Enterprise Strategy Group by TechTarget

OPRACOWANIE EKONOMICZNE

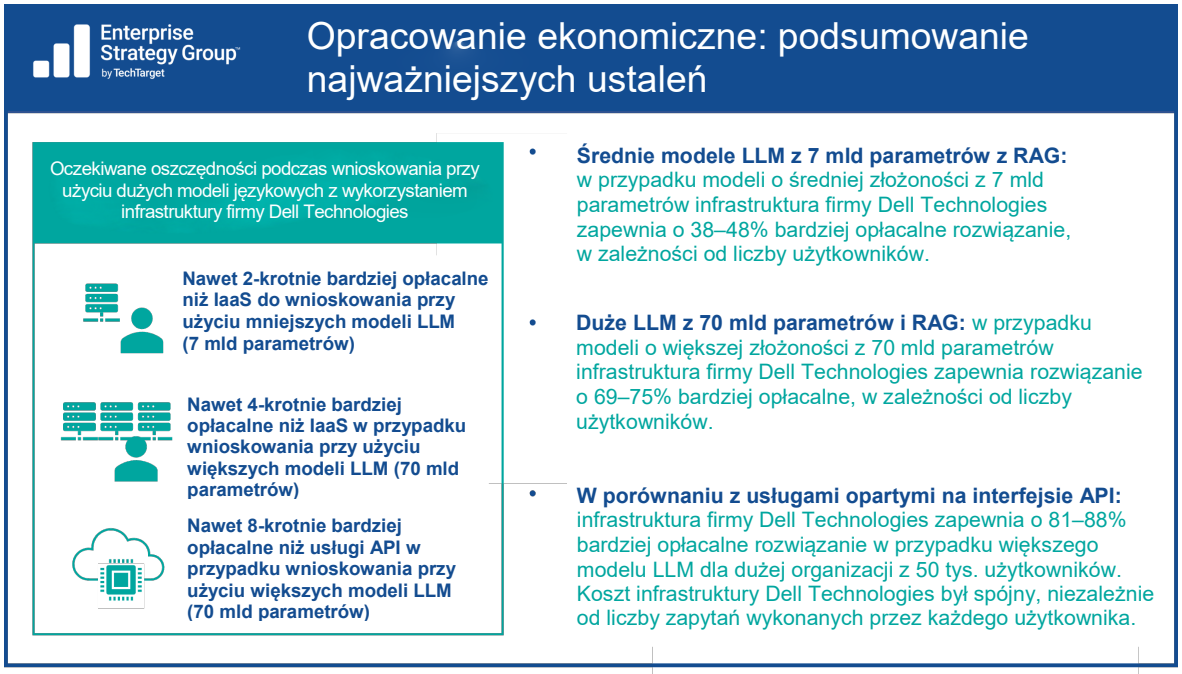
Informacje dotyczące całkowitego kosztu wnioskowania przy użyciu dużych modeli językowych

W jaki sposób wykorzystanie lokalnych rozwiązań firmy Dell Technologies może być o 38–88% bardziej opłacalne w przypadku wnioskowania przy użyciu dużych modeli językowych za pomocą RAG w porównaniu z chmurą publiczną lub interfejsami API opartymi na tokenach

Autor: Aviv Kaufmann, dyrektor placówki i główny analityk ds. weryfikacji Enterprise Strategy Group

Kwiecień 2024 r.

Spis treści



Wstęp

W niniejszym opracowaniu ekonomicznym przedstawiono niektóre opcje i zagadnienia dotyczące dostarczania organizacjom możliwości generatywnej sztucznej inteligencji (GenAI) opartej na tekście. Enterprise Strategy Group TechTarget modelowała i porównywała oczekiwane koszty wnioskowania przy użyciu dużych modeli językowych (LLM) za pomocą techniki retrieval-augmented generation (RAG) w lokalnej infrastrukturze Dell Technologies z wykorzystaniem natywnej infrastruktury chmury publicznej jako usługi (IaaS) lub usługi modelu OpenAI GPT-4 Turbo LLM za pośrednictwem interfejsu API. Okazało się, że firma Dell Technologies może zapewnić wnioskowanie przy użyciu modeli LLM nawet 4 razy bardziej opłacalne niż IaaS i nawet 8 razy taniej niż w przypadku GPT-4 Turbo API.

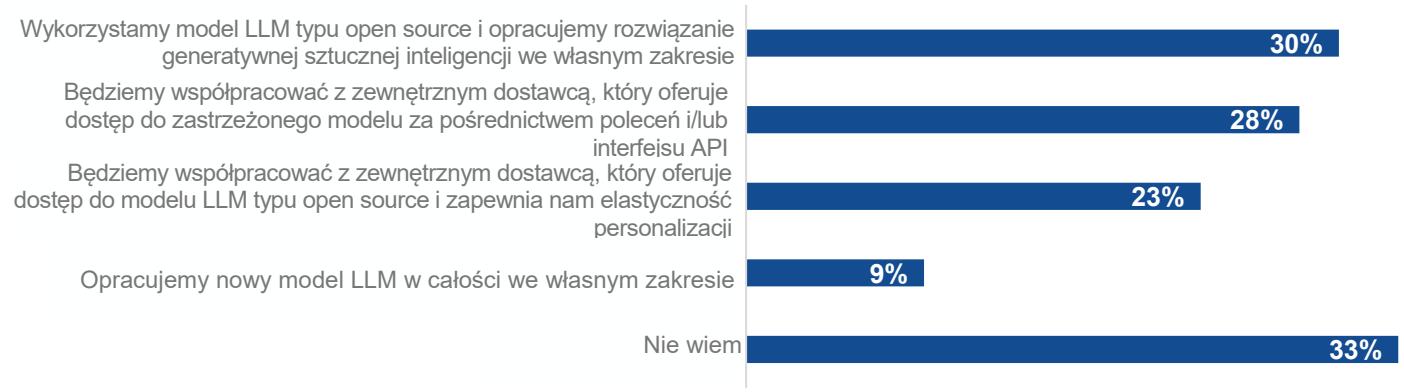
Wyzwania

Organizacje korzystają z mocy generatywnej sztucznej inteligencji i dużych modeli językowych, które wykorzystują dane specyficzne dla firmy i inną własność intelektualną do automatyzacji generowania treści, odpowiadania na pytania i udostępniania szczegółowych informacji decydentom. Oprócz wielu innych korzyści respondenci badania Enterprise Strategy Group stwierdzili, że główne zalety płynące z wykorzystania generatywnej sztucznej inteligencji w ich organizacji obejmują usprawnienie i/lub automatyzację procesów i przepływów pracy, wsparcie analizy danych i analizy biznesowej, zwiększenie produktywności pracowników i poprawę wydajności operacyjnej[1].

Opracowanie modelu LLM może być kosztowne i skomplikowane, ale organizacje mogą łatwo rozszerzyć, dostroić i dostosować istniejące modele LLM typu open source do swoich potrzeb. Gotowe usługi oparte na API, takie jak OpenAI GPT, oferują prostsze rozwiązanie, ale koszty wnioskowania (tj. wykonywania zapytań) mogą szybko wzrastać, szczególnie w przypadku większych organizacji i bardziej złożonych modeli LLM. Organizacje mogą także tworzyć i kontrolować własne rozwiązanie do wnioskowania przy użyciu modeli LLM na potężnych serwerach korporacyjnych z obsługą procesorów graficznych lub równoważnych instancjach chmurowych z obsługą procesorów graficznych oraz na platformie uczenia maszynowego, takiej jak AI Enterprise firmy NVIDIA z systemami LLM typu open source. Nic dziwnego, że Enterprise Strategy Group stwierdziła, że najpopularniejszą strategią opracowywania i wykorzystywania generatywnej sztucznej inteligencji wspieranej przez modele LLM jest wykorzystanie modeli LLM typu open source i opracowanie rozwiązania generatywnej sztucznej inteligencji we własnym zakresie[2].

Rysunek 1. Większość organizacji planuje opracowanie własnego rozwiązania generatywnej sztucznej inteligencji we własnym zakresie

W jaki sposób Twoja organizacja będzie rozwijać/wykorzystywać generatywną sztuczną inteligencję wspieraną przez duży model językowy (LLM)? (Odsetek respondentów, N=670, dopuszczono kilka odpowiedzi)



Źródło: Enterprise Strategy Group, oddział TechTarget, Inc.

[1] Źródło: raport badawczy Enterprise Strategy Group, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, sierpień 2023 r.
[2] Ibid.

Kluczowe zagadnienia dotyczące wnioskowania przy użyciu modeli LLM

Tekstowe modele LLM koncentrują się na uczeniu się, rozumieniu i tworzeniu treści tekstowych, odpowiedzi, podsumowań i pytań, które można dostosować do konkretnej branży, przypadku użycia i organizacji. RAG rozszerza wyniki modeli generatywnej sztucznej inteligencji o niestandardowe dane pobierane z dodatkowych źródeł, co sprawia, że modele są dokładniejsze. Są to najczęściej wdrażane modele LLM dla firm, które można stosować w przypadku chatbotów, asystentów pytań i odpowiedzi, usprawniania i automatyzacji procesów lub jako funkcje wbudowane w niestandardowe narzędzia i aplikacje, a także w wielu innych zastosowaniach. Dostarczając modele LLM, organizacje muszą wziąć pod uwagę infrastrukturę do trenowania (tj. kosztownych pod względem danych i obliczeń analiz wymaganych do stworzenia działającego modelu), wnioskowania (tj. obsługi interakcji użytkowników na wytrenowanym modelu) i dostrajania (tj. ciągłego aktualizowania i optymalizowania modelu). Ten raport koncentruje się na infrastrukturze wymaganej do zmniejszenia obciążeń roboczych związanych z wnioskowaniem. Istnieje kilka metod wdrażania, które można zastosować w przypadku wnioskowania przy użyciu modeli LLM. Zaliczają się do nich:

- **Tradycyjna infrastruktura.** Zakupiona lub dzierżawiona tradycyjna infrastruktura składająca się z zasobów obliczeniowych, pamięci, procesorów graficznych i pamięci masowej może być wdrażana i zarządzana wraz z komercyjną lub otwartą platformą sztucznej inteligencji, zapewniając organizacji kontrolę nad wszystkimi aspektami wdrożenia. Ta metoda może być najbardziej opłacalna w przypadku większych i przewidywalnych obciążeń roboczych.
- **Infrastruktura chmury publicznej jako usługi.** Analogicznie organizacje mogą wdrożyć równoważne wystąpienia chmury obliczeniowej z procesorami graficznymi i pamięcią masową wraz z komercyjną lub otwartą platformą sztucznej inteligencji. Metoda ta zapewnia podobną kontrolę nad platformą, ze zwinnością i łatwą integracją z istniejącymi narzędziami. Może być ona najbardziej opłacalna w przypadku małych wdrożeń oraz wdrożeń o nieprzewidywalnych lub sezonowych wymaganiach.
- **Usługi API do obsługi modeli LLM.** Sprawdzone usługi, takie jak OpenAI GPT, mogą służyć do szybkiego udostępniania funkcji bez konieczności zarządzania infrastrukturą lub platformą AI. Ta metoda może być najlepsza na etapie wstępnym, w mniejszych wdrożeniach i tych, które nie wymagają dużego stopnia dostosowania lub kontroli.

Przed podjęciem decyzji o platformie LLM organizacje powinny dokładnie przyjrzeć się swoim wymaganiom i możliwościom, a także omówić niektóre z następujących kwestii dotyczących wyboru platformy do wnioskowania przy użyciu modeli LLM, takie jak:

- **Koszt/zwrot z inwestycji.** Organizacje powinny rozważyć koszty i korzyści związane z wdrożeniem i wykorzystaniem każdej inwestycji technologicznej. Według badania przeprowadzonego przez Enterprise Strategy Group oszczędności i zwrot z inwestycji są wskaźnikami, za pomocą których organizacje najczęściej mierzą skuteczność swoich inicjatyw związanych ze sztuczną inteligencją[3].
- **Wydajność i skalowalność.** Infrastruktura powinna mieć wystarczająco wiele procesorów, procesorów graficznych, pamięci i pamięci masowej, aby zagwarantować oczekiwaną wielowątkowość wnioskowania przy obciążeniach normalnych i szczytowych oraz akceptowalnie małe opóźnienia. Organizacje powinny również określić, czy intensywne obliczeniowo trenowanie modeli LLM odbędzie się na tej samej platformie czy na dedykowanej platformie treningowej o wyższej wydajności przed jej przeniesieniem na platformę wnioskowania.
- **Proste zarządzanie.** Porównując dowolną infrastrukturę lokalną z infrastrukturą i usługami w chmurze, ważne jest, aby organizacja wzięła pod uwagę swoje wewnętrzne możliwości i zrozumiała koszty obsługi infrastruktury i platform (np. administracji, wsparcia i konserwacji oraz zasilania/chłodzenia). Opcje kolokacji mogą również umożliwić organizacjom uzyskanie wielu korzyści płynących z hostowania we własnych centrach przetwarzania danych przy jednoczesnym odciążeniu zasobów i umiejętności wymaganych do obsługi infrastruktury i platformy.
- **Oczekiwane obciążenie robocze użytkowników.** Ważnym wskaźnikiem, który należy wziąć pod uwagę przy wyborze rozwiązania, jest wiedza i prognoza, ilu użytkowników uzyska dostęp do narzędzia i ile pytań dziennie będą oni zadawać. Jeśli zapotrzebowanie jest niewielkie, usługa API może wystarczyć, ale wraz ze wzrostem liczby obsługiwanych użytkowników i zapytań bardziej opłacalne staje się stworzenie własnej platformy. Ważne jest, aby organizacje brały pod uwagę spodziewany wzrost popularności platformy i częstotliwości jej użytkowania wraz z upływem czasu, aby infrastruktura była odpowiednio duża i mogła być rozbudowywana wraz z potrzebami firmy.

[3] Źródło: Raport z badania Enterprise Strategy Group, *Navigating the Evolving AI Infrastructure Landscape*, wrzesień 2023 r.

- **Zarządzanie danymi.** Organizacje muszą wziąć pod uwagę lokalizację i wymagania w zakresie zarządzania danymi źródeł danych, które są wymagane do trenowania i obsługi modelu. Infrastruktura chmury hybrydowej sprawdzi się najlepiej, gdy dane są przechowywane lokalnie lub łatwo dostępne tam, gdzie są potrzebne, natomiast chmura publiczna może w niektórych przypadkach ułatwić gromadzenie i centralizację danych. Dane lokalne umożliwiają również organizacjom lepszą kontrolę zabezpieczeń i przestrzeganie przepisów w zakresie poufności danych. Szkolenie i utrzymywanie danych, które są aktualne, kompleksowe i bezstronne, podnosi jakość modelu LLM i generowanych odpowiedzi.

Analiza ekonomiczna Enterprise Strategy Group

Enterprise Strategy Group przeprowadziła analizę ekonomiczną, w której porównano oczekiwane koszty wnioskowania dla kilku dużych modeli językowych typu open source wykorzystujących RAG o różnej złożoności (z parametrami w liczbie 7 mld i 70 mld) oraz dla organizacji różnej wielkości (z liczbą użytkowników 5–50 tys.). Założyliśmy, że model zapewnia wewnętrzne tekstowe pytania i odpowiedzi, a wnioskowanie odbywa się tam, gdzie znajdują się dane, więc nie ma wysokiego kosztu migracji danych. Przeanalizowano wszystkie koszty związane z uruchomieniem i wnioskowaniem modeli w okresie trzech lat, w tym zapewnienie i uruchomienie infrastruktury, administrowanie systemami i płacenie za usługi w chmurze, jeśli jest to wymagane.

Infrastruktura lokalna Dell Technologies a IaaS chmury publicznej

W naszych modelach najpierw porównano oczekiwany koszt wnioskowania przy użyciu modeli LLM w tradycyjnej infrastrukturze (lokalnie, w środowiskach kolokacji, w lokalizacjach brzegowych itp.) z uruchomieniem w podobnie skonfigurowanym modelu IaaS chmury publicznej w wystąpieniach Amazon EC2. Wymagania dotyczące konfiguracji serwera węzła wnioskowania i karty graficznej NVIDIA H100 zostały dobrane do każdego obciążenia roboczego na podstawie wyników podstawowych testów wnioskowania, aby zapewnić obsługę wymagań dotyczących współbieżności przy obciążeniu zwykłym i szczytowym (na podstawie maksymalnej liczby zapytań i liczby wystąpień modelu), a także zapewnić odpowiednio małe opóźnienia i wydajność dla każdego oczekiwanego obciążenia roboczego. Następnie przeprowadziliśmy modelowanie kosztów opisanych w Tabeli 1 zarówno dla infrastruktury Dell Technologies, jak i równoważnej konfiguracji EC2.

Tabela 1. Koszty i założenia modelowane dla każdego wymagania dotyczącego obciążenia roboczego wnioskowania przy użyciu modeli LLM

Kategoria kosztów	Dell Technologies (lokalnie)	Infrastruktura chmury publicznej jako usługi (Amazon EC2)
Początkowy koszt nabycia (sprzęt i oprogramowanie)	Cena podana przez firmę Dell Technologies dla serwerów Dell PowerEdge R760xa i R660 z usługami ProDeploy i ProSupport	Nie dotyczy
Dodatkowy koszt kapitału (odsetki) i amortyzacja (korzyść)	Uwzględnione w modelu (8% WACC, 6% rocznej ulgi amortyzacyjnej)	Nie dotyczy
Koszt zasilania i chłodzenia	Obliczono na podstawie specyfikacji systemu (0,173 USD / kWh)	Nie dotyczy
Miesięczne wydatki na chmurę	Nie dotyczy	Koszty instancji EC2 p5.48xlarge obliczone na podstawie 3-letnich rabatów za rezerwacje
Licencja NVIDIA AI Enterprise / karta graficzna	Na podstawie 5-letniej licencji (proporcjonalnie)	Za wystąpienie/godz. przy założeniu 16 godzin/dzień, 5 dni w tygodniu w celu ograniczenia kosztów
Administrowanie infrastrukturą/instancjami	Modelowane (10–100% administratorów systemu w zależności od liczby węzłów)	O 66% mniej niż w modelu lokalnym
Administrowanie modelami i platformami ML	Modelowane (20–100% inżynierów ML na podstawie liczby wystąpień modelu)	Tak samo jak w modelu lokalnym

Źródło: Enterprise Strategy Group, oddział TechTarget, Inc.

Mniejszy model: LLM Mistral z 7 mld parametrów

W przypadku pierwszego porównania przyjęliśmy koszty dla mniejszego modelu zawierającego około 7 miliardów parametrów, podobnego do modelu LLM Mistral z 7 mld parametrów typu open source. Do zmiany wymagań użyliśmy na podstawie wyników testów specjalnego narzędzia. Wskazało ono konfiguracje serwera i procesora graficznego, które byłyby w stanie zapewnić średnie opóźnienie przy poleceniu wynoszące około 0,4 sekundy i szacowaną przepustowość od 2,29 do 6,86 wnioskowania na sekundę. Ogólne założenia, na przykład liczby procesorów graficznych, przedstawiono w tabeli 2.

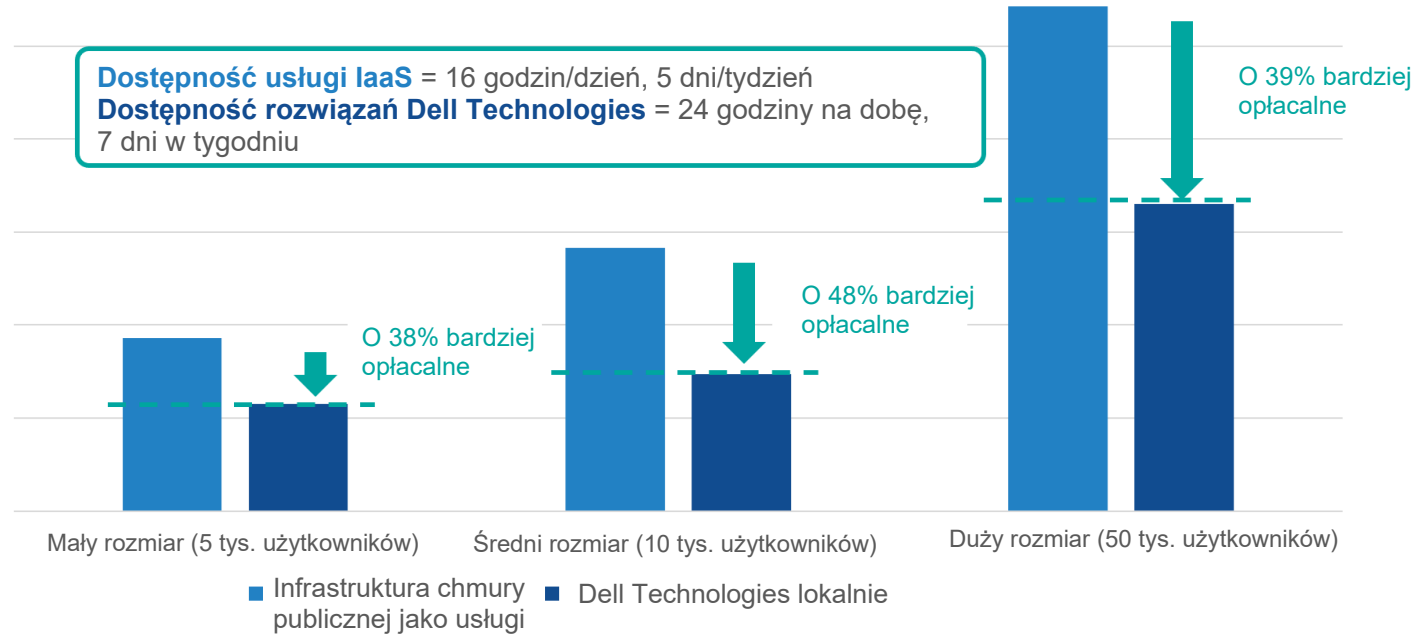
Tabela 2. Założenia konfiguracyjne dla wnioskowania modelu Mistral z 7 mld parametrów

Model LLM (liczba parametrów)	Liczba użytkowników	Liczba węzłów/instancji wnioskowania	Liczba procesorów graficznych H100
Mistral (7 mld)	5000	1	1
	10 000	1	2
	50 000	1	4

Źródło: Enterprise Strategy Group, oddział TechTarget, Inc.

Następnie wymodelowaliśmy wszystkie koszty podsumowane w tabeli 1 dla każdej konfiguracji. Jak pokazano na Rysunku 3, infrastruktura firmy Dell Technologies była 1,6–1,9 razy (38–48%) bardziej opłacalna, jeśli chodzi o stosowanie modeli wnioskowania w organizacji, a jednocześnie jest dostępna dla organizacji całodobowo.

Rysunek 2. Oczekiwany koszt wdrożenia wnioskowania dla modelu LLM Mistral z 7 mld parametrów przy użyciu RAG



Źródło: Enterprise Strategy Group, oddział TechTarget, Inc.

Większy model: LLM Llama 2 z 70 mld parametrów

Następnie oszacowaliśmy koszty większego modelu z 70 miliardami parametrów, podobnego do modelu LLM Llama 2 z 7 mld parametrów typu open source. Ponownie dopasowaliśmy wymagania za pomocą tego samego narzędzia, aby wskazać konfiguracje serwerów i procesorów graficznych, które byłyby w stanie zapewnić nieco wyższe średnie opóźnienie przy poleceniu wynoszące około 1,8 sekundy i szacowaną przepływność od 2,29 do 22,86 wnioskowania na sekundę. Ogólne założenia, na przykład liczby procesorów graficznych, przedstawiono w tabeli 3.

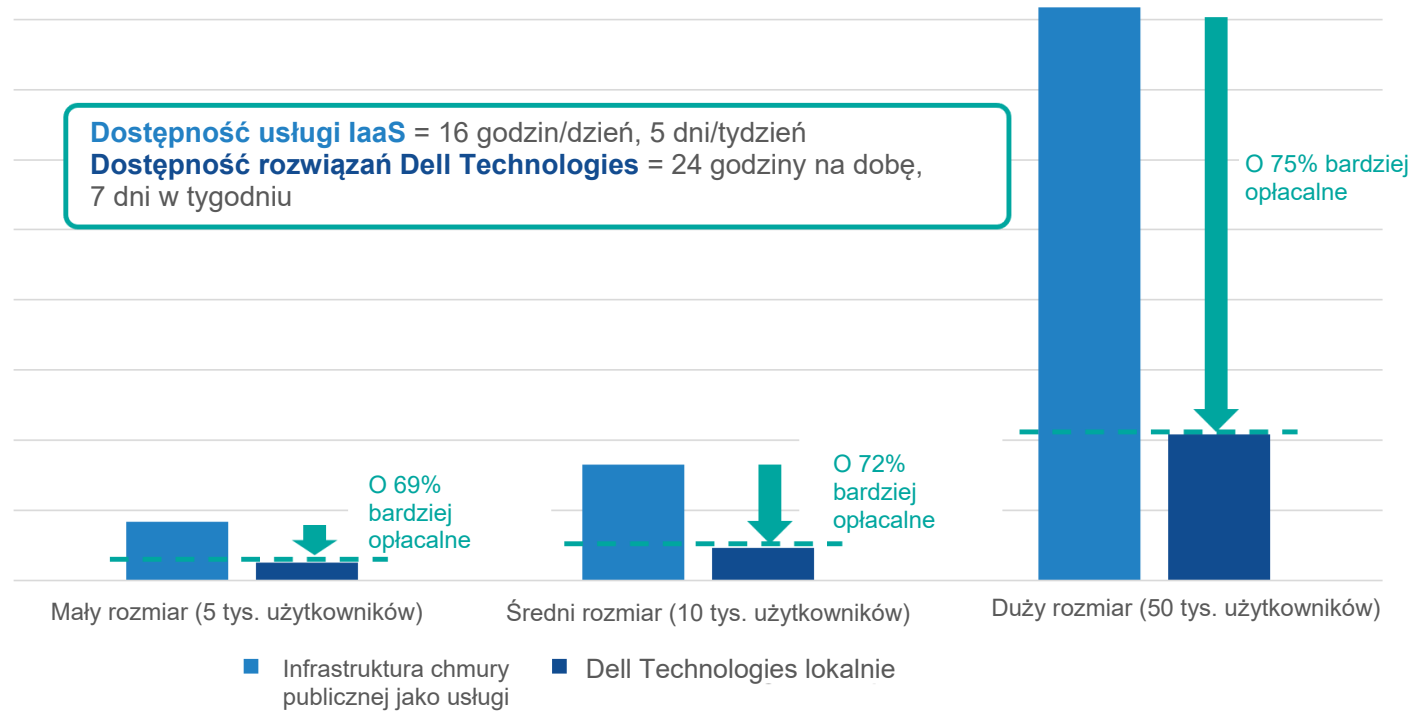
Tabela 3. Założenia konfiguracyjne dla wnioskowania modelu Llama 2 z 70 mld parametrów

Model LLM (liczba parametrów)	Liczba użytkowników	Liczba węzłów/instancji wnioskowania	Liczba procesorów graficznych H100
Llama 2 (70 mld)	5000	2	8
	10 000	4	16
	50 000	20	80

Źródło: Enterprise Strategy Group, oddział TechTarget, Inc.

Po ponownym oszacowaniu wszystkich kosztów podsumowanych w Tabeli 1 dla każdej z przedstawionych powyżej konfiguracji stwierdziliśmy, że infrastruktura Dell Technologies jest 3,3–4,0 razy (69–75%) bardziej opłacalna do wnioskowania w organizacji, a jednocześnie jest dostępna dla organizacji całodobowo.

Rysunek 3. Oczekiwany koszt wdrożenia wnioskowania dla modelu LLM Llama 2 z 70 mld parametrów przy użyciu RAG

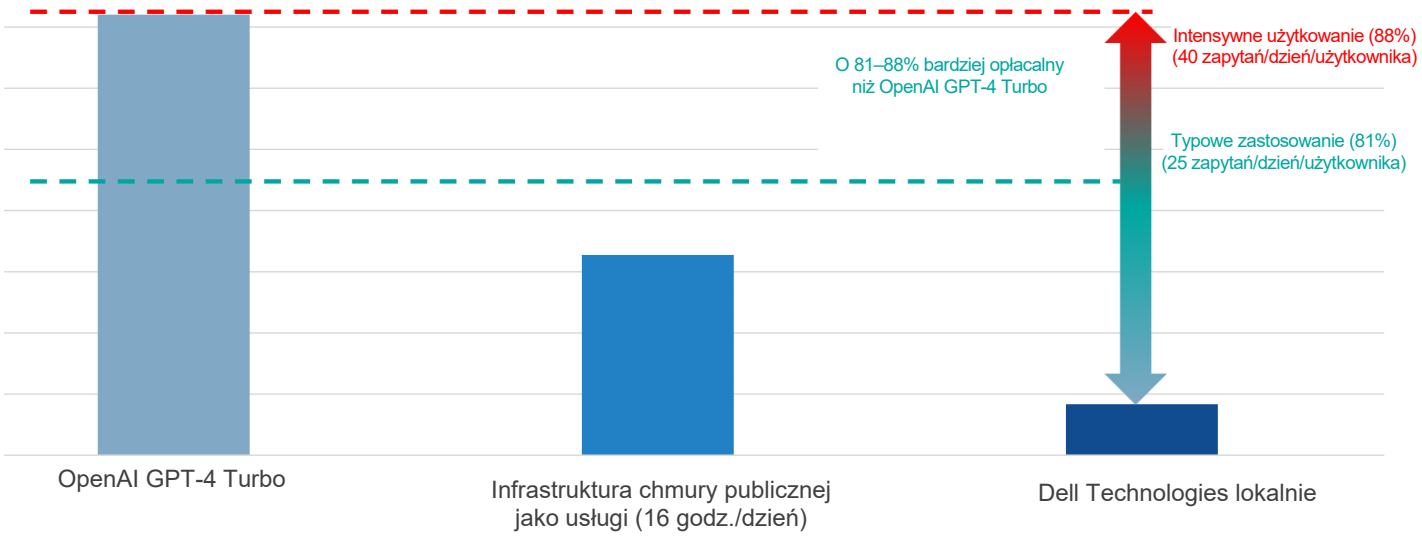


Źródło: Enterprise Strategy Group, oddział TechTarget, Inc.

Infrastruktura lokalna Dell Technologies a usługa generatywnej sztucznej inteligencji oparta na interfejsie API

Następnie porównaliśmy oczekiwane koszty dla dużej organizacji przy założeniu równoważnego modelu z 70 mld parametrów dla jej 50 tys. użytkowników przy użyciu uznanej usługi generatywnej sztucznej inteligencji opartej na interfejsie API OpenAI GPT-4 Turbo, która jest opłacalna w przeliczeniu na „token" wejściowy i wyjściowy. Tekstowe pytania i odpowiedzi wymagają umiarkowanej intensywności tokenu na zapytanie, nie mają wielu wariancji w obciążeniu szczytowym i osiągają względną równowagę między liczbą wymaganych tokenów wejściowych i wyjściowych. Przyjęliśmy łącznie 1500 tokenów (wejściowych i wyjściowych) na zapytanie, ze średnią około 25 zapytań dziennie na użytkownika zarówno w przypadku rozwiązań lokalnych, jak i opartych na interfejsie API. Na podstawie analizy informacji publicznych stwierdziliśmy, że jest to umiarkowana liczba zapytań na użytkownika, przy czym mniej popularne organizacje generują mniej zapytań na użytkownika, a bardziej popularne organizacje średnio aż 40 zapytań na użytkownika dziennie. Nasze obliczenia GPT-4 Turbo przewidywały koszt około 12,50 USD / użytkownika / miesiąc, co wypada korzystnie w porównaniu z narzędziami AI opartymi na pakiecie, które mogą kosztować około 30 USD / użytkownika / miesiąc. Przy tych założeniach stwierdziliśmy, że infrastruktura lokalna firmy Dell Technologies może zapewnić wnioskowanie 5,4–8,6 razy (81–88%) bardziej opłacalnie niż korzystanie z usługi opartej na interfejsie API, zapewniając możliwości generatywnej sztucznej inteligencji za jedyne około 2,31 USD / użytkownika / miesiąc.

Rysunek 4. Oczekiwany koszt wdrożenia wnioskowania dla LLM Llama 2 z 70 mld parametrów dla 50 tys. użytkowników



Źródło: Enterprise Strategy Group, oddział TechTarget, Inc.

Kwestie do rozważenia

Chociaż modele Enterprise Strategy Group powstają w dobrej wierze na podstawie konserwatywnych, wiarygodnych i zweryfikowanych założeń, żaden pojedynczy modelowy scenariusz nigdy nie będzie reprezentował każdego potencjalnego środowiska. Oszczędności klienta będą zależeć od konkretnego zastosowania, charakteru danych, poziomu wiedzy eksperckiej oraz wymagań dotyczących modelu i infrastruktury. Enterprise Strategy Group zaleca przeprowadzenie własnej analizy dostępnych produktów i skonsultowanie się z firmą Dell Technologies w celu poznania i omówienia różnic między rozwiązaniami sprawdzonymi w ramach własnych testów potwierdzających słuszność koncepcji.

Dell Technologies do wnioskowania przy użyciu modeli LLM

Firma Dell Technologies pomaga organizacjom w łatwym wprowadzaniu sztucznej inteligencji do przetwarzania danych, niezależnie od tego, gdzie się one znajdują. Oznacza to oferowanie najszerszego portfolio usług sztucznej inteligencji — od komputerów stacjonarnych, przez centra danych, po chmurę — dzięki czemu organizacje mogą odpowiednio dostosować swoje inwestycje i wykorzystać dane do budowy fabryk sztucznej inteligencji oraz wdrażania zastosowań sztucznej inteligencji w sposób wydajny, bezpieczny i zrównoważony. Aby to osiągnąć, firma Dell zapewnia dostęp do kompleksowego portfolio usług i szerokiego, otwartego ekosystemu partnerów, którzy pomagają organizacjom na każdym etapie wdrażania sztucznej inteligencji, niezależnie od tego, czy opracowują strategie dotyczące sztucznej inteligencji czy przyspieszają i rozbudowują inwestycje w generatywną sztuczną inteligencję.

W przypadku organizacji borykających się z zagrożeniami bezpieczeństwa danych, kwestiami zgodności ze przepisami, silosami danych i niezweryfikowanymi zestawami danych Dell Professional Services dla generatywnej sztucznej inteligencji mogą pomóc w osiągnięciu konsensusu między liderami biznesowymi i informatycznymi w zakresie priorytetowych zastosowań, opracowaniu planu realizacji celów, przygotowaniu danych przedsiębiorstwa do integracji z platformą LLM, przyspieszeniu dojrzałości cyberbezpieczeństwa i ustanowieniu platformy sztucznej inteligencji dostosowanej do konkretnych potrzeb biznesowych. Ponadto dzięki rozwiązaniom Dell APEX organizacje mogą subskrybować rozwiązania sztucznej inteligencji i optymalizować je pod kątem zastosowań wielochmurowych.

Aby dowiedzieć się więcej o rozwiązaniach firmy Dell, odwiedź stronę firmy Dell poświęconą sztucznej inteligencji.

Wnioski

Rozszerzone wykorzystanie generatywnej sztucznej inteligencji w niemal każdym obszarze działalności jest kluczowym czynnikiem zapewniającym usprawnienie operacji i przyszły sukces. Z badań przeprowadzonych przez Enterprise Strategy Group wynika, że najważniejsze obszary, w których organizacje stosują obecnie generatywną sztuczną inteligencję, obejmują badania, marketing, programowanie, tworzenie produktów i informatykę, a potencjał wykorzystania w każdym obszarze ma wzrosnąć[4]. Organizacje mogą osiągać bardziej wymierne i znaczące wyniki poprzez trenowanie i wnioskowanie w oparciu o własną, dostosowaną wersję modelu LLM.

Istnieje kilka metod wdrażania modeli LLM, a każda z nich zapewnia korzyści w konkretnych zastosowaniach i przy określonych wymaganiach. W przypadku większych organizacji z tysiącami użytkowników gotowych do korzystania z możliwości niestandardowych rozwiązań LLM infrastruktura firmy Dell Technologies może zapewnić wydajny model LLM nawet 4 razy tańszy niż IaaS i 8 razy tańszy niż OpenAI GPT-4 Turbo. Enterprise Strategy Group zdecydowanie zaleca, aby firmy wdrażające duże modele językowe w swoich organizacjach rozważyły skorzystanie z opłacalnych technologii i specjalistycznych usług oferowanych przez firmę Dell Technologies w celu uzyskania oczekiwanych wyników, przyspieszenia inicjatyw w zakresie generatywnej sztucznej inteligencji i skrócenia czasu potrzebnego do osiągnięcia oczekiwanych oszczędności.

[4] Źródło: raport badawczy Enterprise Strategy Group, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, sierpień 2023 r.

Informacje o Enterprise Strategy Group
Enterprise Strategy Group firmy TechTarget zapewnia ukierunkowane i praktyczne analizy rynkowe, badania popytu, usługi doradztwa analitycznego, wskazówki dotyczące strategii GTM, weryfikacje rozwiązań oraz niestandardowe treści wspierające kupno i sprzedaż technologii dla przedsiębiorstw.

contact@esg-global.com
www.esg-global.com