sięga dziś większość zespołów, bo dźwigają one ciężar czyszczenia. Najważniejsze zestawiono poniżej.

TABELA 4.1 — PUBLICZNIE DOSTĘPNE KORPUSY, PO KTÓRE SIĘGA SIĘ W PIERWSZEJ KOLEJNOŚCI

ZBIÓR	OBJĘTOŚĆ	ZAWARTOŚĆ I PRZEZNACZENIE
FineWeb	15 bln tokenów 44 TB	Dziewięćdziesiąt sześć migawek archiwum z lat 2013–2024, oczyszczonych i odsianych. Domyślny wybór dla języka angielskiego.
FineWeb-Edu	1,3–5,4 bln	Podzbiór o wysokiej wartości edukacyjnej, wybrany klasyfikatorem. Krótszy trening, lepsze wyniki na zadaniach wiedzy.
FineWeb2	≈ 3 bln słów 8 TB	Wersja wielojęzyczna: 1893 pary język–pismo. Tu znajduje się polszczyzna.
RedPajama-V2	30 bln po odsianiu	Osiemdziesiąt cztery migawki z ponad czterdziestoma gotowymi wskaźnikami jakości do własnego filtrowania.
Dolma	3 bln tokenów	Jawny skład: 2180 mld z archiwum sieci, 342 mld z repozytoriów kodu, 80 mld z forów, 57 mld z prac naukowych, 5,2 mld z książek.
The Stack v2	67,5 TB	Kod źródłowy w ponad sześciuset językach programowania, zbudowany na archiwum Software Heritage.
Realny sufit	18–22 bln	Tyle unikatowych tokenów daje się dziś zebrać legalnie i bezpłatnie. Powyżej trzeba dokładać powtórzenia, syntetykę i dane niepubliczne.

Potok przetwarzania i jego niespodzianki

Droga od surowego archiwum do korpusu treningowego ma pięć etapów, a każdy odsiewa znaczną część materiału. Warto prześledzić ją na przykładzie zbioru FineWeb, którego autorzy opublikowali pełne rozliczenie.

Po ekstrakcji i filtrach bazowych		36 bln
Po odsianiu duplikatów		20 bln
Po filtrach heurystycznych		15 bln
Po filtrze jakości edukacyjnej		5,4 bln
Przy najwyższym progu jakości		1,3 bln

Pierwsza niespodzianka dotyczy ekstrakcji. Archiwum udostępnia gotowe pliki z wyodrębnionym tekstem, ale przetwarzanie surowych plików źródłowych własnym narzędziem daje mierzalnie lepsze modele. Drobiazg techniczny, który przekłada się na jakość gotowego produktu.

Druga jest poważniejsza i przeczy intuicji. Odsiewanie duplikatów prowadzone *w obrębie* każdej migawki usuwa około czterdziestu pięciu procent materiału i poprawia model. Odsiewanie prowadzone *globalnie*, między wszystkimi migawkami, usuwa nawet dziewięćdziesiąt cztery procent najstarszych zrzutów i daje model **gorszy**. Powód jest taki, że materiał wartościowy bywa powielany przez lata, a materiał unikatowy to często reklamy i szum. Agresywniejsze czyszczenie usuwa więc dokładnie to, co chciało się zachować.

GDZIE NAPRAWDĘ LEŻĄ KOSZTY PRZYGOTOWANIA DANYCH

Zastosowanie klasyfikatora jakości do piętnastu bilionów tokenów kosztowało sześć tysięcy godzin pracy kart, czyli około piętnastu tysięcy dolarów. Ustalenie, *jak* filtrować — ponad siedemdziesiąt modeli testowych sprawdzających kolejne decyzje — pochłonęło osiemdziesiąt tysięcy godzin, czyli od stu dwunastu do dwustu pięćdziesięciu tysięcy dolarów. Rozstrzygnięcie metody kosztowało dwadzieścia razy więcej niż jej wykonanie.

Warto odnotować zmianę, która przewróciła dawne kalkulacje. Odsiewanie duplikatów było kiedyś zadaniem dla procesorów głównych i zajmowało tygodnie na klastrze kilkudziesięciordzeniowym. Wykonane na kartach graficznych trwa godziny: prawie dwa biliony tokenów przetwarza się w pół godziny na trzydziestu dwóch kartach, co kosztuje kilkadziesiąt dolarów. Wąskim gardłem przestała być moc obliczeniowa, a stała się pamięć i przepustowość odczytu.

Dane syntetyczne

Generowanie tekstu treningowego przez inny model jest dziś standardowym uzupełnieniem korpusu — zbiór Nemotron-CC zawiera 1,9 biliona tokenów wygenerowanych w ten sposób. Koszt daje się policzyć: wytworzenie miliarda tokenów zajmuje około czterystu godzin pracy karty, czyli około dolara za milion tokenów. Bilion tokenów syntetycznych to zatem wydatek rzędu miliona dolarów na własnym sprzęcie — wobec dwóch do sześćdziesięciu milionów, gdyby zamawiać je przez komercyjne udostępnienie cudzego modelu, zależnie od klasy użytego modelu. Ta różnica, od dwukrotnej do sześćdziesięciokrotnej, wyjaśnia, dlaczego dane pretreningowe generuje się na własnym sprzęcie.

Ile kosztuje kupienie danych

Umowy licencyjne z wydawcami stały się elementem krajobrazu. Najwyższa ujawniona to ponad dwieście pięćdziesiąt milionów dolarów na pięć lat za portfolio prasowe jednej grupy wydawniczej. Serwis społecznościowy sprzedawał dostęp do swoich treści za około sześćdziesiąt milionów dolarów rocznie jednemu odbiorcy i, według doniesień, około siedemdziesięciu drugiemu. Wydawca prasowy otrzymuje dwadzieścia do dwudziestu pięciu milionów rocznie. Cały rynek licencjonowania danych treningowych szacuje się jednak na niecałe sześć miliardów dolarów rocznie — czyli mniej, niż kosztuje trening kilku modeli czołowych.

Kluczowa obserwacja brzmi: *te pieniądze nie kupują objętości*. Przy szacunkowej wielkości archiwum dużego dziennika roczna opłata odpowiada rzędowi tysiąca do dwóch i pół tysiąca dolarów za milion tokenów. Ten sam milion tokenów z publicznego korpusu kosztuje ułamek centa. Różnica pięciu rzędów wielkości oznacza, że przedmiotem transakcji jest bezpieczeństwo prawne, świeżość materiału i prawo do cytowania — a nie surowiec.

ILE KOSZTUJE POMINIĘCIE TEJ OSTROŻNOŚCI

Ugoda zawarta w 2025 r. przez jednego z dostawców modeli opiewa na 1,5 miliarda dolarów za około 482 tysiące książek pozyskanych z nielegalnych repozytoriów — mniej więcej trzy tysiące dolarów za tytuł, wraz z obowiązkiem zniszczenia plików. Roszczenia zgłoszono do dziewięćdziesięciu jeden procent objętych nią tytułów. Dla porównania: legalne zakupienie tych samych książek kosztowałoby ułamek tej kwoty.

Polszczyzna i jej twardy sufit

To najważniejsza część rozdziału dla każdego, kto myśli o modelu dla polskiego rynku. Polski stanowi 2,04 procent stron w archiwum internetu i zajmuje jedenaste miejsce wśród języków — przy 40,6 procent dla angielskiego. W wielojęzycznym zbiorze FineWeb2 polski podzbiór liczy 73,1 miliarda słów w 152 milionach dokumentów i zajmuje 432 gigabajty.

Przeliczenie słów na tokeny wymaga ostrożności, bo zależy od tokenizatora i różnice sięgają dwukrotności. Te 73 miliardy słów to od około 118 miliardów tokenów przy słowniku przystosowanym do polszczyzny

do około 235 miliardów przy słowniku zbudowanym dla angielskiego. Zasoby dedykowane — Narodowy Korpus Języka Polskiego z ponad półtora miliarda słów, blisko trzysta gigabajtów tekstów z konsorcjum CLARIN-PL, polskojęzyczna Wikipedia z około 1,70 miliona artykułów — są w tej skali dodatkiem, nie fundamentem.

TABELA 4.2 — ZASÓB POLSKIEGO TEKSTU I JEGO KONSEKWENCJA

WIELKOŚĆ	WARTOŚĆ	CO Z TEGO WYNIKA
Udział polskiego w archiwum internetu	2,04 %	jedenaste miejsce wśród języków
Polski w zbiorze FineWeb2	73,1 mld słów	432 GB, ósme miejsce w zbiorze
Po przeliczeniu na tokeny	118–235 mld	zależnie od tokenizatora
Sufit jakościowy polszczyzny publicznej	200–300 mld	przy luźniejszych filtrach 400–600 mld
Największy model możliwy wyłącznie z polszczyzny	≈ 780 mln parametrów	przy dzisiejszej proporcji 300 tokenów na parametr

Wyliczenie własne na podstawie statystyk Common Crawl, zbioru FineWeb2 oraz raportów technicznych modeli Bielik v3 i PLLuM.

Ostatni wiersz tej tabeli jest sednem sprawy. Całą dostępną, wysokiej jakości polszczyzną można nasycić model o mniej więcej ośmiuset milionach parametrów — czyli około jednej dziesiątej dzisiejszego modelu małego. Nie jest to ograniczenie budżetowe, którego dałoby się uniknąć większymi pieniędzmi; polskiego tekstu po prostu nie ma więcej.

Stąd bierze się rozwiązanie stosowane przez wszystkie poważne polskie projekty: kompetencje ogólne importuje się z modelu wytrenowanego głównie na angielszczyźnie, a polszczyzny model douczają się w kolejnym etapie. Model Bielik v3 trenowano na korpusie 292 miliardów tokenów, z czego 237 miliardów stanowiła polszczyzna, a resztę angielski materiał przeciwdziałający zapominaniu — ale punktem wyjścia nie były losowe wagi, tylko model gotowy. Projekt PLLuM zbudowano na korpusie 140 miliardów tokenów w analogicznym układzie.

WNIOSEK PRAKTYCZNY

Model Bielik v3, zużywając 237 miliardów polskich tokenów, wyczerpał prawdopodobnie większość dostępnego zasobu organicznego. Dalsze skalowanie polszczyzny musi opierać się na materiale, którego nie ma w otwartym internecie: archiwach wydawców, orzecznictwie, dokumentacji urzędowej i danych syntetycznych. *Kto ma legalny dostęp do dużego, uporządkowanego zbioru polskich tekstów specjalistycznych, dysponuje zasobem rzadszym niż moc obliczeniowa.*

ROZDZIAŁ 5 · KROK 3

Tokenizer i cena polszczyzny

Najkrótszy etap całego przedsięwzięcia — kilkanaście sekund pracy jednego komputera — i zarazem ten, którego skutki finansowe odczuwa się przez cały okres życia produktu. Dla polskojęzycznej usługi to rozdział o tym, dlaczego płaci się dwa razy więcej za to samo.

Model nie przetwarza liter. Przetwarza liczby, a zamianą tekstu na liczby zajmuje się tokenizer — słownik kilkudziesięciu do kilkuset tysięcy fragmentów wyrazów wraz z regułą ich składania. Standardem jest kodo-

wanie parami bajtów: tekst zamienia się na bajty, a następnie iteracyjnie łączy najczęściej sąsiadującą parę jednostek, zapisując kolejność łączeń. Trening polega na przejściu po korpusie z licznikiem par i — wbrew intuicji — trwa sekundy, nie godziny.

Rozmiary słowników rosną systematycznie: trzydzieści dwa tysiące w modelu Mistral 7B, 128 tysięcy w Llamie 3, 129 tysięcy w DeepSeek-V3, 152 tysiące w Qwen3, 248 tysięcy w Qwen3.5, 262 tysiące w Gemmie 3. Powód jest podwójny. Większy słownik skraca sekwencję, a koszt mechanizmu uwagi rośnie z kwadratem jej długości. Zarazem jednak tablice słownika zajmują parametry — i tu pojawia się ograniczenie, które dotyczy właśnie modeli małych.

DLACZEGO MAŁY MODEL NIE MOŻE MIEĆ DUŻEGO SŁOWNIKA

Przy słowniku 128 tysięcy tablice osadzeń stanowią 1,0 % parametrów modelu czterystumiliardowego, ale już 13,1 % modelu ośmiomiliardowego. Podniesienie słownika do ćwierć miliona przy zachowaniu reszty architektury oddałoby na same tablice 22,7 % masy modelu. Dlatego małe modele wiążą tablicę wejściową z wyjściową — a i tak model o 800 mln parametrów ze słownikiem 248 tysięcy przeznacza na to 31,8 % swoich parametrów.

Ile kosztuje polszczyzna

Ta sama treść zapisana po polsku zajmuje istotnie więcej tokenów niż po angielsku, ponieważ słowniki budowano na korpusach, w których angielszczyzna dominuje. Skutek jest podwójny: wyższy rachunek za każde wywołanie modelu oraz mniej treści mieszczącej się w tym samym oknie kontekstu.

Pomiar wykonany na równoległych fragmentach polsko-angielskich — obejmujących między innymi klauzulę umowną, fragment rozporządzenia i tekst potoczny — daje następujące wyniki dla trzech pokoleń słowników.

TABELA 5.1 — KOSZT TEJ SAMEJ TREŚCI PO POLSKU I PO ANGIELSKU

SŁOWNIK	ROZMIAR	TOKENÓW NA SŁOWO POLSKI	TOKENÓW NA SŁOWO ANGIELSKI	NARZUT POLSZCZYZNY
o200k (najnowsze modele)	200 019	2,25	1,08	1,55 ×
cl100k (poprzednie pokolenie)	100 277	2,70	1,09	1,84 ×
GPT-2 (2019 r.)	50 257	3,86	1,10	2,62 ×

Pomiar własny biblioteką tiktoken 0.13.0 z 1 sierpnia 2026 r. na pięciu równoległych fragmentach (1133 znaki po polsku, 1198 po angielsku). Rozstrzał między rodzajami tekstu jest istotny — najgorzej wypada język umów, najlepiej proza potoczna.

Niezależny pomiar wykonany na preambule Konstytucji Rzeczypospolitej daje obraz zgodny co do rzędu wielkości i pozwala porównać modele otwarte: 2,81 tokena na słowo dla Llamy 3.2, 2,69 dla Qwen3, 2,36 dla Mistral Small 3.1, 2,20 dla Gemmy 3, 1,88 dla modelu europejskiego EuroLLM i 1,62 dla tokenizatora zbudowanego specjalnie dla polszczyzny — przy słowniku czterokrotnie mniejszym niż u konkurentów.

Sprawdzenie na własnym tokenizatorze

Powyższe pomiary mają jedną słabość, o której trzeba powiedzieć: wykonano je na krótkich, pojedynczych tekstach. Na potrzeby tego opracowania wytrenowano zatem własny tokenizator o słowniku 16 384 na polskim korpusie i zmierzono go na 2,29 miliona słów z czterech różnych rodzajów tekstu. Wynik jest bardziej stonowany, niż sugerują pomiary punktowe — i ciekawszy.

TABELA 5.2 – POMIAR WŁASNY NA 2,29 MLN SŁÓW, TOKENÓW NA SŁOWO

RODZAJ TEKSTU	SŁÓW	SŁOWNIK WŁASNY 16 TYS.	O200K 200 TYS.	CL100K 100 TYS.	PRZEWAGA
Akty normatywne	734 tys.	1,67	2,38	2,69	1,43 ×
Prawo unijne	669 tys.	2,10	2,61	2,98	1,24 ×
Teksty encyklopedyczne	445 tys.	2,04	2,41	2,78	1,18 ×
Proza literacka	443 tys.	2,26	2,37	2,71	1,05 ×
Razem	2 290 tys.	1,98	2,45	2,80	1,24 ×

Pomiar własny z 1 sierpnia 2026 r. Tokenizator trenowany na korpusie złożonym głównie z aktów normatywnych i tekstów encyklopedycznych, słownik 16 384, czas treningu 17 sekund.

WNIOSEK PRECYZYJNIEJSZY NIŻ OBIEGOWY

Własny tokenizator wygrywa, ale przewaga **zależy od dziedziny i wynika z doboru korpusu**, na którym go wytrenowano. Na aktach normatywnych, których w korpusie treningowym było najwięcej, przewaga sięga 1,43 raza. Na prozie literackiej, której było najmniej, spada do 1,05 — czyli praktycznie znika. Twierdzenie „własny słownik jest lepszy” jest zatem prawdziwe wyłącznie dla tekstów podobnych do tych, na których go zbudowano. *Dla usługi operującej na dokumentach prawnych oznacza to jednak realną, trwałą obniżkę rachunku o ponad trzydzieści procent* — przy koszcie siedemnastu sekund obliczeń.

PRZEŁOŻENIE NA RACHUNEK PRODUKTU

Sto tysięcy słów polskiego tekstu to około 225 tysięcy tokenów przy najnowszym słowniku i 270 tysięcy przy poprzednim. Przy cenniku rzędu trzech dolarów za milion tokenów wejściowych daje to 0,68 albo 0,81 dolara za ten sam materiał — wobec około 0,33 dolara po angielsku. To samo okno kontekstu mieści przy tym o połowę mniej polskiej treści. Każdy budżet tokenowy produktu operującego na polskich aktach, umowach czy fakturach należy powiększyć o ten narzut — i uwzględnić go nie tylko w cenniku, ale i w limitach kontekstu.

Dla przedsięwzięcia budującego własny model jest to zarazem jedna z niewielu okazji do uzyskania realnej przewagi małym nakładem. Wytrenowanie własnego tokenizatora na polskim korpusie kosztuje kilkanaście sekund pracy jednego komputera, a w dziedzinie zgodnej z korpusem obniża koszt obsługi każdego zapytania o ponad trzydzieści procent na cały okres życia modelu. Jest to prawdopodobnie najwyższa stopa zwrotu w całym łańcuchu opisanym w tym opracowaniu.

```
# Trening własnego tokenizatora na polskim korpusie – całość roboty
from tokenizers import ByteLevelBPETokenizer

tok = ByteLevelBPETokenizer()
tok.train(files=["korpus_pl.txt"], vocab_size=32768, min_frequency=2,
          special_tokens=["<|endoftext|>", "<|user|>", "<|assistant|>"])
tok.save_model("tokenizer_pl")

# Sprawdzenie wydajności: ile tokenów na słowo
tekst = open("probka_pl.txt").read()
print(len(tok.encode(tekst).ids) / len(tekst.split()))
```

Architektura modelu

Najbardziej przeceniany etap całego przedsięwzięcia. Schemat jest publicznie znany, kod referencyjny ma pięćset pięćdziesiąt pięć linii, a odstępstwa od sprawdzonych rozwiązań są głównym źródłem ryzyka, nie przewagi.

Współczesny model to stos identycznych bloków, z których każdy wykonuje dwie operacje: mechanizm uwagi, pozwalający każdemu tokenowi odnieść się do pozostałych, oraz sieć w przód, przetwarzającą każdą pozycję osobno. Postać kanoniczna z 2026 roku obejmuje normalizację przed każdym podblokiem zamiast po nim, uproszczoną normalizację rezygnującą z odejmowania średniej, oraz bramkowaną funkcję aktywacji o trzech macierzach zamiast dwóch. Te trzy rozstrzygnięcia są dziś powszechne i nie ma powodu ich kwestionować.

Rachunek, który warto umieć wykonać

Liczbę parametrów modelu daje się policzyć z konfiguracji i warto to zrobić, bo natychmiast widać, gdzie idą pieniądze. Dla modelu Llama 3 o ośmiu miliardach parametrów — trzydzieści dwie warstwy, wymiar 4096, wymiar sieci w przód 14 336, słownik 128 256 — rachunek daje 218,1 miliona parametrów na warstwę, blisko siedem miliardów w korpusie modelu i 1,05 miliarda w tablicach słownika, razem 8,03 miliarda. Zgadza się z deklaracją producenta co do trzeciej cyfry znaczącej. Ten sam wzór odtwarza 70,55 miliarda dla wariantu średniego i 405,85 miliarda dla największego.

TABELA 6.1 — KONFIGURACJE MODELI, NA KTÓRYCH WARTO SIĘ WZOROWAĆ

MODEL	WARSTWY	WYMIAR	GŁOWY Q / KV	SŁOWNIK	PARAMETRY
Llama 3 8B	32	4 096	32 / 8	128 256	8,03 mld
Llama 3 70B	80	8 192	64 / 8	128 256	70,55 mld
Llama 3 405B	126	16 384	128 / 8	128 256	405,85 mld
Qwen3 32B	64	5 120	64 / 8	151 936	32 mld
DeepSeek-V3	61	7 168	128 / uwaga utajona	129 280	671 mld / 37 aktywnych

Pamięć podręczna uwagi decyduje o długim kontekście

Przy generowaniu odpowiedzi model przechowuje stan rozmowy w buforze, którego rozmiar rośnie liniowo z długością kontekstu. To on, a nie liczba parametrów, ogranicza obsługę długich dokumentów. Rachunek jest bezlitosny: model czterystumiliardowy w klasycznej postaci mechanizmu uwagi potrzebowałby 8,06 megabajta na token, czyli 1082 gigabajtów przy kontekście stu dwudziestu ośmiu tysięcy tokenów — wartość absurdalna.

Stąd warianty oszczędne. Zgrupowanie głów mechanizmu uwagi, dziś standardowe, sprowadza tę wartość do 504 kilobajtów na token, czyli 67,7 gigabajta przy pełnym kontekście. Rozwiązanie zastosowane w modelu DeepSeek, polegające na przechowywaniu skompresowanego wektora zamiast pełnych kluczy i wartości, schodzi do 68,6 kilobajta na token — siedmiokrotnie mniej niż u konkurenta, przy modelu większym.

KIEDY UWAGA KOSZTUJE WIĘCEJ NIŻ CAŁA RESZTA MODELU

Koszt mechanizmu uwagi rośnie z kwadratem długości sekwencji, a koszt warstw gęstych — liniowo. Punkt zrównania wypada przy 61 tysiącach tokenów dla modelu ośmiomiliardowego, 108 tysiącach dla siedemdziesięciomiliardowego i 197 tysiącach dla czterystumiliardowego. Oznacza to, że przy deklarowanym oknie 128 tysięcy tokenów mniejsze modele wydają na samą uwagę więcej niż na wszystkie mnożenia macierzy wag. To arytmetyczne uzasadnienie okien przesuwanych i uwagi rzadkiej — nie moda, tylko rachunek.

Mieszanka ekspertów i rosnąca rzadkość

Architektura rozdzielająca sieć w przód na wyspecjalizowane fragmenty, z których przy każdym tokenie uruchamia się tylko kilka, stała się po 2024 roku rozwiązaniem domyślnym dla dużych modeli. Stosunek parametrów całkowitych do aktywnych rośnie systematycznie: 10,7-krotny w Qwen3, 18,1-krotny w DeepSeek-V3, 22,9-krotny w otwartym modelu firmy OpenAI, 23,5-krotny w Llamie 4, a w najnowszych konstrukcjach z 2026 roku przekracza trzydziestokrotność.

Praktyczny wniosek dla planującego: koszt obliczeń rośnie z parametrami aktywnymi, a koszt pamięci z całkowitymi. Mieszanka ekspertów przenosi zatem wąskie gardło z mocy obliczeniowej na pamięć kart i przepustowość sieci między węzłami — co zmienia dobór sprzętu, ale nie zmienia faktu, że przy porównywalnej jakości trening wychodzi kilkakrotnie taniej.

Ile to naprawdę linii kodu

Warto zderzyć wyobrażenie z rzeczywistością. Referencyjna implementacja modelu w projekcie nanochat, zawierająca komplet współczesnych rozwiązań, ma 555 linii. Cały projekt wraz z tokenizatorem, treningiem, dostrajaniem, ewaluacją i interfejsem to 7775 linii w trzydziestu sześciu plikach, a wzorcowy przebieg kosztuje około czterdziestu ośmiu dolarów na węźle z ośmioma kartami. Poprzednik, nanoGPT, to 330 linii samego modelu.

Skala produkcyjna wygląda zupełnie inaczej. Biblioteka torchtitan, rozwijana przy projekcie PyTorch, to ponad sto trzydzieści tysięcy linii i zawiera gotowe definicje wszystkich liczących się modeli. Biblioteka Megatron-LM firmy NVIDIA to ponad czterysta trzydzieści tysięcy linii. Różnica nie polega na architekturze — ta jest w obu przypadkach ta sama. Polega na równoległości, punktach kontrolnych, obsłudze niższej precyzji i odporności na awarie węzłów.

SEDNO ROZDZIAŁU

Architektura to jeden plik, który da się przeczytać w godzinę. Wszystko pozostałe — czyli 98 % kodu, całe ryzyko i cały budżet — to inżynieria uruchomienia tego pliku na tysiącu kart naraz.

ROZDZIAŁ 7 · KROK 5

Infrastruktura obliczeniowa

Rozdział o tym, co realnie kupuje się albo wynajmuje, ile to kosztuje w sierpniu 2026 roku i gdzie leży próg opłacalności posiadania. Kończy się informacją, która dla polskiej firmy jest

prawdopodobnie najważniejsza w całym opracowaniu — o mocy obliczeniowej dostępnej bezpłatnie.

Ceny sprzętu

Rynek rozwarstwił się na trzy półki. Poprzednia generacja jest już towarem schodzącym: rynek wtórny układów H100 mieści się w przedziale 6–22 tysięcy dolarów, czyli od czterdziestu pięciu do osiemdziesięciu pięciu procent poniżej szczytu z 2023 roku, gdy za tę samą kartę płacono około czterdziestu tysięcy. Cena nowego modułu nie jest publikowana przez kanał sprzedaży i szacunki rozchodzą się szeroko — pewna jest wyłącznie relacja: nowy H100 kosztuje dziś istotnie mniej niż bieżąca generacja. To ostrzeżenie dla każdego, kto rozważa zakup: *sprzęt traci wartość szybciej niż większość aktywów przemysłowych*.

Generacja bieżąca kosztuje 40–80 tysięcy dolarów za układ w zależności od kanału zakupu, a kompletny serwer ośmiokartowy 400–500 tysięcy. Wersja wzmocniona to 80–110 tysięcy za układ i od 750 tysięcy do miliona za węzeł. Szafa zintegrowana mieszcząca siedemdziesiąt dwie karty wyceniana jest na dwa do trzech milionów dolarów, co paradoksalnie daje niższą cenę jednostkową niż serwery pojedyncze, ponieważ w cenie szafy amortyzuje się połączenia wewnętrzne i zasilanie.

Wynajem — trzy zupełnie różne rynki

Rozpiętość cen tego samego układu sięga sześciokrotności i nie wynika ze sprzętu, tylko z modelu obsługi. Warto to rozumieć, bo porównywanie ofert bez tego rozróżnienia prowadzi do wniosków fałszywych.

TABELA 7.1 — CENA GODZINY PRACY KARTY, SIERPIEŃ 2026 R. (USD)

RYNEK	H100	H200	B200	CHARAKTER OFERTY
GIEŁDY MOCY — NAJTANIEJ, BEZ GWARANCJI				
Vast.ai	1,49–1,92	—	—	wyłącznie oferty przerywalne
RunPod	1,99–2,99	3,59–4,39	5,89	zależnie od poziomu izolacji
CHMURY WYSPECJALIZOWANE — STANDARD DLA TRENINGU				
Lambda	3,99	—	6,69	klastry rezerwowane drożej
Nebius	3,85	4,50	7,15	tryb przerywalny o 45 % taniej
CoreWeave	6,16	6,31	8,60	pełna obsługa klastra
WIELCY DOSTAWCY CHMUROWI — NAJDROŻEJ				
AWS	6,88	—	14,24	ceny podniesione w 2026 r.
Azure	12,29	13,78	—	rozliczenie za cały węzeł
Google Cloud	—	10,85	—	zależnie od regionu

Ceny na żądanie w przeliczeniu na jedną kartę, odczyt sierpień 2026 r. Warto odnotować kierunek zmian: w styczniu 2026 r. jeden z dużych dostawców podniósł ceny wybranych instancji o 15 %, a od lipca o 20 % podrożały bloki mocy rezerwowanej. Cena wynajmu nie jest trwałą kotwicą, jeśli nie zamrozi jej umowa.

Osobno warto rozważyć sprzęt firmy AMD, który wypada o trzydzieści do pięćdziesięciu procent taniej za godzinę pracy przy porównywalnej klasie pamięci. Ceną tej oszczędności jest przeniesienie całego oprogramowania na inny stos technologiczny — koszt jednorazowy, ale realny i trudny do oszacowania z góry.

Czego nie widać w cenie karty

Sieć wewnętrzna klastra to od dziesięciu do dwudziestu procent jego wartości. Pojedynczy przełącznik wysokiej przepustowości kosztuje 80–140 tysięcy dolarów, karta sieciowa na serwer 2,5–4 tysiące, a gotowa sieć dla szesnastu węzłów 300–500 tysięcy dolarów. W skali tysiąca kart daje to od dziewięciu do trzynastu milionów dolarów za sieć wyspecjalizowaną, wobec około siedmiu milionów za ethernet wysokiej klasy — przy około piętnastoprocentowej przewadze wydajnościowej tej pierwszej.

Wymagania wobec pamięci masowej wyznacza nie zbiór treningowy, lecz zapis stanu treningu. Warto różnić dwie wielkości: stan trzymany w pamięci kart to około szesnastu bajtów na parametr, a zapisywany punkt kontrolny dwanaście do czternastu, bo gradientów się nie utrwała. Punkt kontrolny modelu siedemdziesięcimiliardowego zajmuje zatem 0,84–0,98 terabajta, a czterystumiliardowego 4,8–5,6 terabajta. Zapisanie pierwszej z tych wartości w trzydzieści sekund wymaga przepustowości rzędu trzydziestu gigabajtów na sekundę, co mieści się w możliwościach współczesnych systemów plików, ale wymaga sprzętu z wyższej półki.

BARIERA, O KTÓREJ MAŁO KTO PAMIĘTA

Szafa najnowszej generacji pobiera 132 kilowaty i wymaga chłodzenia cieczą — chłodzenie powietrzem jest fizycznie niewystarczające. To skok trzy- do czterokrotny w jednej generacji. Według badań branżowych 68 % serwerowni korporacyjnych zbudowanych przed 2015 r. nie ma ani gęstości mocy, ani chłodzenia zdolnego obsłużyć takie obciążenie, a 82 % z nich ma jeszcze co najmniej dziesięć lat najmu. Instalacja chłodzenia cieczą kosztuje 3,5–5 tys. dolarów na kilowat.

Energia różnicuje wynik w Europie bardziej niż sprzęt. Realistyczny koszt dużego odbiorcy w Polsce w 2026 roku mieści się w przedziale 620–750 złotych za megawatogodzinę wraz ze składnikami regulowanymi, wobec około połowy tej kwoty w Finlandii i Szwecji. Warto odnotować szczególnie istotny rachunkowo: płaski profil poboru, naturalny dla centrum obliczeniowego pracującego bez przerwy, obniża opłatę mocową o około sto dziesięć złotych na megawatogodzinę względem profilu szczytowego.

Kupić czy wynająć — pełny rachunek

Poniżej rozliczenie klastra tysiąckartowego przez trzy lata, przy sprzęcie bieżącej generacji i polskich kosztach energii.

TABELA 7.2 — KLASTR 1024 KART, TRZY LATA, KOSZT POSIADANIA

POZYCJA	KWOTA (MLN USD)	UWAGI
128 węzłów ośmiokartowych	61,4	480 tys. USD za węzeł
Sieć wewnętrzna klastra	12,0	przełączniki, karty, optyka
Pamięć masowa 3 PB	4,0	nośniki półprzewodnikowe
Szafy, zasilanie, części zamienne	1,5	—
Nakład inwestycyjny	78,9	≈ 295 mln PLN
Energia rocznie	2,85	15 692 MWh przy PUE 1,25
Kolokacja rocznie	2,33	powierzchnia i chłodzenie
Serwis producenta rocznie	3,87	5 % wartości sprzętu

POZYCJA	KWOTA (MLN USD)	UWAGI
Zespół platformowy rocznie	0,86	osiem osób w Polsce
Koszt trzyletni po odjęciu wartości rezydualnej	96,4	≈ 361 mln PLN

Wartość rezydualna przyjęta jako 20 % wartości samych kart po trzech latach, czyli 12,3 mln USD — co odpowiada 15,6 % całego nakładu. Kurs 3,7425 PLN za dolara z 31 lipca 2026 r.

Przeliczone na godzinę pracy karty daje to 3,58 dolara przy pełnym wykorzystaniu, 3,98 przy dostępności dziewięćdziesięcioprocentowej i 5,12 przy realistycznym wykorzystaniu siedemdziesięcioprocentowym. Sama energia to zaledwie trzydzieści siedem centów — *dominuje amortyzacja, nie prąd*, co bywa zaskoczeniem przy publicznej debacie skupionej na zużyciu energii.

Próg opłacalności zakupu zależy więc wprost od tego, z czym się go porównuje. Wobec cen na żądanie u dostawców wyspecjalizowanych wystarczy wykorzystanie na poziomie pięćdziesięciu procent, wobec największych dostawców chmurowych — dwudziestu pięciu. Ale wobec kontraktu wieloletniego zawartego po stawce hurtowej zakup nie zwraca się nigdy. Wniosek: *posiadanie bije ceny katalogowe dwu- do czterokrotnie, ale przegrywa z głębokim zobowiązaniem terminowym*, wiążąc przy tym kapitał i przenosząc ryzyko starzenia się sprzętu na kupującego.

Moc obliczeniowa, za którą nie trzeba płacić

Zanim polska firma wyda własny kapitał, powinna wyczerpać ścieżki publiczne. Jest ich więcej, niż się powszechnie sądzi, a warunki są korzystniejsze, niż większość przedsiębiorców zakłada.

TABELA 7.3 — PUBLICZNE ZASOBY OBLICZENIOWE DOSTĘPNE Z POLSKI

ZASÓB	CZYM DYSONUJE	JAK UZYSKAĆ DOSTĘP	KOSZT
Helios, Cyfronet AGH	440 superukładów GH200, 75 tys. rdzeni, partycja interaktywna	granty obliczeniowe w portalu PLGrid	bezpłatnie dla badań
Gaia, Kraków	ok. 1000 kart, partycje treningowa i wnioskowania	fabryka AI, uruchomiona w maju 2026 r.	tryby preferencyjne
PIAST-AI, Poznań	ponad 1000 kart	fabryka AI, pełnia usług od III kw. 2026 r.	tryby preferencyjne
EuroHPC — Fast Lane	do 50 tys. godzin kart	wniosek rozpatrywany w cztery dni robocze	bezpłatnie dla MŚP
EuroHPC — Playground	zasób na próby i rozeznanie	dostęp w dwa dni robocze	bezpłatnie
EuroHPC — Extreme Scale	LUMI, Leonardo, MareNostrum5, JUPITER	terminy: 4 maja i 19 października 2026 r.	bezpłatnie po ocenie

Ścieżki dostępu obejmują tor naukowy, przemysłowy i dla administracji publicznej. Dla małych i średnich przedsiębiorstw, w tym spółek na wczesnym etapie, tryby innowacyjne są bezpłatne.

WNIOSEK OPERACYJNY

Ścieżka „Fast Lane” daje do pięćdziesięciu tysięcy godzin pracy kart w cztery dni robocze i bezpłatnie. Na rynku komercyjnym ten sam zasób kosztowałby od stu do trzystu tysięcy dolarów. To wystarczy na pełne dostrojenie modelu kilkudziesięciomiliardowego albo na trening własnego modelu rzędu miliarda parametrów od zera. Dla polskiej spółki jest to najtańsze dostępne wejście w poważną skalę — a ceną jest wniosek, kolejka i brak gwarancji terminu.

Na koniec ograniczenie twardsze od ceny: terminy dostaw. Systemy poprzedniej generacji dostarcza się w cztery do sześciu tygodni, bieżącej — w dwanaście do dwudziestu, a wzmocnionej w szesnaście do dwudziestu ośmiu. Większość produkcji odbierają najwięksi odbiorcy, a kanał sprzedaży bywa dla pozostałych zamknięty. Planowanie zakupu bez uwzględnienia półrocznego wyprzedzenia jest planowaniem nierealnym.

ROZDZIAŁ 8 · KROK 6

Pretraining i jego matematyka

To jedyny rozdział, w którym cały budżet przedsięwzięcia daje się wyprowadzić z trzech liczb. Kto zrozumie zawartą tu arytmetykę, przestanie potrzebować czyichkolwiek szacunków — policzy sobie sam koszt dowolnego modelu na dowolnym sprzęcie.

Wzór, od którego wszystko się zaczyna

Nakład obliczeniowy pretrainingu opisuje przybliżenie, w którym liczba operacji zmiennoprzecinkowych równa się sześciokrotnemu iloczynowi liczby parametrów i liczby tokenów treningowych. Szóstka nie jest umowna: przebieg w przód kosztuje dwie operacje na parametr i token, a przebieg wsteczny cztery, ponieważ liczy się w nim dwa gradienty — względem wejść warstwy i względem samych wag.

$$C \approx 6 \cdot N \cdot D$$

C — OPERACJE ZMIENNOPRZECINKOWE · N — PARAMETRY · D — TOKENY TRENINGOWE

Przybliżenie to jest zaskakująco dokładne. Dla modelu GPT-3 iloczyn sześciu, stu siedemdziesięciu pięciu miliardów parametrów i trzystu miliardów tokenów daje $3,15 \times 10^{23}$ operacji, wobec wartości $3,14 \times 10^{23}$ podawanej w opracowaniach technicznych. Dla modelu Llama 3.1 o czterystu pięciu miliardach parametrów, trenowanego na 15,6 biliona tokenów, wychodzi $3,79 \times 10^{25}$ wobec $3,8 \times 10^{25}$ podanych przez firmę Meta. Zgodność do drugiej cyfry znaczącej — na tym można budować budżet.

Wzór pomija koszt mechanizmu uwagi, który nie skaluje się z liczbą parametrów, lecz z długością przetwarzanej sekwencji. Poprawka jest niewielka przy zwykłym kontekście: dla wspomnianego modelu przy oknie ośmiu tysięcy tokenów wynosi 8,3 procent. Przy oknie stu dwudziestu ośmiu tysięcy tokenów rośnie natomiast do 133 procent — sam mechanizm uwagi kosztuje wtedy więcej niż wszystkie mnożenia macierzy razem wzięte. To wyjaśnia praktykę, która z zewnątrz wygląda na dziwaczną: kontekst wydłuża się w krótkim, osobnym etapie na samym końcu pretrainingu, a nie na całym korpusie.

Od operacji do godzin i do pieniędzy

Drugim składnikiem rachunku jest wydajność sprzętu. Układ H100 osiąga szczytowo 989 bilionów operacji na sekundę w precyzji szesnastobitowej — często cytowana wartość dwukrotnie wyższa dotyczy trybu wykorzystującego rzadkość macierzy i nie ma zastosowania w treningu. Realnie wykorzystywana część tej mocy, określana jako wykorzystanie mocy modelu, mieści się między trzydziestoma pięcioma a pięćdziesięcioma procentami.

Tu kryje się pułapka, w którą wpada większość szacunków. Meta raportuje dla swojego największego modelu wykorzystanie na poziomie 41–43 procent. Jeśli jednak podzielić teoretyczny nakład obliczeniowy przez wszystkie rozliczone w karcie modelu godziny pracy kart, wychodzi 34,5 procent. Różnica sześciu punktów to awarie, restarty, rozgrzewka klastra i etapy poboczne. *Do budżetowania należy używać tej drugiej, gorszej liczby*, ponieważ płaci się za godziny wynajmu, a nie za operacje użyteczne.

STAŁA, KTÓRĄ WARTO ZAPAMIĘTAĆ

Pojedynczy układ H100 przy wykorzystaniu 40 % wykonuje 1,42 eksaoperacji na godzinę pracy. Przy cenie 2,50 USD za godzinę-kartę daje to **1,76 USD za eksaoperację**. Mnożąc tę stałą przez wynik wzoru wyrażony w eksaoperacjach, otrzymuje się koszt treningu w dolarach — bez tabel i bez kalkulatora branżowego.

Ceny wynajmu w 2026 roku są rozpięte niemal dziesięciokrotnie i to bywa źródłem nieporozumień przy porównywaniu opracowań. Karta H100 kosztuje od 1,50 do 3,00 dolara za godzinę na giełdach mocy i u pośredników, około 3,85–4,00 dolara u dostawców wyspecjalizowanych, a u największych dostawców chmurowych od 6,88 do 12,29 dolara zależnie od dostawcy i regionu. Nowszy układ B200 kosztuje 4,50–7,00 dolara za godzinę. Pełne zestawienie z podziałem na rynki zawiera tabela 7.1.

Warto zauważyć rzecz nieoczywistą: przy obecnych cenach *nowszy sprzęt nie jest tańszy*, dopóki liczy się w precyzji szesnastobitowej. Koszt eksaoperacji wynosi wtedy 1,76 dolara na H100 i 1,94 dolara na B200. Przewaga pojawia się dopiero po przejściu na precyzję ośmiobitową, gdzie B200 schodzi do 0,97 dolara — mniej więcej o połowę taniej. Wybór sprzętu jest więc pochodną tego, czy zespół potrafi prowadzić trening w niższej precyzji, a nie tego, który układ jest nowszy.

Kupić czy wynająć

Sprzęt własny wychodzi taniej, ale nie tak dramatycznie, jak sugerują rachunki liczące samą kartę. Przy pełnym nakładzie węzła — a więc wraz z obudową, procesorami głównymi, pamięcią, siecią, macierzą dyskową, kolokacją i obsługą — amortyzacja przez cztery lata przy wykorzystaniu dziewięćdziesięcioprocentowym daje około 2,0–2,6 dolara za godzinę-kartę dla sprzętu klasy H100. Energia dokłada do tego niecałe czterdzieści centów: dane o jednym z modeli czołowych pozwalają wyliczyć 1,26 kilowatogodziny na godzinę pracy karty, co stanowi 1,8-krotność jej poboru nominalnego i obejmuje chłodzenie, sieć oraz procesory główne. Pełne rozliczenie klastra znajduje się w rozdziale siódmym; tam ten sam rachunek dla sprzętu bieżącej generacji daje 3,58 dolara przy pełnym wykorzystaniu.

Ta dwukrotność wyjaśnia rozbieżność, która w publikacjach wygląda na sprzeczność. Koszt treningu modelu GPT-4 wyceniany jest na około 40 milionów dolarów metodą amortyzacji sprzętu i na 78 milionów metodą cen chmurowych. Dla modelu Gemini Ultra jest to odpowiednio 30 i 191 milionów. To nie są sprzeczne ustalenia, tylko dwie różne konwencje — i przy czytaniu jakiegokolwiek szacunku należy najpierw sprawdzić, którą przyjęto.

CZEGO TE KWOTY NIE OBEJMUJĄ

Wszystkie publikowane koszty dotyczą **przebiegu finalnego**. Wydatek na badania i rozwój — próby w małej skali, badania wpływu pojedynczych decyzji, przebiegi przerwane — jest wielokrotnie wyższy. Realny budżet programu to trzy- do sześciokrotność kwoty za przebieg finalny. W strukturze wydatków sprzęt stanowi 47–67 %, wynagrodzenia zespołu 29–49 %, a energia zaledwie 2–6 %.

Ile danych na parametr

W 2022 roku zespół DeepMind wykazał, że przy ustalonym budżecie obliczeniowym optimum leży tam, gdzie na każdy parametr przypada około dwadzieścia tokenów. Reguła ta minimalizuje jednak koszt *treningu*, a nie koszt cyklu życia modelu. Generowanie odpowiedzi kosztuje mniej więcej dwie operacje na parametr i wygenerowany token — dla modelu ośmiomiliardowego obsługującego sto milionów użytkowników roczny nakład na obsługę dorównuje nakładowi na jego wytrenowanie. Skoro tak, opłaca się zmniejszyć model i nadrobić jakość dłuższym treningiem, nawet daleko poza punkt teoretycznego optimum.

Stąd dzisiejsze proporcje, które z perspektywy reguły sprzed czterech lat wyglądają na ekstrawagancie. Model Llama 3 o ośmiu miliardach parametrów trenowano na piętnastu bilionach tokenów, czyli 1875 tokenów na parametr — dziewięćdziesięcioczekrotność dawnego optimum. Modele z rodziny Llama 4 sięgają 2353 tokenów na parametr aktywny. Jedynie największe modele pozostają blisko klasycznego punktu, bo przy ich rozmiarze koszt obsługi i tak jest zaporowy.

Mieszanka ekspertów zmienia mianownik

W architekturze rozdzielającej sieć na wyspecjalizowane fragmenty do wzoru wchodzi liczba parametrów *aktywnych*, a nie całkowitych. Model DeepSeek-V3 ma 671 miliardów parametrów, ale przy każdym tokenie uruchamia jedynie 37 miliardów — stosunek osiemnaście do jednego. Skutki dla rachunku są dramatyczne.

TABELA 8.1 — TA SAMA JAKOŚĆ, DWIE ARCHITEKTURY, TEN SAM KORPUS 20 BLN TOKENÓW

WARIANT	PARAMETRY	AKTYWNE	NAKŁAD	GODZINY H100	KOSZT
Gęsty	400 mld	400 mld	$4,8 \times 10^{25}$	33,8 mln	84,5 mln USD
Mieszanka ekspertów	400 mld	40 mld	$4,8 \times 10^{24}$	4,49 mln	11,2 mln USD

Wyliczenie własne wg wzoru 6ND, przy obniżonym do 30 % wykorzystaniu mocy dla mieszanki ekspertów z powodu kosztu komunikacji wszyscy-do-wszystkich. Różnica siedmio- do ośmiokrotna przy porównywalnej jakości.

Oszczędność nie jest darmowa. Zapotrzebowanie na pamięć skaluje się z liczbą parametrów całkowitych, nie aktywnych: stan optymalizatora zajmuje około szesnastu bajtów na parametr, co dla modelu 671-miliardowego daje 10,7 terabajta, które trzeba rozproszyć po setkach kart. Dochodzi komunikacja między wszystkimi węzłami przy każdej warstwie, obniżająca wykorzystanie mocy o kilka do kilkunastu punktów. Mimo to rachunek wychodzi tak korzystnie, że praktycznie wszystkie modele czołowe wydane po 2024 roku mają właśnie taką architekturę.

Co naprawdę kosztowały znane modele

TABELA 8.2 — ROZLICZONE NAKŁADY NA TRENING MODELI, O KTÓRYCH WIADOMO COKOLWIEK PEWNEGO

MODEL	NAKŁAD (FLOP)	GODZINY KART	KOSZT	PODSTAWA
GPT-3 (2020)	$3,14 \times 10^{23}$	—	4,6 mln USD	szacunek przy cenach z 2020 r.
GPT-4	$\approx 2 \times 10^{25}$	—	40 lub 78 mln USD	modelowanie zewnętrzne, dwie konwencje
Gemini Ultra	$\approx 10^{26}$	—	30 lub 191 mln USD	modelowanie zewnętrzne; 35 MW mocy
Llama 3.1 405B	$3,8 \times 10^{25}$	30,84 mln	62–170 mln USD	karta modelu producenta

MODEL	NAKŁAD (FLOP)	GODZINY KART	KOSZT	PODSTAWA
Llama 4 Scout	—	5,00 mln	10–20 mln USD	karta modelu producenta
Llama 4 Maverick	—	2,38 mln	5–10 mln USD	karta modelu producenta
DeepSeek-V3	$3,3 \times 10^{24}$	2,79 mln	5,58 mln USD	raport techniczny autorów
DeepSeek-R1	$6,1 \times 10^{23}$	≈ 147 tys.	0,29–1,0 mln USD	publikacja recenzowana i szacunek zewnętrzny
Grok 4	$\approx 3 \times 10^{26}$	246 mln	490 mln USD	szacunek zewnętrzny, duża niepewność
Czołówka 2026	10^{26} – 10^{27}	—	200–500 mln USD	szacunki branżowe

Godziny kart w przeliczeniu na H100 (dla modeli DeepSeek — H800 o tej samej mocy obliczeniowej). Kwota dla DeepSeek-V3 obejmuje wyłącznie przebieg oficjalny przy założonej przez autorów cenie 2 USD za godzinę-kartę; nie zawiera badań wstępnych, danych ani sprzętu. Kwota dla modelu R1 nie zawiera kosztu modelu bazowego.

Warto zatrzymać się przy dwóch pozycjach. Model DeepSeek-V3 pokazuje, że przy dobrej inżynierii wynik zbliżony do czołówki można osiągnąć za mniej niż sześć milionów dolarów przebiegu — kwotę mieszczącą się w zasięgu średniej firmy technologicznej. Model DeepSeek-R1, który dodał do niego zdolność rozumowania, kosztował dwieście dziewięćdziesiąt cztery tysiące dolarów, czyli pięć procent kosztu modelu bazowego. To najważniejsza liczba w całej tabeli: *zdolności, które robią największe wrażenie na użytkowniku, powstają w najtańszym etapie*.

Druga obserwacja dotyczy tempa. Koszt amortyzowany pojedynczego przebiegu czołowego rośnie 2,4-krotnie rocznie i robi to nieprzerwanie od 2016 roku. Ekstrapolacja tej krzywej wskazuje przekroczenie miliarda dolarów za jeden przebieg około 2027 roku. Kto rozważa gonienie czołówki, ściga się nie z konkurentem, tylko z funkcją wykładniczą.

Realia sześciotygodniowego przebiegu

Trening w dużej skali nie jest uruchomieniem programu i czekaniem. Jest ciągłą walką z awariami sprzętu, których liczba wynika ze statystyki, a nie z pecha. Podczas pięćdziesięciu czterech dni treningu na 16 384 kartach Meta odnotowała 419 nieplanowanych przerw — średnio jedno co trzy godziny. Blisko połowa wynikała z usterek samych kart i ich pamięci: 148 przypadków to wady procesora graficznego, 72 — pamięci wysokoprzepustowej, 35 — przełączników i okablowania.

Z tych liczb wynika czas bezawaryjnej pracy pojedynczej karty: około pięćdziesięciu tysięcy godzin, czyli blisko sześć lat pracy ciągłej. Brzmi solidnie, dopóki nie podzieli się tej wartości przez liczbę kart w klastrze. Przy stu tysiącach kart średni odstęp między awariami wynosi pół godziny.

$$T_{\text{opt}} \approx \sqrt{(2 \cdot \delta \cdot \text{MTBF})}$$

OPTIMALNY ODSTĘP ZAPISU PUNKTU KONTROLNEGO — WZÓR YOUNGA I DALY'EGO

Odpowiedzią na awarie jest regularny zapis stanu treningu. Optymalny odstęp między zapisami wynika z powyższego wzoru, gdzie δ to czas zapisu, a strata czasu wynosi w przybliżeniu pierwiastek z ilorazu podwojonego czasu zapisu i średniego odstępów między awariami. Przy zapisie trwającym minutę i awarii co trzy godziny optimum wypada co dziewiętnaście minut, a strata wynosi 10,5 procent — dokładnie tyle, ile Mecie zabrakło do stu procent czasu użytecznego. Przy stu tysiącach kart optimum spada do ośmiu minut, a

strata rośnie do 25,6 procent. Dlatego przy największych klastrach stosuje się zapis asynchroniczny i przechowywanie stanu w pamięci sąsiednich węzłów; podejście naiwne spaliłoby ćwierć budżetu. Sam zapis nie jest drobiazgiem — punkt kontrolny modelu czterystumiliardowego zajmuje około pięciu terabajtów, przy czym warto rozróżnić dwie wielkości mylone niemal powszechnie: *stan treningu w pamięci* to około szesnastu bajtów na parametr, natomiast *zapisywany punkt kontrolny* dwanaście do czternastu, ponieważ gradientów się nie utrwała — po wznowieniu są przeliczane.

Osobną kategorią kłopotów są nagłe skoki funkcji straty, przy których model w ciągu kilku kroków traci znaczną część osiągniętej jakości. Zespół pracujący nad modelem PaLM odnotował ponad dwadzieścia takich zdarzeń. Przyjęta procedura polegała na cofnięciu się do zapisu sprzed około stu kroków i pominięciu kilkuset porcji danych. Co istotne, ponowne przetworzenie tych samych danych z innego punktu kontrolnego nie odtwarzało problemu — źródłem było zatem współdziałanie konkretnej porcji ze stanem optymalizatora, a nie wadliwe dane.

Trzy rachunki na 2026 rok

Poniżej pełne wyliczenie dla korpusu piętnastu bilionów tokenów, przy założeniach podanych jawnie: karty H100, precyzja szesnastobitowa, wykorzystanie mocy 40 procent, cena 2,50 dolara za godzinę-kartę, czas użyteczny 90 procent.

TABELA 8.3 — KOSZT PRZEBIEGU FINALNEGO, KORPUS 15 BLN TOKENÓW

MODEL	NAKŁAD	GODZINY H100	KOSZT (WIDEŁKI)	CZAS	NA B200 W FP8
8 mld parametrów	$7,2 \times 10^{23}$	507 tys.	1,27 mln USD 0,91–1,77	23 dni 1024 karty	0,70 mln USD
70 mld parametrów	$6,3 \times 10^{24}$	4,44 mln	11,1 mln USD 8,0–15,5	50 dni 4096 kart	6,1 mln USD
400 mld parametrów	$3,6 \times 10^{25}$	25,4 mln	63,4 mln USD 45,6–88,8	72 dni 16 384 karty	34,9 mln USD

Wyliczenie własne. Kalibracja: dla modelu 400-miliardowego rachunek daje 25,4 mln godzin, a Meta rozliczyła 30,84 mln — czyli o 21 % więcej, co odpowiada niższemu efektywnemu wykorzystaniu mocy. Kwoty dotyczą wyłącznie przebiegu finalnego; pełny budżet programu jest 3–6-krotnie wyższy.

WNIOSEK DLA DECYDENTA

Model ośmiomiliardowy to wydatek rządu miliona dolarów za przebieg i pięciu milionów za program — próg osiągalny dla dobrze finansowanej spółki. Model siedemdziesięciomiliardowy to trzydzieści pięć do siedemdziesięciu milionów dolarów programu. Powyżej tego poziomu przedsięwzięcie przestaje być decyzją inżynierską, a staje się decyzją o alokacji kapitału porównywalną z budową zakładu produkcyjnego.

ROZDZIAŁ 9 · KROK 7

Jak z automatu powstaje asystent

Etap, na którym model przestaje dopisywać tekst, a zaczyna odpowiadać. Sprzętowo najtańszy w całym łańcuchu, kadrowo najdroższy — i od 2024 roku decydujący o tym, czy model uchodzi za dobry.

Po pretreningu mamy obiekt, który potrafi jedno: przewidywać kolejny fragment tekstu zgodnie z rozkładem korpusu. Zapytany o sposób napisania pozwu równie chętnie wypisze listę podobnych pytań z forum, co odpowie na to konkretne. Post-trening wycina z tego rozkładu jedno zachowanie — odpowiadanie na polecenie — i dokłada umiejętności, których w korpusie nie było w postaci użytecznej: rozumowanie krok po kroku, korzystanie z narzędzi, odmowę w sytuacjach ryzykownych.

Etap pierwszy: dostrojenie na instrukcjach

Modelowi pokazuje się pary polecenie–wzorcowa odpowiedź. Skala potrzebnych danych jest zaskakująco elastyczna i to najważniejsza obserwacja praktyczna całego rozdziału. Eksperyment opublikowany w 2023 roku wykazał, że *tysiąc starannie dobranych przykładów* — siedemset pięćdziesiąt wybranych z forów i dwieście pięćdziesiąt napisanych ręcznie — wystarczyło, żeby model bazowy wygrywał porównania z modelem komercyjnym dostrojonym pełną procedurą. Autorzy nazwali to hipotezą powierzchownej adaptacji: wiedza powstaje w pretreningu, a dostrojenie jedynie wybiera z niej właściwy podrozkład zachowań.

Nowsze receptury pracują na wyższych rzędach, ale nie dlatego, że tysiąc przykładów przestało wystarczać do nadania formy. Mieszanka użyta w otwartym projekcie Tülu 3 liczy 939 tysięcy przykładów, a Deep-Seek-R1 zbudował korpus około ośmiuset tysięcy — z czego sześćset tysięcy dotyczyło rozumowania. Te wielkości potrzebne są do *dokładania nowych kompetencji*, nie do zmiany formatu.

REGUŁA PRAKTYCZNA

Od 10 do 50 tysięcy przykładów wystarcza, żeby nadać modelowi styl, format i sposób odmowy. Od 100 tysięcy do miliona potrzeba dopiero wtedy, gdy dokłada się nowe umiejętności: obsługę narzędzi, długie łańcuchy rozumowania albo kompetencję w wąskiej dziedzinie.

Ile kosztuje anotacja

To jedyna pozycja post-treningu, która realnie obciąża budżet — i pozycja, w której stawki rozwarstwiły się drastycznie w zależności od wymaganych kompetencji.

TABELA 9.1 — STAWKI ZA PRZYGOTOWANIE DANYCH TRENINGOWYCH, 2026 R.

RODZAJ PRACY	STAWKA (USD/H)	ZASTOSOWANIE
Anotacja prosta, wykonawcy masowi	3–15	oznaczanie, porządkowanie
Anotacja kontraktowa u wyspecjalizowanego dostawcy	18–24	oceny porównawcze
Ocena generalistyczna z weryfikacją	17–40	kontrola jakości
Ekspert dziedzinowy — średnia	81–85	zadania specjalistyczne
Prawnik oceniający odpowiedzi prawne	110–130	dziedziny regulowane
Ekspert z wieloletnim doświadczeniem	ponad 200	zadania najtrudniejsze

Stawki platform pośredniczących w latach 2025–2026. Przychody tego rynku są znaczące: jeden z dostawców podał 1,2 mld USD przychodu w skali roku, drugi około 870 mln USD, przy klienteli liczącej kilkanaście laboratoriów.

$$\text{koszt} = \text{liczba przykładów} \times (\text{czas w minutach} \div 60) \times \text{stawka} \times \text{narzut } 1,3\text{--}1,6$$

NARZUT OBEJMUJE PRZEGLĄD I ODRZUTY — CZĘŚĆ PRACY ZAWSZE IDZIE DO KOSZA

Dwa rachunki dla orientacji. Dwieście tysięcy przykładów prostych po cztery minuty przy stawce dwudziestu dolarów to 267 tysięcy dolarów. Trzydzieści tysięcy przykładów prawniczych po dwadzieścia pięć minut przy stawce stu dwudziestu dolarów to 1,5 miliona dolarów przed narzutem. Różnica pokazuje, gdzie leży cały ciężar budżetu: nie w obliczeniach, lecz w czasie pracy specjalistów.

Etap drugi: dostrojenie preferencji

Klasyczna procedura ma trzy kroki: zebranie porównań, w których człowiek wskazuje lepszą z dwóch odpowiedzi, wytrenowanie na nich modelu nagrody i optymalizacja zachowania modelu względem tej nagrody. Kosztowna jest nie liczba kroków, tylko zapotrzebowanie na pamięć — procedura wymaga jednoczesnej obecności czterech modeli w pamięci kart. Do tego dochodzi kruchość: model szybko uczy się wykorzystywać słabości modelu nagrody, zamiast poprawiać jakość odpowiedzi.

Stąd fala metod tańszych, które w 2026 roku są już standardem. Bezpośrednia optymalizacja preferencji wyprowadza rozwiązanie wprost z par porównawczych, likwidując model nagrody i całą pętlę uczenia ze wzmocnieniem. Wariant oparty na sygnale binarnym — zwykłej ocenie „dobra albo zła odpowiedź” — ma znaczenie ekonomiczne trudne do przecenienia, ponieważ takie dane spływają z produkcji bezpłatnie i nie trzeba zamawiać porównań. Metoda zastosowana w modelach DeepSeek usuwa jeden z czterech modeli, odnosząc ocenę do średniej z grupy kilkunastu odpowiedzi na to samo polecenie, co obcina zapotrzebowanie na zasoby mniej więcej o połowę.

Etap trzeci: nagrody weryfikowalne

Od 2024 roku etapem decydującym stało się uczenie na zadaniach o sprawdzalnym wyniku. Zamiast pytać człowieka, czy odpowiedź jest lepsza, uruchamia się weryfikator: sprawdza wynik zadania matematycznego, uruchamia testy jednostkowe, porównuje stan środowiska. Nagroda staje się wtedy tania, obiektywna i odporna na manipulację — a model uczy się rozumowania, którego nie da się wyćwiczyć samym naśladowaniem.

Droga modelu DeepSeek-R1 jest udokumentowana i warto ją znać, bo stanowi wzorzec: kilka tysięcy przykładów długiego rozumowania na rozgrzewkę, pierwsza runda uczenia ze wzmocnieniem, wygenerowanie ośmiuset tysięcy przykładów przez odsiewanie własnych odpowiedzi, ponowne dostrojenie i druga runda. Koszt samej warstwy uczenia ze wzmocnieniem podano w publikacji recenzowanej jako 294 tysiące dolarów przy pięciuset dwunastu kartach.

DLACZEGO TA FAZA BYWA DROŻSZA OD PRETRENINGU

Po pierwsze, dominuje w niej generowanie, a nie uczenie — model musi wyprodukować kilkanaście długich rozwiązań na każde polecenie, a generowanie wykorzystuje sprzęt w kilku procentach zamiast w kilkudziesięciu. Po drugie, skalowanie jest agresywne: w kolejnych pokoleniach modeli rozumujących nakład na ten etap urósł z jednego procenta kosztu pretreningu do jego równowartości. Po trzecie, doszedł koszt nowy — środowisk treningowych, wycenianych na około 2400 dolarów za jedno zadanie, przy czym jedno z laboratoriów deklaroowało przeznaczenie na nie ponad miliarda dolarów rocznie.

Tańsze przeciwwagi

Zamiast ocen ludzkich można użyć ocen wystawianych przez inny model, kierujący się spisanim zbiorem kilkunastu zasad. Porównanie opublikowane w 2023 roku podaje sześć centów za etykietę wygenerowaną maszynowo wobec sześćdziesięciu siedmiu centów za etykietę ludzką — przewaga jedenastokrotna przy porównywalnej jakości ocen. Ograniczenie jest znane: model oceniający przenosi własne skrzywienia, w szczególności upodobanie do odpowiedzi dłuższych, więc w sprawach spornych i w kwestiach bezpieczeństwa nadal utrzymuje się ocenę ludzką.

Drogą najtańszą jest uczenie modelu mniejszego na odpowiedziach większego. Rachunek jest wprost policzalny: osiemset tysięcy przykładów po cztery tysiące tokenów to 3,2 miliarda tokenów, czyli od siedmiu tysięcy do stu dziewięćdziesięciu tysięcy dolarów zależnie od ceny modelu źródłowego. Droga ta niesie jednak ryzyko prawne, opisane w rozdziale czternastym — warunki korzystania z modeli komercyjnych zwykle zakazują wykorzystywania ich odpowiedzi do budowy modeli konkurencyjnych.

Rachunek całości

Same obliczenia dostrojenia są zaskakująco tanie. Dla modelu ośmiomiliardowego i trzystu tysięcy przykładów po półtora tysiąca tokenów, przy trzech przejściach przez zbiór, wychodzi około czterdziestu pięciu godzin pracy karty — poniżej stu pięćdziesięciu dolarów. Dla modelu siedemdziesięciomiliardowego około czterystu godzin, czyli jeden do dwóch tysięcy dolarów. *Budżet post-treningu jest zdominowany przez dane i przez uczenie ze wzmocnieniem, nie przez samo dostrojenie.*

TABELA 9.2 — PEŁNY KOSZT POST-TRENINGU WEDŁUG SKALI MODELU

SKALA	OBLICZENIA DOSTROJENIA	ETAP WZMOCNIENIA	RAZEM Z DANYMI OTWARTYMI	RAZEM Z ANOTACJĄ EKSPERCKĄ
8 mld	< 150 USD	30–80 tys. USD	20–150 tys. USD	0,5–3 mln USD
70 mld	1–2 tys. USD	150–600 tys. USD	0,3–1 mln USD	1–10 mln USD
Czołówka	—	do 100 % kosztu pretreningu	—	300 mln – 1,5 mld USD

Widelki dla czołówki złożone z ujawnień cząstkowych, nie raportowane przez żadne laboratorium wprost. Obejmują zakup danych, środowiska treningowe i obliczenia etapu wzmocnienia.

WNIOSEK DLA MAŁEGO ZESPOŁU

Post-trening jest jedynym etapem, w którym mały zespół może dorównać dużemu — bo decyduje w nim jakość danych, a nie wielkość klastra. Zespół dysponujący dostępem do rzetelnych, dziedzinowych przykładów rozwiązanych przez specjalistów zbuduje w swojej dziedzinie model lepszy niż laboratorium z tysiącokrotnie większym budżetem obliczeniowym, ale bez tych danych. Dla kogoś, kto pracuje w zawodzie regulowanym, jest to informacja o realnej przewadze konkurencyjnej.

ROZDZIAŁ 10 · KROK 8

Ewaluacja i bezpieczeństwo

Etap, który decyduje o tym, czy wolno model udostępnić — i jedyny, w którym najczęściej popełnianym błędem jest zawierzenie własnym wynikom. Rozdział omawia, jak mierzyć jakość,

dlaczego liczby z materiałów prasowych są niewiarygodne i ile kosztuje rzetelne sprawdzenie modelu.

Dlaczego wyniki publikowane nie służą do decyzji

Liczba podana w karcie modelu jest wynikiem trzech decyzji, z których żadna nie jest oczywista: doboru zbioru zadań, doboru narzędzia uruchomieniowego oraz liczby powtórzeń. Zmiana którejkolwiek przesuwając wynik o kilka punktów procentowych. Zestawienia publikowane przez różne podmioty dla tych samych modeli w tym samym miesiącu potrafią różnić się o dwa punkty procentowe — nie dlatego, że model się zmienił, lecz dlatego, że inaczej go zmierzono.

PIERWSZA ZASADA OPERACYJNA

Wyniki publiczne służą do wstępnej selekcji kandydatów, nigdy do decyzji wdrożeniowej. Decyzję podejmuje się na własnym zbiorze zadań, zbudowanym z realnych przypadków produktu i utrzymywanym w tajemnicy.

Kontaminacja jako stały podatek

Zanieczyszczenie zbiorów testowych — sytuacja, w której model widział pytania egzaminacyjne podczas treningu — nie jest patologią marginesu, tylko stałym obciążeniem każdego zbioru publicznego. Pomiar opublikowany w 2024 roku wykazał ślady zanieczyszczenia w 29,1 procent pozycji jednego z najpopularniejszych zbiorów. Na innym zbiorze podmiana skażonych zadań na czyste odpowiedniki o tej samej strukturze obniżyła wynik modelu o trzynaście punktów procentowych. Sama zmiana kolejności wariantów odpowiedzi potrafi obniżyć trafność o kilkanaście procent, co pokazuje, że część wyniku pochodzi z rozpoznania wzorca, a nie z rozwiązania zadania.

$$\text{delta kontaminacji} = \text{wynik na zbiorze publicznym} - \text{wynik na prywatnym klonie}$$

POWYŻEJ 5 PUNKTÓW PROCENTOWYCH LICZBA PUBLICZNA TRACI WARTOŚĆ PROGNOSTYCZNĄ

Praktyczna miara własna jest prosta: zbuduj prywatny odpowiednik zbioru publicznego, o tej samej trudności, ale utworzony po dacie odcięcia danych treningowych. Różnica wyników mówi, ile z publicznej liczby to realna zdolność, a ile pamięć.

Ewaluacja po polsku

Dla polszczyzny istnieje użyteczny, choć rozproszony zestaw narzędzi, którego znajomość oszczędza kwartał pracy.

TABELA 10.1 — POLSKIE ZBIORY EWALUACYJNE

ZBIÓR	SKALA	DO CZEGO SŁUŻY
LLMzSzŁ	≈ 19 tys. pytań	Archiwum Centralnej Komisji Egzaminacyjnej: cztery typy egzaminów, 154 dziedziny wraz z zawodowymi. Jedyny duży zbiór pozwalający zestawzić wynik modelu ze zdawalnością ludzi na tym samym arkuszu.
KLEJ, LEPISZCZE	zestawy zadań	Rozumienie języka — polski odpowiednik standardowych zestawów anglojęzycznych.
PL-MTEB	30 zadań	Jakość reprezentacji wektorowych w pięciu kategoriach. Właściwe narzędzie do wyboru modelu osadzeń pod wyszukiwanie w dokumentach.
PIRB	41 zadań	Wyszukiwanie, w tym na materiale medycznym, prawnym i biznesowym.
Open PL LLM Leaderboard	ranking otwarty	Porównanie modeli generatywnych po polsku, prowadzone przez środowisko SpeakLeash.

OSTRZEŻENIE Z POLSKIEGO PODWÓRKA

Modele jednego z dużych polskich projektów zostały wyłączone z rankingu, ponieważ trenowano je na treningowych częściach zadań wchodzących w skład tego rankingu. To ilustracja zjawiska, które w małym ekosystemie językowym pojawia się szybciej niż w anglojęzycznym: dostępnych korpusów jest po prostu mniej, więc zbiór treningowy i testowy nachodzą na siebie niemal automatycznie. Kto buduje polski model, musi rozdzielić te zbiory świadomie i udokumentować rozdział.

Ile kosztuje sprawdzenie modelu

Koszt obliczeń jest przewidywalny. Dla realistycznego zestawu produktowego — pięćset przypadków, trzy powtórzenia, osiem tysięcy tokenów wejścia i dwa tysiące wyjścia — jeden pełny przebieg kosztuje około osiemdziesięciu dolarów przy typowym cenniku. Ocena wykonywana przez model sędziowski dokłada trzydzięści do pięćdziesięciu procent, a modele generujące długie łańcuchy rozumowania potrafią zwielokrotnić człon wyjściowy dziesięciokrotnie.

Znacznie poważniejszą pozycją są ludzie. Stawki ekspertów dziedzinowych mieszczą się w 2026 roku w przedziale od pięćdziesięciu do dwustu dolarów za godzinę, przy czym specjaliści z medycyny, prawa i finansów wyceniani są na 175–300 dolarów, a lekarze na 250–450. Pojedynczy przykład oceniony ekspercko kosztuje od pięćdziesięciu do stu dolarów, wobec pół dolara do pięciu w wariancie masowym.

Rozwiązaniem pośrednim jest ocena przez model sędziowski, wymagająca jednak kalibracji. Zgodność najlepszych sędziów maszynowych z ludźmi sięga 79,9 procent — przy zgodności dwóch ludzi między sobą wynoszącej 79,2 procent. Przyjęta praktyka to pilotaż na około dwustu przykładach ocenionych podwójnie, doprowadzenie miary zgodności powyżej wartości 0,6 i utrzymanie pięciu do dziesięciu procent weryfikacji ludzkiej w trybie ciągłym.

Czerwony zespół

Testowanie odporności modelu na próby wymuszenia zachowań niepożądanych przestało być czynnością jednorazową. Najlepiej udokumentowany przykład publiczny obejmował blisko stu osiemdziesięciu pięciu uczestników, którzy przez dwa miesiące poświęcili około trzech tysięcy godzin na próby przełamania zabezpieczeń. Skuteczność przełamań spadła w wyniku wprowadzonych poprawek z 86 do 4,4 procent, kosztem 23,7-procentowego narzutu obliczeniowego. Kolejna generacja zabezpieczeń, testowana w ponad tysiącu

siedmiuset godzinach i stu dziewięćdziesięciu ośmiu tysiącach prób, osiągnęła ten sam skutek przy narzucie rzędu jednego procenta.

TABELA 10.2 — KOSZT TESTÓW ODPORNOŚCIOWYCH DLA FIRMY PRODUKTOWEJ

ZAKRES	KOSZT (USD)
Jednorazowy audyt podstawowy	8–25 tys.
Pojedyncze pełne zaangażowanie zespołu	16–100 tys.
System wieloagentowy o rozbudowanej powierzchni ataku	50–150 tys.
Testowanie ciągłe, miesięcznie	od 5 tys.

Dla porównania: 1700 godzin pracy specjalisty przy średniej stawce rynkowej około 68 USD za godzinę to około 115 tys. USD samej pracy, bez kosztu obliczeń, koordynacji i obsługi zgłoszeń.

NAJCZĘSTSZY BŁĄD METODYCZNY

Badanie opublikowane w maju 2026 r. wykazało skuteczność ataków rozłożonych na wiele wiadomości sięgającą 88 % wobec piętnastu flagowych modeli czterech dostawców — a technika polegająca na stopniowej eskalacji, bez ani jednego jawnie wrogiego polecenia, osiągała do 98 %. Wniosek jest jednoznaczny: *zestaw testowy złożony z pojedynczych szkodliwych zapytań nie mierzy realnego ryzyka*. Filtry muszą działać na kontekście całej rozmowy, a nie na ostatniej wiadomości. W systemach korzystających z narzędzi dochodzi wektor wstrzykiwania poleceń przez treść pobieraną z sieci i z dokumentów, którego klasyczne filtry w ogóle nie obejmują.

Co z tego wynika dla planującego

Ewaluacja i bezpieczeństwo to zwykle półtora do dwóch miesięcy pracy zespołu przed udostępnieniem modelu, a następnie stały koszt bieżący. Dla przedsięwzięcia w skali opisanej w tym opracowaniu rozsądny budżet to od pięćdziesięciu do trzystu tysięcy dolarów na pierwsze udostępnienie, przy czym większość tej kwoty stanowi praca specjalistów, nie obliczenia. Warto tę pozycję zaplanować od początku — jest to bowiem etap, który przy presji terminowej skraca się jako pierwszy, a którego pominięcie ujawnia się publicznie.

ROZDZIAŁ 11 · KROK 9

Wdrożenie i ekonomika działania

Trening kończy się raz, a serwowanie trwa. W skali usługi koszt obsługi zapytań przewyższa koszt treningu, a rządzą nim zupełnie inne prawa. Rozdział zamyka rachunkiem, który dla większości firm rozstrzyga sprawę: własny hosting czy cudze udostępnienie modelu.

Dwie fazy o przeciwnych profilach

Obsługa zapytania rozpada się na dwie części, które obciążają zupełnie inne zasoby. Przetworzenie treści pytania jest ograniczone mocą obliczeniową — kosztuje około dwóch operacji na parametr i token, więc karta bieżącej generacji przetwarza pięć do sześciu tysięcy tokenów zapytania na sekundę. Generowanie odpowiedzi jest natomiast ograniczone przepustowością pamięci, ponieważ przy każdym kolejnym tokenie trzeba odczytać komplet wag.

prędkość generowania $\approx$ przepustowość pamięci $\div$ rozmiar wag w bajtach

NAJBARDZIEJ UŻYTECZNY WZÓR W CAŁEJ EKONOMICZNEJ SERWOWANIA

Rachunek sprawdza się co do jedności. Model siedemdziesięciomiliardowy w precyzji szesnastobitowej zajmuje 140 gigabajtów; przy przepustowości 3,35 terabajta na sekundę daje to dwadzieścia cztery tokeny na sekundę — i dokładnie tyle mierzą publiczne testy przy pojedynczym zapytaniu. Po przejściu na precyzję ośmiobitową wychodzi około czterdziestu ośmiu.

Pamięć podręczna bywa większa niż model

Zapotrzebowanie na pamięć karty to nie tylko wagi. Bufor przechowujący stan rozmowy rośnie z długością kontekstu i bywa niedoszacowany. Dla modelu siedemdziesięciomiliardowego wynosi 320 kilobajtów na token, więc jedna rozmowa o kontekście stu dwudziestu ośmiu tysięcy tokenów zajmuje około czterdziestu jeden gigabajtów — *więcej niż wagi tego samego modelu po kwantyzacji do czterech bitów*. Sto równoległych rozmów po osiem tysięcy tokenów to 256 gigabajtów samego bufora.

To jest właściwy powód, dla którego długi kontekst jest drogi: nie moc obliczeniowa, tylko pamięć i pismo jej odczytu. Środki zaradcze obejmują kwantyzację samego bufora, architektury z kompresją stanu, współdzielenie wspólnych początków zapytań oraz przenoszenie bufora do pamięci komputera głównego.

Grupowanie zapytań i stukrotna różnica kosztu

Obsługiwanie zapytań pojedynczo marnuje sprzęt, bo wagi trzeba odczytać dla każdego z osobna. Grupowanie ciągłe, w którym ukończone zapytania natychmiast opuszczają grupę, a nowe dołączają w trakcie, pozwala odczytać wagi raz na całą grupę. Próg, powyżej którego dokładanie kolejnych zapytań przestaje pomagać, daje się policzyć ze stosunku mocy obliczeniowej do przepustowości pamięci i wypada w okolicach dwustu dziewięćdziesięciu — dlatego testy przemysłowe zbiegają się na grupie liczącej dwieście pięćdziesiąt sześć zapytań.

LICZBA, KTÓRA PRZESĄDZA O MODELU BIZNESOWYM

Te same osiem kart obsługujących model siedemdziesięciomiliardowy daje około 2800 tokenów na sekundę przy pełnej grupie i około 25 przy obsłudze pojedynczego zapytania. To **stukrotna różnica kosztu jednostkowego na identycznym sprzęcie**. Usługa o małym, nierównomiernym ruchu płaci zatem za token sto razy więcej niż usługa o ruchu stałym — i to, a nie cena sprzętu, przesądza zwykle o nieopłacalności własnego hostingu na wczesnym etapie.

TABELA 11.1 — PRZEPUSTOWOŚĆ I KOSZT JEDNOSTKOWY, SPRZĘT 2026 R.

KONFIGURACJA	TOKENÓW NA SEKUNDĘ	KOSZT ZA MLN TOKENÓW
Model 8 mld, jedna karta H100	12–16 tys.	—
Model 70 mld, osiem H100, precyzja 16-bitowa	2 800	2,30 USD
Model 70 mld, osiem H100, precyzja 8-bitowa	5 600	1,15 USD
Model 70 mld, osiem H200	3 600	2,28 USD
Model 70 mld, osiem B200, ceny przerywane	5 200	0,88 USD
Model 671 mld z mieszanką ekspertów, osiem H100	1 800	3,58 USD

Warunki: grupowanie ciągle, 512 tokenów wejścia i 512 wyjścia, grupa 256. Skalowanie między generacjami sprzętu jest w przybliżeniu skalowaniem przepustowości pamięci.

Kwantyzacja i destylacja

Zapis wag z mniejszą precyzją jest dziś decyzją domyślną. Zejście do ośmiu bitów zmniejsza wagi o połowę, podwaja prędkość generowania i kosztuje poniżej dwóch procent jakości. Zejście do czterech bitów daje kolejne oszczędności, ale straty są wyraźniejsze i nierównomierne — najbardziej cierpią zadania wymagające wieloetapowego rozumowania i rachunków, najmniej klasyfikacja i wyciąganie informacji z tekstu.

Destylacja działa na innej osi: zamiast ścisnąć ten sam model, uczy się mniejszy na odpowiedziach większego. Zastąpienie modelu siedemdziesięciomiliardowego ośmiomiliardowym, wyuczonym pod wąską dziedzinę, obniża koszt obsługi mniej więcej dziesięciokrotnie i zwykle zachowuje dziewięćdziesiąt do dziewięćdziesięciu pięciu procent jakości zadaniowej. To szacunek zależny od dziedziny i wymagający własnego zbioru sprawdzającego przed wdrożeniem — ale dla usługi o ustabilizowanym zakresie zadań jest to zwykle najlepszy dostępny ruch optymalizacyjny.

Własny hosting czy cudza usługa

Rachunek dla typowej firmy średniej wielkości: tysiąc aktywnych użytkowników dziennie, po dwanaście zapytań, średnio dwa tysiące tokenów wejścia i pięćset wyjścia. Daje to dziewięćset milionów tokenów miesięcznie.

Minimalna sensowna konfiguracja własna dla modelu siedemdziesięciomiliardowego to dwie karty w jednym węźle, a przy zapewnieniu ciągłości działania — dwa takie węzły, czyli cztery karty. Przy medianie rynkowej daje to około 9200 dolarów miesięcznie za sprzęt. Doliczając ułamek etatu inżyniera platformy oraz monitoring i sieć, koszt całkowity wynosi około 12 200 dolarów miesięcznie, czyli mniej więcej 44 tysiące złotych.

GDZIE LEŻY PRÓG OPŁACALNOŚCI

Przy powyższym wolumenie zakup tych samych tokenów przez udostępnienie cudzego modelu kosztuje — zależnie od klasy modelu — od kilkuset dolarów za modele tanie do kilkunastu tysięcy za modele czołowe. Próg opłacalności własnego hostingu przebiega zatem nie przy określonej liczbie użytkowników, lecz przy określonej *równomierności* ruchu i wymaganiach dotyczących miejsca przetwarzania danych. Dla usługi o zmiennym obciążeniu, mniejszej niż

kilka miliardów tokenów miesięcznie, własny hosting jest zwykle droższy — chyba że przesądzą o nim wymogi poufności albo regulacyjne.

Na koniec pozycje, które w kalkulacjach znikają, a w rachunku bankowym nie: dyżury i obsługa awarii, monitoring, automatyczne skalowanie, moc zapasowa na wypadek szczytu, testy przed każdą podmianą wersji oraz koszt utrzymania jakości po zmianie modelu bazowego. Rozsądnie jest przyjąć, że koszt całkowity własnego serwowania to półtorakrotność ceny samego sprzętu.

CZĘŚĆ III

Rachunek

Pełny kosztorys czterech przedsięwzięć o rosnącej skali, z rozbiem na pozycje i harmonogramem. Tu widać, gdzie przebiega granica między projektem wykonalnym a niewykonalnym.

ROZDZIAŁ 12 · RACHUNEK

Kosztorys w czterech scenariuszach

Wszystko, co poprzednie rozdziały rozliczały osobno, zebrane w cztery budżety o rosnącej skali. Kwoty w złotych, przy kursie 3,7425 za dolara z 31 lipca 2026 roku. Każda pozycja ma odsyłacz do rozdziału, z którego pochodzi.

Scenariusz A — własny model na laptopie

Model o dwudziestu do pięćdziesięciu milionów parametrów, wytrenowany od zera na polskim korpusie, na komputerze, który już się posiada. Wagi w stu procentach własne, pochodzenie każdego bitu znane. Pełny przepis wykonawczy zawiera załącznik A.

TABELA 12.1 — SCENARIUSZ A

POZYCJA	KOSZT	CZAS
Korpus — pobranie gotowych zbiorów polskich	0 zł	1 dzień
Tokenizer — trening na własnym korpusie	0 zł	2 godziny
Trening — energia elektryczna	30–150 zł	1–4 dni
Dostrojenie na instrukcjach	0 zł	kilka godzin
Razem	30–150 zł	tydzień

Efekt: model piszący poprawną gramatycznie polszczyznę i naśladowujący styl korpusu, niezdolny do rzetelnej rozmowy. Wartość edukacyjna bardzo wysoka, użytkowa żadna.

Scenariusz B — model dziedzinowy przez adaptację

Kontynuowany pretrening otwartego modelu ośmiomiliardowego na pięćdziesięciu miliardach tokenów polskich, własny tokenizer, własny post-trening na danych dziedzinowych. To scenariusz realny dla polskiej spółki i to on jest przedmiotem rekomendacji w rozdziale siedemnastym.

TABELA 12.2 — SCENARIUSZ B, PIERWSZY ROK

POZYCJA	WARIANT OSZCZĘDNY	WARIANT PEŁNY	ROZDZIAŁ
PRZYGOTOWANIE			
Zebranie i oczyszczenie korpusu polskiego	15 tys. zł	80 tys. zł	4
Trening własnego tokenizatora	1 tys. zł	5 tys. zł	5
TRENING			
Kontynuowany pretrening, 50 mld tokenów	0 zł przydział publiczny	19 tys. zł	15
Próby wstępne i przebiegi nieudane	5 tys. zł	40 tys. zł	8
POST-TRENING			
Obliczenia dostrojenia	1 tys. zł	4 tys. zł	9
Dane instrukcyjne — synteza i praca własna	40 tys. zł	150 tys. zł	9
Anotacja ekspercka w dziedzinie regulowanej	—	600 tys. zł	9
SPRAWDZENIE I UDOSTĘPNIENIE			
Ewaluacja i zbiory testowe	20 tys. zł	90 tys. zł	10
Testy odpornościowe	30 tys. zł	200 tys. zł	10
Dokumentacja zgodności	25 tys. zł	80 tys. zł	14
LUDZIE I EKSPLOATACJA			
Zespół — 3 do 5 osób przez 6–9 miesięcy	420 tys. zł	1 400 tys. zł	3
Serwowanie modelu — rok	120 tys. zł	530 tys. zł	11
Razem pierwszy rok	677 tys. zł	3 198 tys. zł	—

Wariant oszczędny zakłada uzyskanie bezpłatnego przydziału mocy obliczeniowej, dane instrukcyjne budowane siłami własnymi i syntetycznie oraz serwowanie modelu małego. Wariant pełny zakłada zakup mocy na rynku i anotację przez specjalistów dziedzinowych.

GDZIE W TYM BUDŻECIE LEŻY CIĘŻAR

Moc obliczeniowa to od zera do dwóch procent całości. Ponad połowę pochłania zespół, a przy wariantcie pełnym drugą co do wielkości pozycją jest anotacja ekspercka. *To nie jest przedsięwzięcie sprzętowe, tylko kadrowe* — i tak samo wygląda struktura kosztów u największych.

Scenariusz C — model ośmiomiliardowy od zera

Pełny pretrening na piętnastu bilionach tokenów, własna architektura, własny tokenizator, pełny post-trening. Model porównywalny z otwartymi modelami średniej klasy z lat 2024–2025.

TABELA 12.3 — SCENARIUSZ C

POZYCJA	KOSZT (MLN ZŁ)	UDZIAŁ
Przebieg finalny pretreningu	3,4–6,6	18 %
Próby wstępne, badania wpływu decyzji, przebiegi nieudane	6,8–13,2	36 %

POZYCJA	KOSZT (MLN ZŁ)	UDZIAŁ
Pozyskanie i przygotowanie danych	0,4–1,2	3 %
Post-trening z anotacją	1,9–11,0	20 %
Ewaluacja, bezpieczeństwo, zgodność	0,3–1,3	3 %
Zespół 8–11 osób przez 12–18 miesięcy	2,0–4,1	17 %
Serwowanie w pierwszym roku	0,5–1,0	3 %
Razem	15,3–38,4	100 %

Mnożnik trzykrotny dla prób wstępnych i przebiegów nieudanych przyjęty za dolną granicę przedziału podawanego w literaturze branżowej (trzy- do sześciokrotności przebiegu finalnego). Zespół, który robi to po raz pierwszy, powinien przyjąć górną granicę.

Scenariusz D — model klasy czołowej

Model porównywalny z tym, co w 2026 roku oferują komercyjnie największe laboratoria. Podawany dla porządku, nie jako opcja.

TABELA 12.4 — SCENARIUSZ D

POZYCJA	KOSZT (MLN USD)
Przebieg finalny pretreningu	200–500
Badania i rozwój — trzy- do sześciokrotność przebiegu	600–3 000
Post-trening: dane, środowiska, uczenie ze wzmocnieniem	300–1 500
Zespół badawczy	100–400 rocznie
Ewaluacja, bezpieczeństwo, zgodność regulacyjna	20–80
Razem program	1 200–5 500

Kwoty złożone z ujawnień częściowych; żadne laboratorium nie publikuje pełnego rachunku. Koszt przebiegu czołowego rośnie 2,4-krotnie rocznie, co ekstrapoluje się do przekroczenia miliarda dolarów za pojedynczy przebieg około 2027 r.

Cztery scenariusze obok siebie

A — laptop		150 zł
B — adaptacja		0,7–3,2 mln zł
C — 8 mld od zera		15–38 mln zł
D — klasa czołowa		4,5–20 mld zł

Skala tego wykresu jest logarytmiczna w treści, choć liniowa w rysunku — między pierwszym a ostatnim wierszem leży osiem rzędów wielkości. Warto to widzieć naraz, bo publiczna dyskusja o budowie własnych modeli miesza te scenariusze, traktując je jako warianty tego samego przedsięwzięcia. Nie są. Scenariusz A to weekend, scenariusz D to zakład przemysłowy.

WNIOSEK, WOKÓŁ KTÓREGO ZORGANIZOWANE JEST CAŁE OPRACOWANIE

Dla polskiej firmy technologicznej realnym przedziałem jest scenariusz B, a jego wariant oszczędny — poniżej siedemset tysięcy złotych w pierwszym roku — mieści się w budżecie jednego dobrze finansowanego projektu

produktowego. Scenariusz C jest osiągalny dla spółki po rundzie finansowania, ale wymaga uzasadnienia, którego samo „chcemy własny model” nie dostarcza. Scenariusz D nie jest dostępny dla nikogo w Polsce i nie będzie w dającej się przewidzieć przyszłości.

ROZDZIAŁ 13 · RACHUNEK

Harmonogram i ryzyka

Budżet mówi, ile to kosztuje. Harmonogram mówi, kiedy pieniądze wychodzą, a rejestr ryzyk — dlaczego wychodzą w większej ilości, niż zaplanowano. Rozdział zamyka listą najczęstszych przyczyn niepowodzenia, spisana z publicznych relacji zespołów, którym się nie udało.

Rozkład prac w czasie

Etapy nakładają się na siebie, więc czas kalendarzowy jest krótszy niż suma czasów częściowych. Poniżej realistyczny przebieg dla scenariusza B, czyli adaptacji modelu otwartego.

TABELA 13.1 — SCENARIUSZ B, PRZEBIEG DZIEWIĘCIOMIESIĘCZNY

ETAP	MIESIĄCE	CO MUSI BYĆ GOTOWE NA KONIEC	WYDATEK
Rozeznanie i decyzja	1	Policzony koszt alternatywy na cudzym modelu, wybrana baza, wniosek o przydział mocy	niski
Korpus	1–3	Oczyszczony, odsiany, udokumentowany co do pochodzenia zbior	średni
Tokenizer i próby małej skali	3–4	Własny słownik, zmierzona wydajność kodowania polszczyzny, przebiegi próbne	niski
Kontynuowany pretrening	4–5	Wagi po adaptacji językowej, zmierzona poprawa na polskich zbiorach	wysoki
Dane instrukcyjne	3–7	Zbiór rozmów dziedzinowych, ocenionych i podwójnie sprawdzonych	najwyższy
Post-trening	6–7	Model odpowiadający na polecenia, odmawiający w sytuacjach ryzykownych	średni
Ewaluacja i testy odpornościowe	7–8	Własny zbiór testowy, raport z prób przełamania zabezpieczeń	średni
Dokumentacja i udostępnienie	8–9	Komplet dokumentacji zgodności, uruchomiona usługa	średni

Dla scenariusza C — treningu od zera — ten sam przebieg rozciąga się na dwanaście do osiemnastu miesięcy dla zespołu, który robi to po raz pierwszy, i na sześć do dziewięciu dla zespołu doświadczonego. Różnica nie leży w mocy obliczeniowej, tylko w liczbie prób potrzebnych do pierwszego udanego przebiegu.

REGUŁA PLANOWANIA, KTÓRA SPRAWDZA SIĘ ZAWSZE

Pierwszy przebieg treningu nigdy nie jest udany. Harmonogram, w którym zaplanowano jedno okno na trening właściwy, jest harmonogramem nierealnym. Rozsądnie jest przyjąć trzy podejścia i zabudżetować moc obliczeniową

na wszystkie trzy — dokładnie tak, jak robi to mnożnik trzy- do sześciokrotności obecny w każdym rachunku w rozdziale dwunastym.

Rejestr ryzyk

TABELA 13.2 — RYZYKA UPORZĄDKOWANE WEDŁUG ILOCZYNU PRAWDOPODOBIENSTWA I SKUTKU

RYZYKO	PRAWDOPOD.	SKUTEK	SPOSÓB OGRANICZENIA
Dane okazują się gorsze, niż zakładano	wysokie	poważny	Zbadać jakość korpusu przebiegami próbnymi na modelu stumilionowym, zanim uruchomi się cokolwiek większego.
Brak wiarygodnej ewaluacji	wysokie	poważny	Zbudować własny, tajny zbiór testowy <i>przed</i> treningiem, a nie po nim.
Anotacja ekspercka przekracza budżet	wysokie	umiarkowany	Rozliczać po przykładach, nie po godzinach; zacząć od tysiąca przykładów i zmierzyć efekt.
Odejście osoby z doświadczeniem treningowym	średnie	krytyczny	Dokumentować konfiguracje i decyzje na bieżąco; nie dopuszczać do sytuacji, w której wiedza jest w jednej głowie.
Wnioskowany przydział mocy nie zostaje przyznany	średnie	poważny	Składać wniosek z półrocznym wyprzedzeniem; mieć wycenioną ofertę komercyjną jako plan zapasowy.
Model bazowy zmienia licencję lub znika	niskie	poważny	Zachować lokalną kopię wag i tekstu licencji z datą pobrania.
Zmiana obowiązków regulacyjnych w trakcie prac	średnie	umiarkowany	Prowadzić dokumentację zgodności od pierwszego dnia, nie od dnia udostępnienia. Szerzej w rozdziale 14.

Najczęstsze przyczyny niepowodzenia

Poniższa lista pochodzi z publicznych relacji zespołów, którym się nie udało, oraz z powtarzalnych wzorców opisywanych w literaturze technicznej. Uderzające jest to, że żadna pozycja nie dotyczy braku pieniędzy ani mocy obliczeniowej.

Pisanie własnej biblioteki treningowej. Najbardziej niezawodny sposób na stratę pół roku. Istnieją dojrzałe rozwiązania otwarte, obsługujące równoległość, punkty kontrolne i odporność na awarie węzłów; napisanie tego samemu jest projektem na kilkaset tysięcy linii kodu.

Odstępstwa architektoniczne bez uzasadnienia. Zmiana sprawdzonego rozwiązania na własne jest w tym przedsięwzięciu źródłem ryzyka, nie przewagi. Odstępstwa wprowadza się pojedynczo i po zmierzeniu skutku na małej skali.

Trening bez wcześniejszej ewaluacji. Zespół, który nie umie zmierzyć, czy model jest lepszy, nie dowie się tego po treningu. Ewaluacja jest warunkiem wstępnym, nie etapem końcowym.

Zbyt duży model przy zbyt małych danych. Klasyczny błąd ambicjonalny. Przy dostępnym korpusie polskim sufit wielkości modelu wynika z arytmetyki opisaną w rozdziale czwartym i nie da się go przekroczyć budżetem.

Zbyt późne uruchomienie potoku danych. Przygotowanie korpusu zajmuje więcej czasu niż trening. Zespoły, które zaczynają od konfiguracji klastra, zwykle czekają potem miesiąc na dane.

Pominięcie dokumentacji pochodzenia danych. Odtworzenie po fakcie, skąd wziął się każdy fragment korpusu, bywa niewykonalne — a jest wymagane. Zapisuje się to w trakcie zbierania albo nie zapisuje wcale.

JEDNO ZDANIE, KTÓRE WARTO ZAPAMIĘTAĆ Z TEGO ROZDZIAŁU

Przedsięwzięcia tego typu nie kończą się niepowodzeniem dlatego, że zabrakło kart graficznych. Kończą się niepowodzeniem dlatego, że po ośmiu miesiącach nikt nie potrafi rozstrzygnąć, czy powstały model jest lepszy od tego, który można było pobrać za darmo w pierwszym tygodniu.

CZĘŚĆ IV

Prawo

Od 2 sierpnia 2025 r. twórca modelu ogólnego przeznaczenia jest w Unii adresatem konkretnych obowiązków. Rozdział omawia je z przywołaniem przepisów.

ROZDZIAŁ 14 · PRAWO

Obowiązki prawne twórcy modelu

Stan prawny zmienił się zasadniczo na tydzień przed datą, którą przez dwa lata traktowano jako punkt odniesienia. Rozdział ustala, co obowiązuje 1 sierpnia 2026 roku, co przesunięto, a co — wbrew rozpowszechnionemu przekonaniu — przesunięte nie zostało.

Zmiana z lipca 2026 roku

Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2026/1744 z 8 lipca 2026 r., zmieniające akt o sztucznej inteligencji, opublikowano 24 lipca, a weszło w życie 27 lipca 2026 r. Przebudowało ono art. 113 rozporządzenia 2024/1689 i przesunęło terminy — ale nie te, o których najczęściej się mówi.

DWA BŁĘDY, KTÓRE KOSZTUJĄ NAJWIĘCEJ

Błąd pierwszy: twierdzenie, że 2 sierpnia 2026 r. zaczyna obowiązywać reżim systemów wysokiego ryzyka. Nie zaczyna — terminy przesunięto na 2 grudnia 2027 r. dla systemów z załącznika III i 2 sierpnia 2028 r. dla załącznika I.

Błąd odwrotny, równie kosztowny: przyjęcie, że nowelizacja odroczyła także obowiązki dostawców modeli ogólnego przeznaczenia. *Nie odroczyła.* Rozdział V aktu stosuje się nieprzerwanie od 2 sierpnia 2025 r.; nowelizacja zmieniła w nim wyłącznie art. 56 ust. 6.

TABELA 14.1 — HARMONOGRAM STOSOWANIA PO NOWELIZACJI, STAN NA 1 SIERPNIA 2026 R.

DATA	CO ZACZYNA OBOWIĄZYWAĆ	PODSTAWA
1.08.2024	Wejście w życie rozporządzenia	art. 113 akapit pierwszy

DATA	CO ZACZYNA OBOWIĄZYWAĆ	PODSTAWA
2.02.2025	Rozdziały I i II — przepisy ogólne i zakazane praktyki, kompetencje w zakresie AI	art. 113 akapit trzeci lit. a); art. 4
2.08.2025	Obowiązki dostawców modeli ogólnego przeznaczenia — rozdział V, art. 51–56. Bez zmian.	art. 113 akapit trzeci lit. b)
27.07.2026	Wejście w życie nowelizacji: nowe brzmienie art. 4, art. 50 ust. 7 i art. 56 ust. 6, nowy art. 4a, nowe art. 75 i nast.	art. 4 rozporządzenia 2026/1744
2.08.2026	Obowiązki przejrzystości, w tym oznaczanie treści generowanych; kary Komisji wobec dostawców modeli	art. 50; art. 101
2.12.2026	Nowe zakazane praktyki; koniec okresu przejściowego na oznaczanie treści dla systemów już na rynku	art. 5 ust. 1 lit. ba), bb); art. 111 ust. 4
2.08.2027	Modele ogólnego przeznaczenia wprowadzone do obrotu przed 2.08.2025 — termin dostosowania	art. 111 ust. 3
2.12.2027	Systemy wysokiego ryzyka z załącznika III — <i>termin przesunięty</i>	art. 113 po zmianie
2.08.2028	Systemy wysokiego ryzyka z załącznika I — <i>termin przesunięty</i>	art. 113 po zmianie

Cztery obowiązki dostawcy modelu ogólnego przeznaczenia

Artykuł 53 ust. 1 nakłada na dostawcę cztery obowiązki podstawowe, obowiązujące od 2 sierpnia 2025 roku. Litera a) — sporządzenie i aktualizowanie dokumentacji technicznej modelu wraz z opisem procesu trenowania i testowania, obejmującej co najmniej elementy z załącznika XI, oraz udostępnianie jej Urzędowi do spraw Sztucznej Inteligencji na żądanie. Litera b) — przygotowanie informacji dla dostawców, którzy będą model integrować, zgodnie z załącznikiem XII. Litera c) — wprowadzenie polityki przestrzegania prawa autorskiego, w tym poszanowania zastrzeżeń dotyczących eksploracji tekstów i danych. Litera d) — sporządzenie i podanie do publicznej wiadomości wystarczająco szczegółowego streszczenia treści wykorzystanych do trenowania.

Ten ostatni obowiązek ma gotowy wzór: Urząd do spraw Sztucznej Inteligencji opublikował 24 lipca 2025 r. objaśnienia wraz ze wzorem streszczenia i dokumentem pytań i odpowiedzi, w dwudziestu czterech językach urzędowych. Nie ma zatem podstaw do twierdzenia, że wymóg jest nieprecyzyjny — jest sformalizowany.

Osobno funkcjonuje kodeks praktyk dla modeli ogólnego przeznaczenia, opublikowany 10 lipca 2025 r. i podpisany przez dwudziestu jeden dostawców. Warto rozwiać rozpowszechnione nieporozumienie: *przystąpienie do kodeksu nie daje domniemania zgodności* — stwierdza to wprost motyw 41 nowelizacji, która przy okazji usunęła upoważnienie Komisji do zatwierdzania kodeksów aktem wykonawczym. Domniemanie zgodności daje wyłącznie europejska norma zharmonizowana. Kodeks pozostaje natomiast użytecznym punktem odniesienia przy wykazywaniu należytej staranności.

Ryzyko systemowe i próg mocy obliczeniowej

Model kwalifikuje się jako obciążony ryzykiem systemowym, jeżeli ma zdolności dużego oddziaływania albo jeżeli tak postanowi Komisja. Artykuł 51 ust. 2 wprowadza domniemanie: przyjmuje się, że model ma takie

zdolności, gdy łączna moc obliczeniowa użyta do jego trenowania przekracza 10^{25} operacji zmiennoprzecinkowych.

TRZY RZECZY, KTÓRE WARTO O TYM PROGU WIEDZIEĆ

Po pierwsze, jest to **domniemanie wzruszalne** — dostawca może je obalić uzasadnionymi argumentami na podstawie art. 52 ust. 2. Nie jest to definicja.

Po drugie, to **dostawca zawiadamia Komisję** w terminie dwóch tygodni od spełnienia progu albo od powzięcia wiadomości, że zostanie spełniony — nie odwrotnie. Odwrócenie tej relacji jest częstym błędem w opracowaniach wtórnych.

Po trzecie, dla porównania ze skalą z rozdziału ósmego: próg 10^{25} przekracza model czterystumiliardowy trenowany na piętnastu bilionach tokenów ($3,6 \times 10^{25}$). Model ośmiomiliardowy osiąga $7,2 \times 10^{23}$, a więc leży ponad rząd wielkości poniżej progu. *Żaden projekt opisany w scenariuszach A, B i C tego opracowania nie zbliża się do tej granicy.*

Dostawca modelu obciążonego ryzykiem systemowym ma obowiązki dodatkowe z art. 55: ocenę modelu według znormalizowanych protokołów wraz z udokumentowanymi testami kontradyktoryjnymi, ocenę i ograniczanie ryzyk systemowych, śledzenie i zgłaszanie poważnych incydentów oraz zapewnienie odpowiedniego poziomu cyberbezpieczeństwa modelu i infrastruktury.

Wyjątek dla modeli otwartych i jego granice

Artykuł 53 ust. 2 zwalnia dostawców modeli udostępnianych na wolnej i otwartej licencji — umożliwiającej dostęp, wykorzystywanie, modyfikowanie i dystrybucję, przy publicznie dostępnych parametrach — z obowiązków z liter a) i b). Granice tego wyjątku bywają pomijane, a są istotne.

Wyjątek *nie obejmuje* litery c) ani d): polityka przestrzegania prawa autorskiego i publiczne streszczenie treści treningowych obowiązują zawsze, także model w pełni otwarty. Wyjątek nie działa również wobec modeli obciążonych ryzykiem systemowym ani — zgodnie z motywem 103 — wobec modeli monetyzowanych.

Kiedy dotrenowanie cudzego modelu czyni z nas dostawcę

To pytanie ma bezpośrednie znaczenie dla scenariusza rekomendowanego w rozdziale siedemnastym. Motyw 109 wskazuje, że w przypadku modyfikacji lub dostrojenia modelu obowiązki dostawcy powinny ograniczać się do samej modyfikacji — na przykład przez uzupełnienie istniejącej dokumentacji o informacje o wprowadzonych zmianach.

Wytyczne Komisji dotyczące zakresu obowiązków dostawców modeli ogólnego przeznaczenia doprecyzowują to progiem ilościowym: podmiot dotrenowujący staje się dostawcą, gdy moc obliczeniowa użyta do modyfikacji przekracza jedną trzecią mocy użytej do wytrenowania modelu pierwotnego. Dopiero wtedy, gdy tej wartości nie da się poznać ani oszacować, stosuje się regułę zastępczą odwołującą się do ułamka progów bezwzględnych. Należy odnotować, że *próg ten pochodzi z wytycznych, a nie z tekstu rozporządzenia* — jest instrumentem miękkim, wyznaczającym praktykę Urzędu, a nie normą, na którą można się powołać przed sądem z pewnością wiążącego skutku. Warto też zweryfikować sygnaturę i datę egzemplarza wytycznych przed powołaniem się na nie: strona Komisji i sam plik podają rozbieżne daty.

UWAGA WARSZTATOWA PRZY CYTOWANIU

Na 1 sierpnia 2026 r. w bazie EUR-Lex **nie ma wersji skonsolidowanej** aktu o sztucznej inteligencji uwzględniającej nowelizację — jedyna dostępna nosi datę 12 lipca 2024 r. Powołując się na art. 4, 4a, 50 ust. 7, 56 ust. 6, 75, 99, 111 i 113, trzeba zatem odsyłać bezpośrednio do rozporządzenia 2026/1744, a nie do tekstu pierwotnego.

Kary

TABELA 14.2 — SANKCJE I DATY ICH STOSOWANIA

NARUSZENIE	KARA	OD KIEDY	PODSTAWA
Zakazane praktyki z art. 5	do 35 mln EUR lub 7 % obrotu	2.08.2025	art. 99 ust. 3
Naruszenia dostawcy modelu ogólnego przeznaczenia	do 15 mln EUR lub 3 % obrotu	2.08.2026	art. 101
Naruszenie obowiązku kompetencji w zakresie AI (art. 4)	brak kary pieniężnej	—	katalog art. 99 ust. 4

Kwota wyższa z dwóch wskazanych. Warto odnotować, że do 1 sierpnia 2026 r. włącznie obowiązki z art. 53 i 55 obowiązywały bez sankcji pieniężnej — art. 101 był wyłączony z pakietu stosowanego od 2 sierpnia 2025 r. Nadzór i egzekucja wobec modeli ogólnego przeznaczenia należą wyłącznie do Komisji działającej przez Urząd do spraw Sztucznej Inteligencji (art. 88 ust. 1), a nie do organów krajowych.

Polska ustawa wdrażająca

Prace legislacyjne są zakończone. Obowiązuje ustawa z dnia 3 lipca 2026 r. o systemach sztucznej inteligencji (Dz.U. z 2026 r. poz. 1003), ogłoszona 27 lipca 2026 r. Wchodzi w życie stopniowo: zasadnicza część 11 sierpnia 2026 r., natomiast art. 8–18 oraz rozdziały 3–5, 8 i 9 — dopiero 28 października 2026 r.

Organem nadzoru rynku jest Komisja Rozwoju i Bezpieczeństwa Sztucznej Inteligencji: przewodniczący powoływany przez Sejm za zgodą Senatu na pięcioletnią kadencję, dwóch zastępców i czterech członków delegowanych przez organy nadzoru finansowego, ochrony konkurencji, komunikacji elektronicznej i mediów. Przewodniczący ma być powołany w ciągu dwóch miesięcy od wejścia ustawy w życie, pełny skład w ciągu trzech — co realistycznie oznacza działający organ w listopadzie 2026 roku. *Do tego czasu nie ma w Polsce organu zdolnego wydać decyzję.*

TRZY SZCZEGÓŁY PROCEDURALNE WARTO ZAPAMIĘTANIA

Odwołanie od decyzji Komisji rozpoznaje **Sąd Okręgowy w Warszawie — sąd ochrony konkurencji i konsumentów**, a nie sąd administracyjny.

Równowartość kar wyrażonych w euro przelicza się według średniego kursu Narodowego Banku Polskiego **z dnia 28 stycznia danego roku** — nie z dnia decyzji ani naruszenia.

Do postępowania w sprawie administracyjnych kar pieniężnych wyłączono art. 189f Kodeksu postępowania administracyjnego: nie ma odstąpienia od nałożenia kary na zasadach ogólnych, a jedynym mechanizmem łagodzącym jest obniżenie kary przewidziane w samej ustawie.

Komisja prowadzi także piaskownice regulacyjne, bezpłatne dla mikroprzedsiębiorców oraz małych i średnich przedsiębiorstw, wydaje opinie indywidualne i rozpatruje skargi.

Prawo autorskie i dane treningowe

Warstwa prawnoautorska jest starsza i lepiej osadzona niż akt o sztucznej inteligencji. Dyrektywa 2019/790 przewiduje dwa wyjątki na eksplorację tekstów i danych. Artykuł 3 obejmuje organizacje badawcze i instytucje dziedzictwa kulturowego, do celów badań naukowych, i nie może zostać wyłączony umownie — przesądza o tym art. 7 ust. 1 dyrektywy. Artykuł 4 ma charakter ogólny i dopuszcza cele komercyjne, ale ustępuje wobec zastrzeżenia uprawnionego. *Oba wymagają zgodnego z prawem dostępu do treści* i pominięcie tej przesłanki zniekształca obraz drugiego z nich jako wyjątku swobodnego.

Warto przy tym uważać na często powtarzany skrót myślowy. Artykuł 4 ust. 3 dyrektywy wymaga zastrzeżenia „w stosowny sposób, na przykład za pomocą środków nadających się do odczytu maszynowego w przypadku treści udostępnionych publicznie online” — forma maszynowa jest tu *przykładem*, a nie bezwzględnym wymogiem. Bezwzględny obowiązek formatu odczytywalnego maszynowo wprowadza dopiero przepis polski.

Polską implementacją jest ustawa z dnia 26 lipca 2024 r. o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi (Dz.U. 2024 poz. 1254), ogłoszona 19 sierpnia i obowiązująca od 20 września 2024 r., z jednym przepisem wchodzącym w życie dopiero 20 lutego 2025 r. Dodała ona do ustawy o prawie autorskim dwa przepisy.

Artykuł 26² odpowiada art. 3 dyrektywy: obejmuje instytucje dziedzictwa kulturowego oraz podmioty wskazane w art. 7 ust. 1 pkt 1, 2 i 4–8 Prawa o szkolnictwie wyższym i nauce, przy czym zwielokrotnianie musi służyć badaniom naukowym i *nie może być podejmowane w celu osiągnięcia bezpośredniej lub pośredniej korzyści majątkowej*. Artykuł 26³ odpowiada art. 4: nie ogranicza celu ani podmiotu, ale dotyczy wyłącznie utworów *rozpowszechnionych* — tym kryterium polski ustawodawca zastąpił dyrektywalną przesłankę zgodnego z prawem dostępu. Zastrzeżenia dokonuje się wyraźnie i odpowiednio do sposobu udostępnienia, a dla treści dostępnych publicznie w wybranym miejscu i czasie — w formacie przeznaczonym do odczytu maszynowego. Zwielokrotnione utwory wolno przechowywać wyłącznie w celu eksploracji i tylko tak długo, jak to konieczne.

KONSEKWENCJA PRAKTYCZNA DLA BUDUJĄCEGO KORPUS

Podmiot komercyjny działa na podstawie art. 26³, a więc *musi respektować zastrzeżenia*. Oznacza to obowiązek odczytywania i honorowania maszynowych zastrzeżeń przy pobieraniu treści oraz — co ważniejsze — obowiązek udokumentowania, że się to robiło. Odtworzenie tego po fakcie jest zwykle niewykonalne, a art. 53 ust. 1 lit. c) aktu o sztucznej inteligencji wymaga posiadania odpowiedniej polityki.

Orzecznictwo

Sądy orzekają szybciej, niż zmienia się regulacja. W sprawie Bartz przeciwko Anthropic sąd federalny dla Północnego Dystryktu Kalifornii udzielił 20 lipca 2026 r. ostatecznego zatwierdzenia ugody opiewającej na 1,5 miliarda dolarów wraz z odsetkami — największej w historii amerykańskiego prawa autorskiego — obejmującej uprawnionych do utworów pobranych z nielegalnych repozytoriów, wraz z obowiązkiem zniszczenia plików i kopii pochodnych. Szacowana wypłata to około trzech tysięcy dolarów na utwór, czyli czterokrotność ustawowego minimum za naruszenie umyślne. Zgłoszeniami roszczeń objęto 91,3 procent pozycji z wykazu utworów według stanu na 16 kwietnia 2026 r. — nie należy tego mylić z zasięgiem kampanii

zawiadomień, który wyniósł 99,5 procent. Uгода nie obejmuje roszczeń opartych na wynikach generowanych przez model.

Istotna jest linia podziału przyjęta w tej sprawie. Postanowieniem z 23 czerwca 2025 r. sąd uznał trenowanie na legalnie nabytych egzemplarzach — a także destrukcyjne skanowanie kupionych książek — za dozwolony użytek o charakterze wysoce przekształcającym, natomiast pobranie ponad siedmiu milionów pirackich egzemplarzy do stałej biblioteki wewnętrznej za dozwolonym użytkowaniem nieobjęte. To rozróżnienie — *legalność źródła, nie legalność treningu* — jest dziś najważniejszą wskazówką praktyczną płynącą z orzecznictwa.

W Wielkiej Brytanii High Court rozstrzygnął 4 listopada 2025 r. spór Getty Images przeciwko Stability AI. Wygrana powoda jest znacznie węższa, niż sugerują doniesienia prasowe: naruszenie stwierdzono wyłącznie na podstawie sekcji 10(1) i 10(2) ustawy o znakach towarowych z 1994 r., i to tylko co do historycznych wersji modelu odtwarzających znak wodny agencji; roszczenie z sekcji 10(3) oddalono. Główne roszczenia autorskie powód porzucił w toku procesu, a co do naruszenia wtórnego sąd orzekł na korzyść pozwanego, przyjmując, że wagi modelu nie są egzemplarzem naruszającym w rozumieniu ustawy o prawie autorskim. Zgody na apelację udzielono 16 grudnia 2025 r. odrębnym orzeczeniem; styczeń 2026 r. to jedynie data upublicznienia. Pozwanemu odmówiono przy tym zgody na apelację w zakresie znaków towarowych.

W sprawie New York Times przeciwko OpenAI sąd dla Południowego Dystryktu Nowego Jorku oddalił 4 kwietnia 2025 r. wniosek o umorzenie w przeważającej części, pozostawiając w sprawie roszczenia oparte na wynikach modelu; postępowania skonsolidowano, a nakaz wydania około dwudziestu milionów zapisów rozmów użytkowników został utrzymany w styczniu 2026 r. Dwie inne sprawy bywają przedstawiane jako rozstrzygnięte, choć nie są: w postępowaniu Kadrey przeciwko Meta nierozpoznane pozostaje roszczenie dotyczące rozpowszechniania utworów w sieci wymiany plików, a wyrok w sprawie Thomson Reuters przeciwko Ross został zaskarżony i na 1 sierpnia 2026 r. sąd odwoławczy nie orzekł.

Ochrona danych osobowych

To warstwa najmniej stabilna. Opinia 28/2024 Europejskiej Rady Ochrony Danych z 17 grudnia 2024 r. przesądziła, że model wytrenowany na danych osobowych *nie może być z założenia uznany za anonimowy* — anonimowość wymaga, by prawdopodobieństwo wydobycia danych osób, których dane posłużyły do trenowania, w tym poprzez zapytania do modelu, było znikome. Powołanie się na prawnie uzasadniony interes wymaga przeprowadzenia i udokumentowania testu trójstopniowego.

Tu przydaje się rozróżnienie, które w doniesieniach prasowych bywa zacierane. Pakiet upraszczający z listopada 2025 r. składał się z dwóch bliźniaczych wniosków: pierwszy dotyczył aktu o sztucznej inteligencji i *został już przyjęty* jako omawiane wyżej rozporządzenie 2026/1744; drugi, obejmujący ochronę danych — w tym projektowaną wyraźną podstawę przetwarzania na potrzeby trenowania modeli wraz z bezwarunkowym prawem sprzeciwu — pozostaje w negocjacjach. Odrębnie i na dalszym etapie procedowana jest jeszcze jedna zmiana, dotycząca obowiązku prowadzenia rejestru czynności przetwarzania.

Stan na 1 sierpnia 2026 r. jest zatem taki, że ogólne rozporządzenie o ochronie danych **nie zostało dotąd znowelizowane** — przy tym przepisie w bazie EUR-Lex figurują wyłącznie sprostowania.

Warto też odnotować regułę, która przesądza status podmiotu przetwarzającego: procesor, który wykorzysta powierzone mu dane do trenowania własnego modelu, wykracza poza polecenie administratora i z

mocy art. 28 ust. 10 rozporządzenia o ochronie danych staje się administratorem w odniesieniu do tego celu — ze wszystkimi wynikającymi z tego obowiązkami.

TRENOWANIE NA AKTACH KANCELARII — BARIERA DALEJ IDĄCA NIŻ OCHRONA DANYCH

Materiał z dokumentacji kancelarii to co do zasady dane osobowe, w tym niemal zawsze szczególne kategorie oraz dane dotyczące wyroków. Niezależnie od reżimu ochrony danych obowiązuje jednak bariera mocniejsza: tajemnica zawodowa. Adwokat jest obowiązany zachować w tajemnicy wszystko, o czym dowiedział się w związku z udzielaniem pomocy prawnej; obowiązek nie jest ograniczony w czasie, a art. 6 ust. 4 Prawa o adwokaturze zna tylko dwa wyłączenia — przepisy o przeciwdziałaniu praniu pieniędzy i finansowaniu terroryzmu oraz rozdział 11a działu III Ordynacji podatkowej. Odrębnym mechanizmem, którego nie należy stawiać z nimi w jednym szeregu, jest art. 180 § 2 Kodeksu postępowania karnego; tajemnica obrończa pozostaje bezwzględna.

Po stronie radców prawnych stan zmieniła ustawa z 15 maja 2026 r. (Dz.U. 2026 poz. 731), obowiązująca od 18 czerwca 2026 r., zgodnie z którą obowiązek zachowania tajemnicy nie ustaje, gdy z żądaniem ujawnienia występuje organ nadzorczy w rozumieniu przepisów o ochronie danych.

Trenowanie modelu na aktach klienta na własne cele dostawcy jest zatem drogą zamkniętą. Otwarte pozostają trzy: powierzenie przetwarzania z umownym i technicznym zakazem uczenia się na danych, materiał skutecznie zanonimizowany albo syntetyczny, oraz model wdrożony lokalnie u samego administratora. Bezpiecznym źródłem korpusu pozostaje materiał publiczny — orzecznictwo i akty normatywne.

Oznaczanie treści generowanych

Artykuł 50 stosuje się od 2 sierpnia 2026 r. i nie został odroczony. Dostawca systemu generującego treść syntetyczną — dźwięk, obraz, wideo albo tekst — ma obowiązek oznakować wynik w formacie odczytywalnym maszynowo, a system przeznaczony do bezpośredniej interakcji z człowiekiem musi być zaprojektowany tak, żeby osoba wiedziała, że rozmawia z maszyną. Dla systemów wprowadzonych do obrotu przed 2 sierpnia 2026 r. nowelizacja dodała okres przejściowy do 2 grudnia 2026 r.

Komplet dokumentów przed udostępnieniem modelu

LISTA KONTROLNA

1. Dokumentacja techniczna modelu według załącznika XI — zadania modelu, dopuszczalne zasady wykorzystania, architektura, liczba parametrów, sposób trenowania i testowania, zużyte zasoby obliczeniowe.
2. Informacje dla dostawców integrujących model, według załącznika XII.
3. Polityka przestrzegania prawa autorskiego, obejmująca sposób respektowania maszynowych zastrzeżeń przy pozyskiwaniu danych.
4. Publiczne streszczenie treści treningowych, sporządzone na wzorze opublikowanym 24 lipca 2025 r.
5. Dokumentacja pochodzenia danych — prowadzona *od pierwszego dnia zbierania*, nie odtwarzana na końcu.
6. Jeżeli model jest udostępniany jako system generujący treści: rozwiązanie znakujące wyniki w formacie odczytywalnym maszynowo oraz informowanie o interakcji z systemem.
7. Dokumentacja oceny wpływu na ochronę danych, jeżeli w korpusie znalazły się dane osobowe, wraz z udokumentowanym testem prawnie uzasadnionego interesu.

Stan prawny ustalony na 1 sierpnia 2026 r. w oparciu o teksty pierwotne oraz niezależną weryfikację przeprowadzoną na potrzeby tego opracowania. Zastrzeżenie: opracowanie ma charakter informacyjny i nie stanowi opinii prawnej w rozumieniu przepisów o świadczeniu pomocy prawnej. Kwalifikacja konkretnego modelu i konkretnego korpusu wymaga odrębnej analizy.

Decyzja

Trening od zera jest jedną z pięciu możliwych dróg i prawie nigdy najlepszą. Ta część porównuje je kosztowo i kończy się rekomendacją.

ROZDZIAŁ 15 · DECYZJA

Pięć dróg i ich ceny

Między korzystaniem z cudzego modelu a trenowaniem własnego od zera leży pięć wyraźnie oddzielonych szczebli, różniących się kosztem o trzy rzędy wielkości. Najczęstszy błąd decyzyjny polega na przeskakiwaniu szczebli — w obie strony.

TABELA 15.1 — PIĘĆ SZCZEBLI DRABINY

SZCZEBEL	NA CZYM POLEGA	KOSZT MOCY	KIEDY JEST WŁAŚCIWYM WYBOREM
I	Polecenie systemowe i baza wiedzy na cudzym modelu	kilka–kilkanaście tys. zł miesięcznie	Problemem jest wiedza dziedzinowa i aktualność informacji.
II	Dostrojenie adapterami niskiego rzędu	poniżej 200 zł na przebieg	Problemem jest styl, format odpowiedzi, konsekwencja terminologiczna, wąska klasyfikacja.
III	Pełne dostrojenie wszystkich wag	10–15 tys. zł na przebieg	Zmienia się coś strukturalnego — słownik, długość kontekstu, format narzędzi.
IV	Kontynuowany pretrening po polsku	10–100 tys. zł	Model źle radzi sobie z polską morfologią, nieefektywnie tokenizuje polszczyznę, nie zna polskiego kontekstu.
V	Trening od zera	1,6–13 mln zł	Istnieje powód architektoniczny albo licencyjny — nie „lepsze odpowiedzi po polsku”.

Szczebel I — policz to najpierw

Zanim ktokolwiek policzy koszt treningu, powinien policzyć, ile kosztowałby ten sam wolumen na cudzym modelu przez trzy lata. Dla realistycznego zastosowania prawniczego — sto tysięcy zapytań miesięcznie, osiem tysięcy tokenów wejścia i sześćset tokenów odpowiedzi na modelu klasy średniej — rachunek wynosi około 3300 dolarów miesięcznie, czyli mniej więcej 12 350 złotych.

Dwie dźwignie potrafią ten rachunek istotnie obniżyć. Pamięć podręczna zapytań wycenia odczyt stałego prefiksu na około jedną dziesiątą ceny bazowej; przy dziewięćdziesięcioprocentowej trafności bufora efektywna cena stałego kontekstu spada do jednej piątej, co w powyższym przykładzie sprowadza rachunek do około 7060 złotych — obniżka o czterdzieści trzy procent. Druga dźwignia to tryb wsadowy z pięćdziesięcioprocentowym rabatem, sensowny wszędzie tam, gdzie odpowiedź nie musi przyjść natychmiast.

LICZBA, OD KTÓREJ TRZEBA ZACZAĆ

Trzy lata działania powyższej usługi na cudzym modelu to około 250 tysięcy złotych — *mniej niż jeden porządny przebieg treningu małego modelu od zera*. Każda decyzja o wejściu wyżej na drabinę musi się z tą liczbą zmierzyć.

Szczebel II — najbardziej niedoceniany

Dostrojenie adapterami modyfikującymi ułamek procenta wag jest zaskakująco tanie. Model siedmio- lub ośmiomiliardowy dostraja się w dwie do czterech godzin na jednej karcie za trzy do dziesięciu dolarów; ten sam zabieg na karcie konsumenckiej trwa sześć do ośmiu godzin i kosztuje niecałe trzy dolary. Model siedemdziesiętomiliardowy — osiem do dwunastu godzin i piętnaście do trzydziestu dolarów.

Sekret niskich kosztów tkwi w pamięci. Model siedmiomiliardowy, który w pełnej precyzji potrzebuje czterestu gigabajtów, po kwantyzacji do czterech bitów mieści się w pięciu-sześciu; model siedemdziesiętomiliardowy schodzi ze stu czterdziestu gigabajtów do czterdziestu sześciu, czyli na jedną kartę. Ceną jest jakość: metoda kwantyzowana osiąga typowo osiemdziesiąt do dziewięćdziesięciu procent wyniku pełnego dostrojenia, sama metoda adapterowa dziewięćdziesiąt do dziewięćdziesięciu pięciu.

GDZIE NAPRAWDĘ IDZIE BUDŻET TEGO SZCZEBLA

Danych trzeba zaskakująco mało: trzy do dziesięciu tysięcy starannie dobranych par polecenie–odpowiedź daje więcej niż sto tysięcy zebranych byle jak. Przy tych cenach karta graficzna przestaje być pozycją kosztową. Wąskim gardłem jest przygotowanie i oczyszczenie zbioru oraz zbudowanie wiarygodnej ewaluacji — i tam idzie dziewięćdziesiąt procent budżetu, liczonego w roboczogodzinach, nie w godzinach karty.

Narzędziowo standardem są cztery pakiety: Unsloth, skracający czas treningu dwu- do pięciokrotnie i najlepszy przy jednej karcie; Axolotl z konfiguracją w plikach tekstowych, wygodny do powtarzalnych potoków; torchtune, natywny dla biblioteki PyTorch i czytelny jako kod referencyjny; oraz LLaMA-Factory o najszerszej obsłudze architektur.

Szczebel III — rzadko się opłaca

Pełne dostrojenie modelu siedmiomiliardowego wymaga stu do stu dwudziestu gigabajtów pamięci, bo do wag dochodzą gradienty i stan optymalizatora. Oznacza to minimum dwie karty najwyższej klasy, a realnie węzeł ośmiokartowy. Czterdziestogodzinny przebieg kosztuje około trzech i pół tysiąca dolarów. Za dziesięć procent tej kwoty metoda adapterowa daje dziewięćdziesiąt do dziewięćdziesięciu pięciu procent efektu — pełne dostrojenie ma zatem sens wyłącznie przy zmianie strukturalnej.

Szczebel IV — najciekawszy dla polskiego podmiotu

Kontynuowany pretrening polega na dołożeniu kilkudziesięciu miliardów tokenów polskiego tekstu do gotowego modelu. Koszt liczy się tym samym wzorem co pretrening od zera, tylko z mniejszą liczbą tokenów.

TABELA 15.2 — KOSZT ADAPTACJI JĘZYKOWEJ MODELU OŚMIOMILIARDOWEGO

OBJĘTOŚĆ KORPUSU	GODZINY KARTY	KOSZT (USD)	KOSZT (PLN)
30 mld tokenów	≈ 1 150	2 300–3 100	8,6–11,6 tys.
50 mld tokenów	≈ 1 900	4 000–5 000	15–19 tys.

Wyliczenie z wzoru 6ND przy wykorzystaniu mocy 35 % i cenie 2,0–2,7 USD za godzinę-kartę. To są kwoty mieszczące się w budżecie marketingowym średniej firmy — i jest to najważniejsza liczba w tym rozdziale.

Szczebel V — i ile naprawdę kosztuje

Trening małego modelu od zera jest wyraźnie droższy, choć nie kosmiczny. Model trzymiliardowy wytrenowany na jedenastu bilionach tokenów pochłonął około 220 tysięcy godzin karty, czyli 1,65–2,2 miliona złotych. Model ośmiomiliardowy z rodziny Llama 3 — 1,3 miliona godzin, czyli 9,7–13,1 miliona złotych. I to wyłącznie przebieg finalny, bez prób wstępnych, nieudanych startów i kosztu zespołu.

REGUŁA DECYZYJNA

Szczebel piąty ma sens wyłącznie wtedy, gdy istnieje powód architektoniczny albo licencyjny — inny tokenizator, inna licencja bazowa, wymogi zgodności co do pochodzenia wag. *Nie ma sensu wtedy, gdy chodzi o jakość odpowiedzi po polsku.* Dowód empiryczny jest polski i przedstawia go następny rozdział: model jedenastomiliardowy zbudowany na cudzej bazie za około dziewięć tysięcy godzin karty przewyższa na polskich testach modele dwu- do sześciokrotnie większe.

ROZDZIAŁ 16 · DECYZJA

Polskie punkty odniesienia

Dwa polskie projekty pokazują dwie skrajnie różne drogi do tego samego celu — jeden powstał za kilka tysięcy złotych gotówki i pracę wolontariuszy, drugi za blisko czterdzieści milionów złotych publicznych. Porównanie ich jest pouczające, bo rozstrzyga, co w tym przedsięwzięciu naprawdę kosztuje.

Bielik — model zbudowany przez społeczność

Pierwsza wersja powstała przez kontynuowany pretrening gotowego modelu francuskiego na ponad siedemdziesięciu miliardach tokenów, z czego trzydzieści sześć miliardów stanowiła polszczyzna przepuszczona dwukrotnie. Trening prowadzono na 256 kartach GH200 superkomputera Helios w Akademickim Centrum Komputerowym Cyfronet, we własnym oprogramowaniu treningowym, przy przepustowości przekraczającej 9200 tokenów na kartę na sekundę.

Z tej przepustowości wynika wprost skala przedsięwzięcia: przebieg trwał około ośmiu godzin zegarowych i pochłonął mniej więcej 2100 godzin pracy kart. Na rynku komercyjnym kosztowałoby to szesnaście do dwudziestu jeden tysięcy złotych. Podawany publicznie budżet gotówkowy wynosił jednak *od jednego do dwóch tysięcy złotych* — różnica bierze się stąd, że moc obliczeniowa pochodziła z grantu, a pracowało nad tym ponad czterysta osób w trybie wolontariackim przez niecałe półtora roku.

TABELA 16.1 — KOLEJNE WERSJE MODELU BIELIK I ICH SKALA

WERSJA	PARAMETRY	TOKENY	SPRZĘT	WARTOŚĆ RYNKOWA MOCY
7B v0.1	7 mld	70 mld	256 × GH200	16–21 tys. zł
11B v2	11 mld	198 mld	256 × GH200	68–92 tys. zł
11B v3	11 mld	—	128 × GH200	—
v3 Small	1,5 i 4,5 mld	292 mld	—	—

Wersja 11B v2 powstała przez pogłębienie modelu bazowego do pięćdziesięciu warstw. Rodzina v3 Small korzysta z własnego tokenizatora o słowniku 32 tysięcy, kodującego polski tekst po 4,78 znaku na token wobec 2,87 dla porównywanego modelu ogólnego — to właśnie przewaga, której nie da się osiągnąć samym dostrojeniem adapterów.

NAJWAŻNIEJSZY WNIOSEK Z TEGO PROJEKTU

Model o jedenastu miliardach parametrów, zbudowany na cudzej bazie za około dziewięć tysięcy godzin pracy kart, przewyższa na polskich testach modele od dwóch do sześciu razy większe. Dowodzi to empirycznie tezy z poprzedniego rozdziału: *o jakości w danym języku decyduje adaptacja, a nie skala pretreningu od zera.*

PLLuM — model finansowany publicznie

Przeciwieństwo pod względem modelu finansowania. Ministerstwo Cyfryzacji przeznaczyło na pierwszą fazę 14,5 miliona złotych, następnie dziewiętnaście milionów na rozwój i wdrożenie w administracji, a bieżąca faza finansowana jest z grantu celowego o wartości blisko sześciu milionów. Konsorcjum tworzyły początkowo Politechnika Wrocławska jako lider, instytuty NASK i OPI, Instytut Podstaw Informatyki Polskiej Akademii Nauk, Uniwersytet Łódzki i Instytut Slawistyki; później dołączyły Centralny Ośrodek Informatyki oraz Cyfronet.

W maju 2026 roku udostępniono jedenaście modeli w rozmiarach cztery, osiem, dwanaście i siedemdziesiąt miliardów parametrów, opartych na dwóch zagranicznych rodzinach bazowych, wytrenowanych na około siedmiu milionach polskich tekstów, z około siedemnastoma i pół tysiąca nowych instrukcji tworzonych ręcznie i ponad sześcioma tysiącami dialogów wielotururowych.

Warto zauważyć proporcję: niecałe czterdzieści milionów złotych na dwa lata pracy siedmiu instytucji dało rodzinę modeli, w której *sam koszt mocy obliczeniowej stanowi prawdopodobnie mniejszość budżetu*. Reszta to ludzie, dane i licencje — dokładnie tak, jak przewiduje rozdział trzeci.

Pozostałe polskie projekty

Obraz domykają dwa wcześniejsze przedsięwzięcia. Rodzina Qra, opracowana wspólnie przez instytut OPI i Politechnikę Gdańską w marcu 2024 roku, to adaptacja modelu amerykańskiego w wariantach jedno-, siedmio- i trzynastomiliardowym, trenowana na dwudziestu jeden kartach superkomputera Kraken; z blisko dwóch terabajtów surowego polskiego tekstu po oczyszczeniu i odsianiu duplikatów została mniej więcej połowa, czyli około dziewięćdziesięciu miliardów tokenów. Model TRURL 2, wydany przez prywatną spółkę w 2023 roku, był dostrojeniem na blisko milionie próbek konwersacyjnych — czyli klasycznym drugim szczeblem drabiny, zrealizowanym komercyjnie.

Poziom europejski

Konsorcjum OpenEuroLLM, obejmujące dwadzieścia instytucji, dysponuje budżetem 37,4 miliona euro. Kluczowa jest jednak nie kwota pieniężna, lecz przydział mocy: ponad dziesięć milionów godzin pracy kart w ramach strategicznego dostępu do europejskiej infrastruktury obliczeniowej. Przy stawce komercyjnej odpowiada to około osiemdziesięciu sześciu milionom złotych — *ponad dwukrotności budżetu pieniężnego projektu*. To dobrze ilustruje, jak w tej dziedzinie wygląda realna struktura finansowania w Europie: moc obliczeniowa bywa większym wkładem niż gotówka.

Co z tego wynika dla polskiej firmy

Najważniejsza informacja dotyczy dostępu do mocy, nie do pieniędzy. Przydział pięćdziesięciu tysięcy godzin pracy kart, dostępny bezpłatnie dla małych i średnich przedsiębiorstw w trybie przyspieszonym, odpowiada wartości rynkowej około 430 tysięcy złotych. Podstawiając do wzoru z rozdziału ósmego: taki przydział pozwala przepuścić przez model ośmiomiliardowy około 1,3 biliona tokenów — *wielokrotnie więcej, niż liczył cały korpus, na którym trenowano pierwszego Bielika*.

ŚCIEŻKA, KTÓRĄ POLSKA FIRMA MOŻE PRZEJŚĆ W 2026 ROKU

Złożyć wniosek o przydział w trybie przyspieszonym, przygotować własny korpus dziedzinowy, przeprowadzić kontynuowany pretrening wybranego modelu otwartego, dołożyć własny post-trening na danych, których nikt inny nie ma. Koszt gotówkowy: praca zespołu przez kilka miesięcy plus kilkanaście tysięcy złotych na moc uzupełniającą. Efekt: własne wagi, własny tokenizator, model działający bez połączenia z internetem. Jest to droga przetarta i udokumentowana — dwa polskie zespoły przeszły ją przed nami.

ROZDZIAŁ 17 · DECYZJA

Co zrobić naprawdę

Wszystkie poprzednie rozdziały opisywały możliwości. Ten wskazuje jedną i uzasadnia, dlaczego akurat ją. Rekomendacja jest sformułowana dla polskiej spółki technologicznej dysponującej dostępem do materiału dziedzinowego, a nie dla laboratorium badawczego.

Odpowiedź na zadane pytanie

Pytanie brzmiało: jak zbudować własny model taki jak ChatGPT. Odpowiedź uczciwa składa się z dwóch części, z których pierwsza jest rozczarowująca, a druga — moim zdaniem — ciekawsza.

Część pierwsza: **modelu klasy ChatGPT nie da się zbudować w Polsce**. Nie z powodu braku umiejętności ani zamknięcia wiedzy — architektura jest publiczna, kod referencyjny ma kilkaset linii, a wszystkie potrzebne narzędzia są dostępne bezpłatnie. Powód jest wyłącznie kapitałowy: program takiego modelu to wydatek rzędu od 1,2 do 5,5 miliarda dolarów, a koszt pojedynczego przebiegu rośnie 2,4-krotnie rocznie. Nie ma w Polsce podmiotu — publicznego ani prywatnego — zdolnego wyłożyć tę kwotę, a gdyby był, ścigałby się z funkcją wykładniczą.

Część druga: **prawie nikt tego nie potrzebuje, łącznie z firmami, które o tym mówią**. Rozdział pierwszy pokazał, że model to pięć z siedmiu warstw produktu, a warstwy szóstą i siódmą — te, które użytkownik

faktycznie ocenia — buduje się zwykłą pracą programistyczną. Rozdział dziewiąty pokazał, że zdolności robiące największe wrażenie powstają w post-treningu, czyli w najtańszym etapie, i decyduje o nich jakość danych. Rozdział szesnasty pokazał polski model, który za dziewięć tysięcy godzin pracy kart wygrywa na polskich testach z modelami sześciokrotnie większymi.

Rekomendacja

SCENARIUSZ B, WARIANT OSZCZĘDNY

Wytrenować własny tokenizator dla polszczyzny, przeprowadzić kontynuowany pretrening otwartego modelu ośmiomiliardowego na przydziale publicznym, a cały wysiłek zespołu skierować na dane dziedzinowe, których nikt inny nie ma. Budżet pierwszego roku: poniżej siedmiuset tysięcy złotych. Rezultat: własne wagi, własny słownik, model działający bez połączenia z internetem.

Uzasadnienie tej rekomendacji da się sprowadzić do czterech obserwacji, z których każda została w tym opracowaniu udokumentowana liczbowo.

Pierwsza dotyczy tokenizatora. Polski tekst kosztuje w słownikach ogólnych od 1,55 do 1,84 raza więcej tokenów niż angielski, a w dziedzinie prawa — najgorzej ze wszystkich rodzajów tekstu. Słownik zbudowany dla polszczyzny bije każdy słownik ogólny przy czterokrotnie mniejszym rozmiarze. Kosztuje kilka godzin pracy jednego komputera i obniża koszt każdego zapytania przez cały okres życia produktu. Jest to najwyższa stopa zwrotu w całym łańcuchu opisanym w tym opracowaniu i nie ma powodu z niej rezygnować.

Druga dotyczy mocy obliczeniowej. Ścieżka przyspieszona europejskiej infrastruktury obliczeniowej daje do pięćdziesięciu tysięcy godzin pracy kart w cztery dni robocze, bezpłatnie dla małych i średnich przedsiębiorstw. To odpowiednik 430 tysięcy złotych i wystarcza na przepuszczenie przez model ośmiomiliardowy około 1,3 biliona tokenów — wielokrotnie więcej, niż potrzeba do pełnej adaptacji językowej. Do tego dochodzą dwie polskie fabryki sztucznej inteligencji i krajowe granty obliczeniowe. Wydawanie własnego kapitału na karty graficzne przed wyczerpaniem tych ścieżek jest po prostu błędem.

Trzecia dotyczy danych i jest najważniejsza. Cała dostępna publicznie polszczyzna wysokiej jakości to 200–300 miliardów tokenów, a jeden z polskich modeli zużył już 237 miliardów — zasób organiczny jest zatem w praktyce wyczerpany. Dalsze różnicowanie może się opierać wyłącznie na materiale, którego nie ma w otwartym internecie: archiwach, orzecznictwie, dokumentacji specjalistycznej. *Kto ma legalny dostęp do dużego, uporządkowanego zbioru polskich tekstów dziedzinowych, dysponuje zasobem rzadszym niż moc obliczeniowa* — i to jest jedyna przewaga w tej dziedzinie, której konkurent nie kupi za pieniądze.

Czwarta dotyczy anotacji. Rynkowa stawka za ocenę odpowiedzi prawnych przez prawnika wynosi 110–130 dolarów za godzinę, a przy specjalistach z wieloletnim doświadczeniem przekracza dwieście. Trzydzieści tysięcy przykładów prawniczych to na rynku wydatek rzędu półtora miliona dolarów. Podmiot, który dysponuje tą kompetencją wewnątrz, ma zatem dostęp do najdroższego składnika post-treningu po koszcie własnym — i jest to przewaga, której nie da się zrównoważyć ani większym klastrem, ani większym zespołem inżynierów.

Czego nie robić

Nie trenować od zera dla jakości po polsku. To nie działa i jest udokumentowane empirycznie: model zbudowany na cudzej bazie wygrywa z modelami wielokrotnie większymi trenowanymi od podstaw. Pretre-

ning odtwarza wiedzę ogólną, którą ktoś już opłacił i udostępnił bezpłatnie.

Nie zaczynać od sprzętu. Zespoły, które zaczynają od konfiguracji klastra, czekają potem miesiąc na dane. Kolejność jest odwrotna: dane, ewaluacja, próby w małej skali, dopiero potem moc obliczeniowa.

Nie pomijać ewaluacji. Zespół, który nie umie zmierzyć, czy model jest lepszy, nie dowie się tego po treningu. Własny, tajny zbiór testowy buduje się przed treningiem, nie po nim.

Nie nazywać modelem własnym cudzych wag pod nową nazwą. Poza kwestiami licencyjnymi i przepisami o nieuczciwych praktykach rynkowych jest to ryzyko reputacyjne — odbiorca instytucjonalny coraz częściej pyta wprost o pochodzenie wag, a od sierpnia 2025 roku dostawca modelu ogólnego przeznaczenia ma obowiązek opublikować streszczenie danych treningowych.

Pierwsze trzy ruchy

TABELA 17.1 — CO ZROBIĆ W PIERWSZYM KWARTALE

RUCH	DZIAŁANIE	KOSZT	CZAS
1	Przejść załącznik B — wytrenować model miliardowy od zera na wynajętym węźle. Nie po to, żeby go używać, tylko żeby mieć zmierzone, a nie zasłyszane pojęcie o skali.	≈ 1 200 zł	2 dni
2	Policzyć koszt alternatywy: ile kosztowałby trzyletni wolumen zapytań produktu na cudzym modelu, z uwzględnieniem pamięci podręcznej i trybu wsadowego.	—	1 dzień
3	Złożyć wniosek o przydział mocy w trybie przyspieszonym. Rozpatrywany w cztery dni robocze, bezpłatny, niezobowiązujący.	0 zł	4 dni

Trzy ruchy o łącznym koszcie poniżej dwóch tysięcy złotych, po których decyzja o wejściu w scenariusz B albo o rezygnacji będzie oparta na własnych pomiarach, a nie na cudzych opracowaniach — łącznie z tym.

ZDANIE ZAMYKAJĄCE

Budowa własnego modelu przestała być problemem inżynierskim, a stała się problemem alokacji: czego dokładnie chcemy, ile to warto i co mamy takiego, czego nie ma nikt inny. Dla polskiego podmiotu odpowiedź na trzecie pytanie brzmi prawie zawsze: *dane i kompetencja dziedzinowa* — nie karty graficzne. Cały wysiłek warto skierować tam.

ZAŁĄCZNIK A

Własny model od zera na MacBooku

Przepis wykonawczy: model o 33,8 miliona parametrów, wytrenowany od zera na polskim korpusie, na komputerze, który już Pan ma. Koszt — nieco ponad złotówka za prąd. Efekt — pełne zrozumienie mechanizmu, którego opisowi poświęcono poprzednie czternaście rozdziałów.

TEN PRZEPIS ZOSTAŁ WYKONANY, NIE TYLKO OPISANY

Wszystkie liczby w tym załączniku pochodzą z rzeczywistego przebiegu przeprowadzonego na MacBooku z układem Apple M5 Max i 48 GB pamięci — tej samej maszynie, dla której przepis jest napisany. Wykres krzywej straty niżej to

pomiar, nie ilustracja. Tam, gdzie pomiar rozminął się z wcześniejszym szacunkiem, zostawiłem obie wartości.

CZEGO SIĘ PO TYM SPODZIEWAĆ

Model będzie poprawnie odmieniał rzeczowniki i uzgadniał przypadki w krótkich frazach. Nie będzie miał żadnej wiedzy o świecie, a spójność wypowiedzi rozpadnie się po dwóch–trzech zdaniach. To narzędzie do zrozumienia mechanizmu, nie produkt — ale jest to *w pełni własny model*, którego każdy parametr powstał na Pana danych.

Krok 1. Środowisko

Zmienna środowiskowa na końcu sprawia, że brakujące jądra obliczeniowe układu Apple spadają na procesor główny zamiast przerywać trening. Co do wersji Pythona — obiegowa rada każe przypiąć się do 3.12, bo nowszą bywają wyprzedzone koła binarne zależności; w przebiegu kontrolnym cały łańcuch zadziałał jednak bez zarzutu na Pythonie 3.14 z biblioteką PyTorch 2.13, więc przypinanie okazało się zbędne.

```
xcode-select --install
curl -LsSf https://astral.sh/uv/install.sh | sh
mkdir -p ~/pl-lm && cd ~/pl-lm
uv venv --python 3.12 .venv && source .venv/bin/activate
uv pip install torch numpy pyarrow tqdm tokenizers datasets huggingface-hub
export PYTORCH_ENABLE_MPS_FALLBACK=1

# sprawdzenie, czy układ graficzny jest widoczny dla PyTorch
python -c "import torch; print(torch.__version__, torch.backends.mps.is_available())"
```

Krok 2. Korpus

Najwygodniejszym punktem wyjścia przy tej skali jest zbiór o jawnych licencjach, obejmujący między innymi Wikipedię, Wolne Lektury, Dziennik Ustaw, materiał z bazy EUR-Lex i stenogramy parlamentarne — łącznie 2,71 miliona dokumentów i 6,88 miliarda tokenów w 6,6 gigabajta. Dla modelu tej wielkości to znacznie więcej, niż potrzeba.

```
hf download SlayerLab/polish-dynaword --repo-type dataset \
--local-dir data/dynaword --include "data/**/*.parquet"
```

Warto znać alternatywy o innej skali: polski podzbiór zbioru FineWeb2 liczy 73,1 miliarda słów w 432 gigabajtach, jego wersja przefiltrowana jakościowo obejmuje 13,4 miliona dokumentów, a zbiór HPLT — 89,5 miliarda słów. Do treningu samego tokenizatora wystarczy polska Wikipedia: 1,59 miliona artykułów w 2,95 gigabajta.

Krok 3. Tokenizer

Słownik szesnastu tysięcy jest właściwy dla tej skali modelu — większy zjadłby zbyt dużą część jego parametrów, co omawia rozdział piąty. Trening jest szybszy, niż się powszechnie sądzi: na czterystu milionach znaków zajęł **17 sekund**, bo biblioteka rozkłada pracę na wszystkie rdzenie i wykorzystała ich siedem i pół. Tokenizacja całego korpusu do postaci binarnej — 80,8 miliona tokenów — zajęła kolejnych 18 sekund.

```

import glob, pyarrow.parquet as pq
from tokenizers import Tokenizer, models, pre_tokenizers, decoders, trainers, processors

def teksty(sciezki, max_znakow=2_000_000_000, limit_dok=20_000):
    n = 0
    for p in sciezki:
        pf = pq.ParquetFile(p)
        for i in range(pf.num_row_groups):
            for t in pf.read_row_group(i, columns=["text"]).column("text").to_pylist():
                t = t[:limit_dok]; n += len(t); yield t
            if n > max_znakow: return

tok = Tokenizer(models.BPE())
tok.pre_tokenizer = pre_tokenizers.ByteLevel(add_prefix_space=False)
tok.decoder = decoders.ByteLevel()
tok.post_processor = processors.ByteLevel(trim_offsets=False)
trener = trainers.BpeTrainer(
    vocab_size=16384, min_frequency=2, show_progress=True,
    special_tokens=["<|endoftext|>"],
    initial_alphabet=pre_tokenizers.ByteLevel.alphabet())
tok.train_from_iterator(
    teksty(sorted(glob.glob("data/dynaword/data/*/*.parquet"))), trainer=trener)
tok.save("tokenizer-pl-16k.json")

```

Krok 4. Model i trening

Konfiguracja: osiem warstw, wymiar 512, osiem głów mechanizmu uwagi, kontekst 512 tokenów, słownik 16 384, wagi wejścia związane z wyjściem. Daje to dokładnie 33 816 576 parametrów. Trening prowadzi się repozytorium nanoGPT, po zapisaniu strumieni treningowego i walidacyjnego oraz pliku opisującego rozmiar słownika — bez tego ostatniego biblioteka przyjmie wartość domyślną i trening się nie powiedzie.

```

python train.py --dataset=polish --out_dir=out-pl-34m \
  --device=mps --dtype=float32 --compile=False \
  --n_layer=8 --n_head=8 --n_embd=512 --block_size=512 \
  --bias=False --dropout=0.0 \
  --batch_size=32 --gradient_accumulation_steps=4 \
  --learning_rate=6e-4 --min_lr=6e-5 --warmup_iters=200 \
  --max_iters=10000 --lr_decay_iters=10000 \
  --weight_decay=0.1 --beta2=0.95 --grad_clip=1.0 --eval_interval=250

```

TRZY PUŁAPKI, NA KTÓRE TRZEBA UWAGAĆ

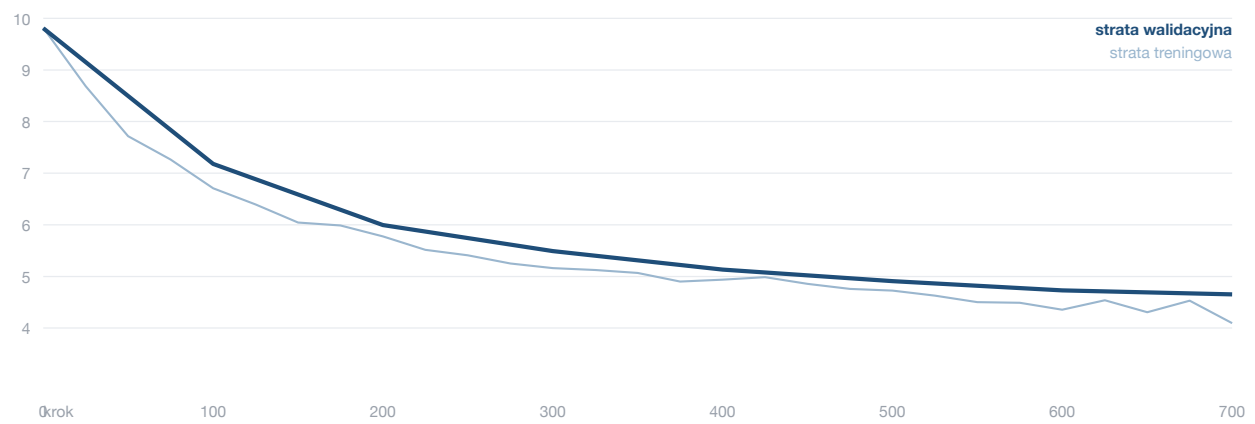
Kompilacja grafu obliczeniowego bywa na układach Apple zawodna — stąd wyłączenie jej w poleceniu. Biblioteka nanoGPT traktuje wszystko poza kartami NVIDIA jako procesor główny, więc automatyczne obniżanie precyzji się nie włącza i model liczy się w pełnej precyzji; kto chce precyzji połówkowej, musi samodzielnie opakować krok treningowy. Trzecia rzecz: w nowszych wersjach PyTorch'a jedna z konstrukcji użytych w pliku treningowym jest oznaczona jako przestarzała i wypisuje ostrzeżenie — trening działa mimo tego.

Krok 5. Zmierzona krzywa straty

Jeden krok przetwarza 65 536 tokenów, więc dziesięć tysięcy iteracji daje 0,655 miliarda tokenów — dokładnie dwudziestokrotność liczby parametrów, czyli klasyczny punkt optymalny. Cały budżet obliczeniowy to $1,33 \times 10^{17}$ operacji.

Przy uruchomieniu biblioteka potwierdza konfigurację, i warto na to spojrzeć, bo jest to sprawdzian, czy rozumie się własny model: raportuje **33 816 576 parametrów** — co do jedności tyle, ile daje rachunek z rozdziału szóstego.

RYSUNEK A.1 — KRZYWA STRATY ZMIERZONA NA APPLE M5 MAX



Ciemna linia to strata walidacyjna liczona co sto kroków, jasna — treningowa co dwadzieścia pięć. Wartość początkowa 9,81 odpowiada modelowi zgadującemu losowo; teoretyczne minimum dla słownika 16 384 wynosi 9,70.

TABELA A.1 — POMIAR WOBEC WCZEŚNIEJSZEGO SZACUNKU

WIELKOŚĆ	SZACUNEK	POMIAR	KOMENTARZ
Trening tokenizatora	kilkanaście minut	17 s	zrównoleglenie na rdzeniach
Tokenizacja korpusu	—	23 s	80 mln tokenów
Liczba parametrów	33 816 576	33 816 576	zgodność co do jedności
Strata początkowa	9,70	9,81	losowa inicjalizacja, nie rozkład jednostajny
Strata po 100 krokach	≈ 6,3	7,18	—
Strata po 700 krokach	—	4,65	koniec przebiegu
Przepustowość początkowa	—	31 000 tok./s	2,1 s na krok
Przepustowość ustalona	—	12 000 tok./s	5,5 s na krok — ograniczenie cieplne
Czas przebiegu 700 kroków	—	≈ 60 min	0,046 mld tokenów
Pełny przebieg 10 000 kroków	7,5–18 h	9–14 h	mieści się w szacunku

Wiersz o przepustowości jest najważniejszy przy planowaniu: laptop utrzymuje pełną wydajność przez kilkanaście minut, po czym schodzi mniej więcej o połowę. Rachunek z wartości początkowej zaniża czas przebiegu dwukrotnie.

DLACZEGO STRATA Z DWÓCH PRZEBIEGÓW JEST NIEPORÓWNYWALNA

Ten sam model, ta sama konfiguracja i ta sama liczba kroków dały na korpusie złożonym głównie z aktów prawnych stratę 3,18 po czterystu krokach, a na korpusie ogólnym — 5,13. Model nie stał się gorszy; tekst stał się trudniejszy. Akty normatywne są schematyczne i łatwe do przewidzenia, proza i teksty encyklopedyczne mają znacznie większą entropię. *Porównywanie wartości straty między modelami trenowanymi na różnych korpusach jest bezwartościowe*, choć wygląda na porównywanie tej samej wielkości — to ta sama pułapka, którą rozdział dziesiąty opisuje przy benchmarkach.

CZEGO UCZY SAM KSZTAŁT KRZYWEJ

Cały spadek z 9,81 do 6,0 dokonuje się w pierwszych dwustu krokach, czyli w pierwszym kwadransie. Kolejne pięćset kroków przynosi poprawę o niecałe półtora punktu. To jest w miniaturze ta sama zależność, która w rozdziale

ósmym każe wydawać dziesiątki milionów dolarów na ostatnie kilka procent jakości — *tanie jest dojście do poprawnej polszczyzny, drogie jest wszystko powyżej.*

DIAGNOSTYKA, GDY COŚ NIE IDZIE

Strata tkwi przy 9,7 po pięciuset krokach — dane są uszkodzone albo krok uczenia wynosi zero. Po zjeździe do 6,3 nic się nie dzieje przez tysiąc kroków — model uczy się wyłącznie częstości, sprawdzić kodowanie pozycji. Strata przyjmuje wartość nieokreśloną — krok uczenia za duży albo obniżona precyzja bez skalowania gradientu.

Krok 6. Co model naprawdę pisze

Poniżej surowy wynik z opisanego przebiegu, po siedmiuset krokach i godzinie treningu, przy stracie walidacyjnej 4,65. Tekst nie jest wybrany ani poprawiony — to trzy pierwsze próbki przy ustalonym ziarnie losowości. Warto je przeczytać uważnie, bo pokazują sedno całego opracowania lepiej niż którykolwiek rozdział.

- » Warszawa jest miastem miastem, a od II lidze, przy ulicy Śróda, a od rzeki do Ow. Długoszczono tutaj od rzeki Bałkanów, pod względem wielkości wsi Ow. W skład tego miejscowości wchodziły w skład Nowe Woodia (1826), co stanowiła siedzibą gminy Zagród i Osieka...
- » W dziewiętnastym wieku (Urokitny) i „Więć” (1) „Part” (1826) „Uczestnicy” (1934–1984), „Skutki” (1957–1939) „Ręcza”. Zarówno w „Wiado”, „Młody”. Życiorys.
- » Podstawową zasadą w tej pory lub przedstawieniem w tym samym czasie. W Polsce wystanie na posłusznej na wyprawie w Polsce. Dopiero w 1895 w Bukowskim Wschodzie przybrzegnął na terenie Warszawy...

CO DOKŁADNIE WIDAĆ W TYCH TRZECH AKAPITACH

Model opanował **formę**: odmianę, uzgodnienie przypadków, interpunkcję, a nawet konwencje redakcyjne hasła encyklopedycznego — nawiasy z datami, zakresy lat, cudzysłowy wokół tytułów, śródtytuł „Życiorys”. Trzydzieści trzy miliony parametrów i godzina na laptopie wystarczyły, żeby nauczyć się kształtu polskiego tekstu.

Nie opanował **treści**. Nie istnieje wieś Ow, miejscowość Nowe Woodia ani rzeka Bałkanów. Zakres lat „1957–1939” biegnie wstecz. Zdania rozpadają się po dwóch przecinkach, a „miastem miastem” to powtórzenie, którego model nie zauważa.

To jest, obojętna z bliska, ta sama właściwość, którą przy modelach tysięcy razy większych nazywa się konfabulacją. *Model uczy się rozkładu prawdopodobieństwa nad tekstem, a nie zbioru faktów* — przekonująca forma pojawia się dużo wcześniej niż prawdziwa treść i jest dużo tańsza. Kto raz zobaczy to na własnym modelu, przestaje pytać, dlaczego duże modele potrafią zmyślić datę albo sygnaturę.

Żeby model odpowiadał na polecenia, a nie dopisywał tekst, potrzebuje jeszcze krótkiego dostrojenia na kilku tysiącach par polecenie–odpowiedź, sformatowanych znacznikami wprowadzonymi do słownika. Przy tej skali efekt będzie skromny, ale różnica w zachowaniu przed i po jest wyraźna — i to właśnie jest zjawisko opisane w rozdziale dziewiątym.

Ostatnia pułapka: nanoGPT dekoduje wyniki słownikiem angielskim, więc do obejrzenia próbki trzeba podstawić własny tokenizator. Skrypt jest krótki — wczytuje punkt kontrolny, zdejmuje przedrostki z nazw wag, przełącza model w tryb wnioskowania i generuje kilkadziesiąt tokenów przy temperaturze 0,8.

Model miliardowy na wynajętym klastrze

Przepis na model o blisko miliardzie parametrów, wytrenowany od zera na dwudziestu miliardach polskich tokenów. Koszt — od dziewięciuset do tysiąca czterystu złotych. Czas — jeden dzień. To jest realny próg wejścia w samodzielny trening i granica, poniżej której nie warto schodzić, jeśli celem jest cokolwiek użytkowego.

Wybór skali

Współczesne repozytoria referencyjne sprowadzają konfigurację do jednego pokrętła — głębokości modelu, z której wynika reszta wymiarów. Poniżej trzy sensowne punkty pracy.

TABELA B.1 — TRZY WARIANTY I ICH RACHUNEK

WARSTWY	WYMIAR	PARAMETRY	TOKENY	SPRZĘT I CZAS	KOSZT
20	1 280	477 mln	10 mld	1 × H100, 17,9 godz.	< 300 zł
26	1 664	973 mln	20 mld	8 × H100, 9,1 godz.	900–1 400 zł
30	1 920	1,45 mld	30 mld	8 × H100, ~20 godz.	2–3 tys. zł

Kalkulacja przy wykorzystaniu mocy 45 % i cenach najmu z sierpnia 2026 r.: 2,69–3,99 USD za kartę i godzinę, czyli 21,52–31,92 USD za węzeł ośmiokartowy. Kurs 3,7425 PLN za dolara. Wariant przerywalny schodzi do 1,05–1,87 USD za kartę, ale wymaga zapisu punktów kontrolnych co 250 kroków.

Krok 1. Środowisko i korpus

```
git clone https://github.com/karpathy/nanochat && cd nanochat
uv sync --extra gpu && source .venv/bin/activate

# dziewięć plików polskiego FineWeb2 – ok. 37 GB, ok. 20 mld tokenów
hf download HuggingFaceFW/fineweb-2 --repo-type dataset --local-dir /data/fw2 \
  --include "data/pol_Latn/train/00000*.parquet"
```

Podmiana korpusu jest prostsza, niż się wydaje: biblioteka czyta ponumerowane pliki z kolumną tekstową z jednego katalogu, a ostatni z nich alfabetycznie służy jako zbiór walidacyjny. Wystarczy przepakować pobrane pliki pod oczekiwane nazwy.

```
python -c "
import glob, os, pyarrow.parquet as pq
d = os.path.expanduser('~/.cache/nanochat/base_data_climbmix')
os.makedirs(d, exist_ok=True)
for i, p in enumerate(sorted(glob.glob('/data/fw2/data/pol_Latn/train/*.parquet'))):
    t = pq.read_table(p, columns=['text'])
    pq.write_table(t, f'{d}/shard_{i:05d}.parquet', row_group_size=8192)
"
```

Krok 2. Tokenizer, pretrening, dostrojenie

Trzy polecenia. Pierwsze trenuje własny tokenizer o słowniku 32 768 na dwóch miliardach znaków polskiego tekstu — to właśnie ten element daje przewagę opisaną w rozdziale piątym. Drugie prowadzi pretrening w precyzji ośmiobitowej. Trzecie dostroja model na rozmowach.

```
python -m scripts.tok_train --vocab-size=32768 --max-chars=2000000000

torchrun --standalone --nproc_per_node=8 -m scripts.base_train -- \
  --depth=26 --target-param-data-ratio=20 --device-batch-size=16 --fp8 --run=pl1b

torchrun --standalone --nproc_per_node=8 -m scripts.chat_sft -- --run=pl1b-sft
python -m scripts.chat_cli
```

Krok 3. Dane instrukcyjne — tu jest wąskie gardło

Otwartych polskich zbiorów rozmów jest niewiele. Praktyczny plan to od dwudziestu do pięćdziesięciu tysięcy rozmów: mniej więcej połowa z dostępnych zbiorów publicznych, reszta wygenerowana większym modelem na własnych zadaniach dziedzinowych. Zasady techniczne, które warto zapamiętać: znaczniki ról muszą trafić do słownika *przed* pretreningiem, strata liczona jest wyłącznie na tokenach odpowiedzi, dwa do trzech przejść przez zbiór, krok uczenia równy jednej dziesiątej szczytowego kroku z pretreningu i bez rozgrzewki.

Koszt tego etapu jest pomijalny — dostrojenie modelu miliardowego na trzydziestu tysiącach rozmów to około czterdziestu minut na jednej karcie, czyli dwa dolary. Cały ciężar leży w przygotowaniu samych rozmów.

Krok 4. Kontrola przebiegu

Funkcja straty startuje przy wartości 10,40, czyli logarytmie naturalnym rozmiaru słownika. Po pierwszej setce kroków spada poniżej siedmiu, po dziesięciu procentach przebiegu wynosi 3,4–3,6, a na końcu 2,6–2,9 na token. Jeżeli po połowie przebiegu strata przekracza 3,5, prawdopodobnie zbyt mały jest efektywny rozmiar porcji danych albo dane są niedostatecznie przemieszane.

Krok 5. Uruchomienie u siebie

Gotowe wagi można wyeksportować do powszechnie używanego formatu i uruchomić przez standardową bibliotekę, albo skonwertować do formatu skompresowanego i uruchomić lokalnie. Model miliardowy w kwantyzacji ośmiobitowej zajmuje około 1,1 gigabajta i generuje na wydajnym MacBooku ponad sto tokenów na sekundę — czyli szybciej, niż da się czytać.

REALISTYCZNE OCZEKIWANIA

Model 477-milionowy po dostrojeniu nadaje się do klasyfikacji, wydobywania pól ze struktury, przepisywania stylu i krótkich odpowiedzi na pytania zamknięte — przy wynikach na testach wiedzy na poziomie losowym. Model 973-milionowy jest wyraźnie sprawniejszy językowo i utrzymuje spójność akapitu, ale nadal konfabuluje i przegrywa z każdym modelem dostępnym komercyjnie.

Wartość takiego przebiegu leży gdzie indziej: w pełnej kontroli nad danymi, w policzalnym koszcie i w tym, że po dwudziestu godzinach ma się *zmierzone*, a nie *zasłyszane* pojęcie o tym, gdzie przebiega granica między modelem-

zabawką a modelem-narzędziem. Dla kogoś, kto ma podejmować decyzje inwestycyjne w tej dziedzinie, jest to prawdopodobnie najlepiej wydany tysiąc złotych.

GDZIE KOŃCZY SIĘ SENS TEJ DROGI

Model klasy jedenastu miliardów parametrów, trenowany na 1,5 biliona tokenów, to 7724 godziny pracy węzła ośmiokartowego — 322 dni pracy ciągłej i około 622 tysięcy złotych samego najmu. To jest granica, za którą samodzielny trening od zera przestaje mieć uzasadnienie ekonomiczne, a właściwą drogą staje się kontynuowany pretrening opisany w rozdziale piętnastym.

ZAŁĄCZNIK C

Słownik pojęć

Terminy używane w opracowaniu, z polskimi odpowiednikami i angielskimi oryginałami. Uporządkowane rzeczowo, nie alfabetycznie — pojęcia pokrewne stoją obok siebie.

Budulec modelu

POJĘCIE	ORYGINAŁ	ZNACZENIE
Parametr, waga	parameter, weight	Pojedyncza liczba wewnątrz sieci, zmieniana w trakcie treningu. Ich liczba — miliardy — jest podstawową miarą wielkości modelu.
Token	token	Najmniejsza jednostka tekstu przetwarzana przez model. Zwykle fragment wyrazu. Rozliczenia z dostawcami odbywają się w tokenach, nie w znakach.
Tokenizer	tokenizer	Słownik i algorytm zamieniający tekst na ciąg liczb. Trenowany na korpusie przed modelem właściwym.
Osadzenie	embedding	Reprezentacja tokenu jako wektora liczb. Warstwa wejściowa modelu; osobno wykorzystywana w wyszukiwaniu semantycznym.
Kontekst, okno kontekstu	context window	Maksymalna liczba tokenów, którą model przetwarza jednorazowo. Koszt jego obsługi rośnie szybciej niż liniowo.
Transformer	transformer	Architektura sieci neuronowej z 2017 r., podstawa wszystkich współczesnych modeli językowych.
Uwaga	attention	Mechanizm pozwalający każdemu tokenowi odnieść się do pozostałych. Główne źródło kosztu obliczeniowego przy długim kontekście.
Mieszanka ekspertów	Mixture of Experts, MoE	Architektura, w której dla każdego tokenu uruchamia się tylko część sieci. Pozwala zwiększyć liczbę parametrów bez proporcjonalnego wzrostu kosztu.
Parametry aktywne	active parameters	W modelu mieszanki ekspertów: część parametrów faktycznie użyta przy jednym tokenie. To ona wyznacza koszt, nie suma całkowita.

Etapy powstawania

POJĘCIE	ORYGINAŁ	ZNACZENIE
Pretrening	pretraining	Zasadniczy trening na wielkim korpusie, polegający na przewidywaniu kolejnego tokenu. Pochłania większość budżetu.
Trening ciągły	continued pretraining	Dołożenie kolejnych danych do modelu już wytrenowanego, np. w celu adaptacji do języka lub dziedziny.

POJĘCIE	ORYGINAŁ	ZNACZENIE
Post-trening	post-training	Zbiorcza nazwa etapów zamieniających model surowy w asystenta: dostrojenie na instrukcjach i uczenie z informacji zwrotnej.
Dostrojenie nadzorowane	supervised fine-tuning, SFT	Trening na parach polecenie–wzorcowa odpowiedź, przygotowanych przez ludzi.
Uczenie z informacji zwrotnej	RLHF	Uczenie ze wzmocnieniem na podstawie ludzkich ocen porównawczych. Nadaje modelowi przydatność i sposób wypowiedzi.
Bezpośrednia optymalizacja preferencji	DPO	Tańsza metoda osiągająca podobny skutek bez trenowania osobnego modelu nagrody.
Uczenie na nagrodach weryfikowalnych	RLVR	Trening na zadaniach o sprawdzalnym wyniku — matematyka, kod. Podstawa modeli rozumujących.
Destylacja	distillation	Uczenie modelu mniejszego na odpowiedziach większego. Tanie, ale ograniczone warunkami licencji modelu źródłowego.
Dostrojenie niskiego rzędu	LoRA, QLoRA	Technika modyfikująca ułamek procenta wag. Umożliwia dostrojenie dużego modelu na jednej karcie.

Skala, sprzęt i koszt

POJĘCIE	ORYGINAŁ	ZNACZENIE
Operacja zmiennoprzecinkowa	FLOP	Jednostka pracy obliczeniowej. Skala treningu wyraża się w potęgach dziesięciu — 10^{25} to próg regulacyjny w prawie unijnym.
Wykorzystanie mocy modelu	MFU	Odsetek teoretycznej mocy kart faktycznie wykorzystany. Realnie 35–50 %; reszta ginie w komunikacji i oczekiwaniu na dane.
Prawa skalowania	scaling laws	Empiryczne zależności wiążące wielkość modelu, ilość danych i moc obliczeniową z osiąganą jakością.
Karta obliczeniowa	GPU, akcelerator	Procesor graficzny wyspecjalizowany w mnożeniu macierzy. Podstawowa jednostka rozliczeniowa całego przedsięwzięcia.
Godzina karty	GPU-hour	Praca jednej karty przez godzinę. Waluta, w której podaje się koszty treningu.
Sieć wewnętrzna klastra	InfiniBand, RoCE	Szybkie połączenie między węzłami. Przy tysiącach kart jej przepustowość decyduje o wydajności bardziej niż same karty.
Punkt kontrolny	checkpoint	Zapis stanu treningu, pozwalający wznowić pracę po awarii. Przy dużych modelach zajmuje terabajty i zapisywany jest co kilkadziesiąt minut.
Kwantyzacja	quantization	Zapis wag z mniejszą precyzją. Zmniejsza zapotrzebowanie na pamięć kilkukrotnie kosztem niewielkiego spadku jakości.
Podręczna pamięć uwagi	KV cache	Bufor przechowujący stan rozmowy w pamięci karty. Przy długim kontekście bywa większy niż same wagi.
Wnioskowanie	inference	Praca modelu już wytrenowanego — generowanie odpowiedzi. Koszt bieżący, w skali usługi przewyższający koszt treningu.

Dane, jakość i prawo

POJĘCIE	ORYGINAŁ	ZNACZENIE
Korpus	corpus	Zbiór tekstów, na którym trenuje się model.
Odsiewanie duplikatów	deduplication	Usuwanie powtórzeń z korpusu. Znacząco poprawia jakość i skraca trening.
Odkazanie zbioru	decontamination	Usuwanie z korpusu zadań testowych, żeby wyniki ewaluacji nie były zawyżone.

POJĘCIE	ORYGINAŁ	ZNACZENIE
Zanieczyszczenie zbioru testowego	contamination	Sytuacja, w której model widział pytania testowe w treningu. Główna przyczyna zawyżonych wyników w publikacjach.
Konfabulacja	hallucination	Wygenerowanie treści fałszywej, przedstawionej w formie stanowczej. Właściwość statystycznej natury modelu, nie usterka do naprawienia.
Czerwony zespół	red teaming	Celowe próby wymuszenia na modelu zachowań niepożądanych, prowadzone przed udostępnieniem.
Model ogólnego przeznaczenia	GPAI	Kategoria z aktu o sztucznej inteligencji, obejmująca modele o szerokim zastosowaniu. Wiąże się z konkretnymi obowiązkami dostawcy.
Ryzyko systemowe	systemic risk	Kwalifikacja modelu przekraczającego próg mocy obliczeniowej użytej do treningu, uruchamiająca obowiązki dodatkowe.
Eksploracja tekstów i danych	TDM	Zautomatyzowana analiza utworów, objęta wyjątkiem od prawa autorskiego, z możliwością zastrzeżenia przez uprawnionego.

ZAŁĄCZNIK D

Źródła i dalsza lektura

Opracowanie opiera się na ponad trzystu pozycjach źródłowych. Materiał przeszedł trzy niezależne kontrole: 186 twierdzeń liczbowych sprawdzono przeliczeniem i u źródła, z czego 44 skorygowano, a rozdział prawny poddano dwóm osobnym weryfikacjom na tekstach autentycznych, które przyniosły kolejnych szesnaście poprawek. Poniżej wybór najważniejszych źródeł, uporządkowany według tego, do czego się przydają. Priorytet mają źródła pierwotne: raporty techniczne, karty modeli, cenniki producentów i teksty aktów prawnych.

Podstawy teoretyczne — jeśli czytać tylko cztery pozycje

Kaplan i in., <i>Scaling Laws for Neural Language Models</i> arXiv:2001.08361, 2020	Źródło przybliżenia sześciu operacji na parametr i token, na którym opiera się cała arytmetyka rozdziału ósmego.
Hoffmann i in., <i>Training Compute-Optimal Large Language Models</i> arXiv:2203.15556, 2022	Prawa skalowania i reguła dwudziestu tokenów na parametr, od której współczesna praktyka odeszła — i wyjaśnienie, dlaczego.
Meta, <i>The Llama 3 Herd of Models</i> arXiv:2407.21783, 2024	Najbardziej szczegółowy publiczny opis treningu w dużej skali: konfiguracja równoległości, statystyka awarii, rozliczone godziny kart.
DeepSeek-AI, <i>DeepSeek-V3 Technical Report</i> arXiv:2412.19437, 2024	Dowód, że wynik zbliżony do czołówki da się osiągnąć za mniej niż sześć milionów dolarów przebiegu. Pełne rozliczenie godzin.

Koszty i skala

Serwis **Epoch AI** jest jedynym systematycznym źródłem szacunków kosztów modeli czołowych; szczególnie przydatne opracowania to *How much does it cost to train frontier AI models?*, analiza zasobów modelu Grok 4 z września 2025 oraz materiał wykazujący, że przebiegi finalne stanowią mniejszość wydatku obliczeniowego. Uzupełnieniem jest **Stanford AI Index**, stosujący odmienną metodykę wyceny — rozbieżność między obiema seriami wynosi około dwukrotności i wynika z konwencji, nie z błędu.

Ceny sprzętu i najmu: cenniki producentów (Lambda, Nebius, CoreWeave, RunPod, Vast.ai) odczytane bezpośrednio w sierpniu 2026 roku, agregatory *getdeploying* i *IntuitionLabs Data Center GPU Pricing 2026*. Kurs walutowy — tabela 147/A/NBP/2026 z 31 lipca 2026 roku.

Dane treningowe

Raport techniczny zbioru **FineWeb** (arXiv:2406.17557) jest najlepszym istniejącym opisem potoku przetwarzania danych, wraz z pełnym rozliczeniem, ile materiału odsiewa każdy etap i ile to kosztowało. Dla wielojęzyczności — **FineWeb2** (arXiv:2506.20920), z którego pochodzą wszystkie liczby o polskim korpusie. Skład zbioru **Dolma** (arXiv:2402.00159) jako jedyny jest ujawniony pozycja po pozycji. Statystyki językowe archiwum internetu — *commoncrawl.github.io/cc-crawl-statistics*.

Post-trening

Praca **LIMA** (arXiv:2305.11206) wykazująca, że tysiąc dobranych przykładów wystarcza do nadania modelowi formy asystenta. Receptura **Tulu 3** (arXiv:2411.15124) jako najbardziej kompletny otwarty opis post-treningu. **DeepSeek-R1** (arXiv:2501.12948 oraz publikacja recenzowana w „Nature” z września 2025) — wzorzec budowy modelu rozumującego wraz z kosztem. **Constitutional AI** (arXiv:2212.08073) dla metody opartej na spisanych zasadach.

Polskie zasoby

Bielik v3 Small: Technical Report arXiv:2505.02550, 2025	Korpus 292 mld tokenów, własny tokenizator, pomiary wydajności kodowania polszczyzny.
Bielik 11B v2 Technical Report arXiv:2505.02410, 2025	Pogłębienie modelu bazowego, pełne hiperparametry treningu.
PLLuM arXiv:2511.03823, 2025	Korpus, instrukcje, skład konsorcjum, zaplecze obliczeniowe.
LLMzSzł arXiv:2501.02266	Blisko 19 tys. pytań z egzaminów państwowych — jedyny zbiór pozwalający zestawić model ze zdawalnością ludzi.
Zasoby praktyczne	<i>nkjp.pl</i> — Narodowy Korpus Języka Polskiego; <i>clarin-pl.eu</i> — korpusy specjalistyczne; <i>speakleash.org</i> — Spichlerz; <i>github.com/sdadas/polish-nlp-resources</i> — zbiorczy wykaz.

Kod, od którego zacząć

nanoGPT — 330 linii modelu, właściwy punkt wyjścia do zrozumienia mechanizmu. **nanochat** — pełny łańcuch od tokenizatora po interfejs rozmowy w 7775 liniach; podstawa załącznika B. **torch titan** — skala produkcyjna z gotowymi definicjami wszystkich liczących się architektur. **Megatron-LM** — najbardziej dojrzała biblioteka do treningu wielowęzłowego. Do dostrajania: **Unsloth**, **Axolotl**, **torchtune**, **LLaMA-Factory**. Do serwowania: **vLLM**, **SGLang**, **TensorRT-LLM**.

Zasoby obliczeniowe i finansowanie

access.eurohpc-ju.europa.eu — tryby dostępu do europejskiej infrastruktury obliczeniowej, w tym ścieżki bezpłatne dla małych i średnich przedsiębiorstw. *plgrid.pl* — krajowe granty obliczeniowe, przez które uzyskano

czas na superkomputerze Helios dla polskich modeli. Fabryki sztucznej inteligencji: krakowska Gaia oraz poznańska PIAST-AI.

Prawo

Teksty pierwotne w bazie EUR-Lex: rozporządzenie 2024/1689 o sztucznej inteligencji oraz dyrektywa 2019/790 o prawie autorskim na jednolitym rynku cyfrowym. Materiały pomocnicze: *artificialintelligenceact.eu* z zestawieniem artykułów i harmonogramu, wytyczne oraz kodeks postępowania publikowane przez Urząd do spraw Sztucznej Inteligencji, opinia Europejskiej Rady Ochrony Danych numer 28/2024 w sprawie modeli sztucznej inteligencji a ochrony danych osobowych. Stan spraw sądowych — publiczne rejestry postępowań dotyczących wykorzystania utworów w treningu modeli.

UWAGA O WIARYGODNOŚCI

Najślabszym ogniwem tego opracowania są ceny sprzętu i stawki anotacji pochodzące z portali branżowych o nieujawnionej metodyce. Weryfikatorzy zdobili potwierdzić u źródła pierwotnego cenniki chmurowe, dane energetyczne i wszystkie rozliczenia godzin kart z raportów technicznych, natomiast twierdzenia o cenach zakupu sprzętu i o wartościach umów licencyjnych opierają się na źródłach wtórnych. Wszystkie liczby w opracowaniu są opatrzone wskazaniem podstawy — tam, gdzie mowa o szacunku, użyto tego słowa.