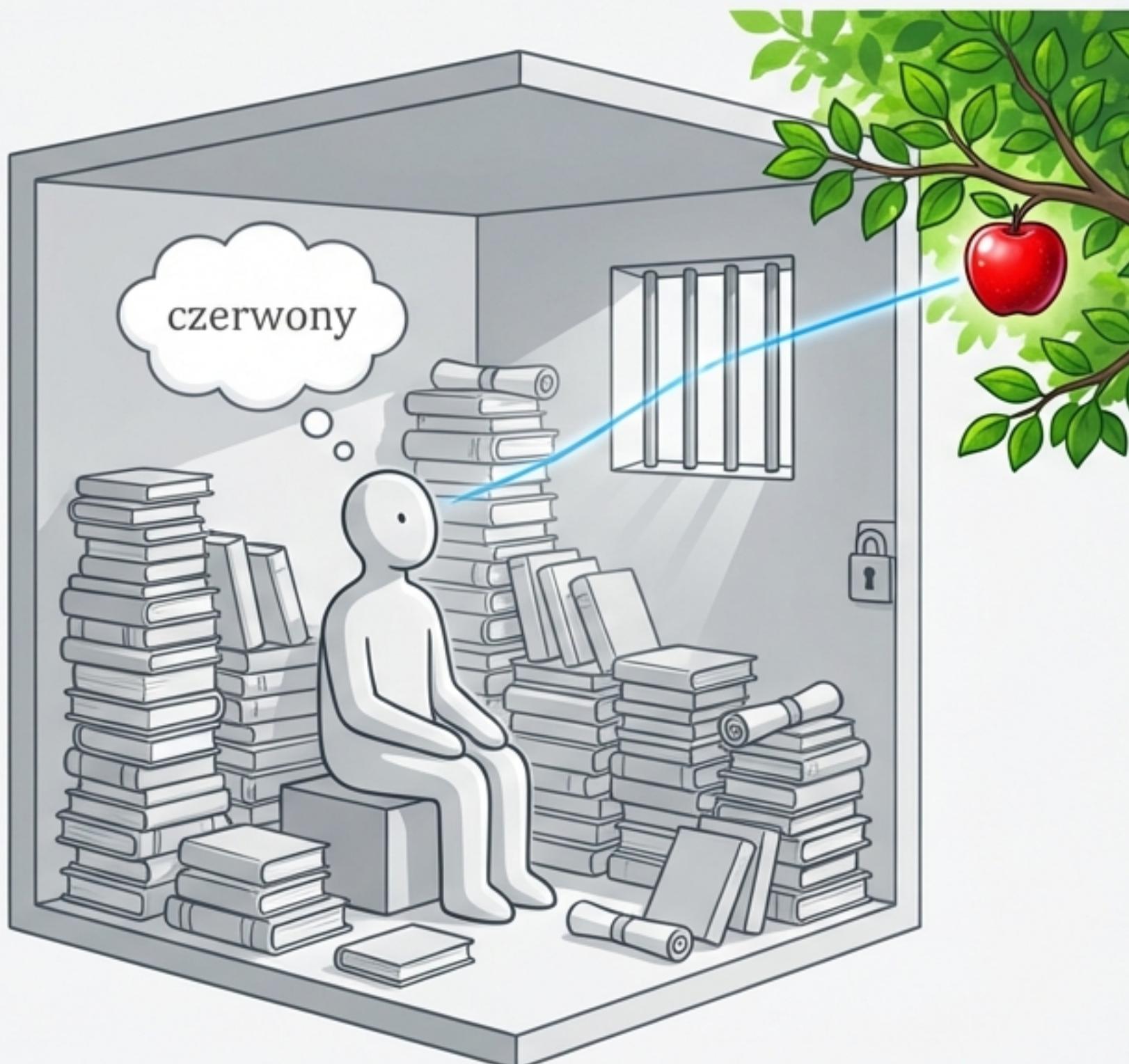


Problem Uziemienia: Dlaczego sam język to za mało?



Główna idea i analogia:

Duże Modele Językowe (LLM) są jak 'osoba zamknięta w pokoju pełnym książek'. Mogą przeczytać wszystko o kolorze czerwonym, ale nigdy go nie widziały.

Definicja problemu:

Uziemienie (grounding): Wiedza LLM jest czysto abstrakcyjna, pozbawiona odniesienia do realnego świata. Brak jest kluczowego połączenia symboli (słów) z sensorycznym doświadczeniem.

Kontekst multymodalny:



Wzrok



Słuch



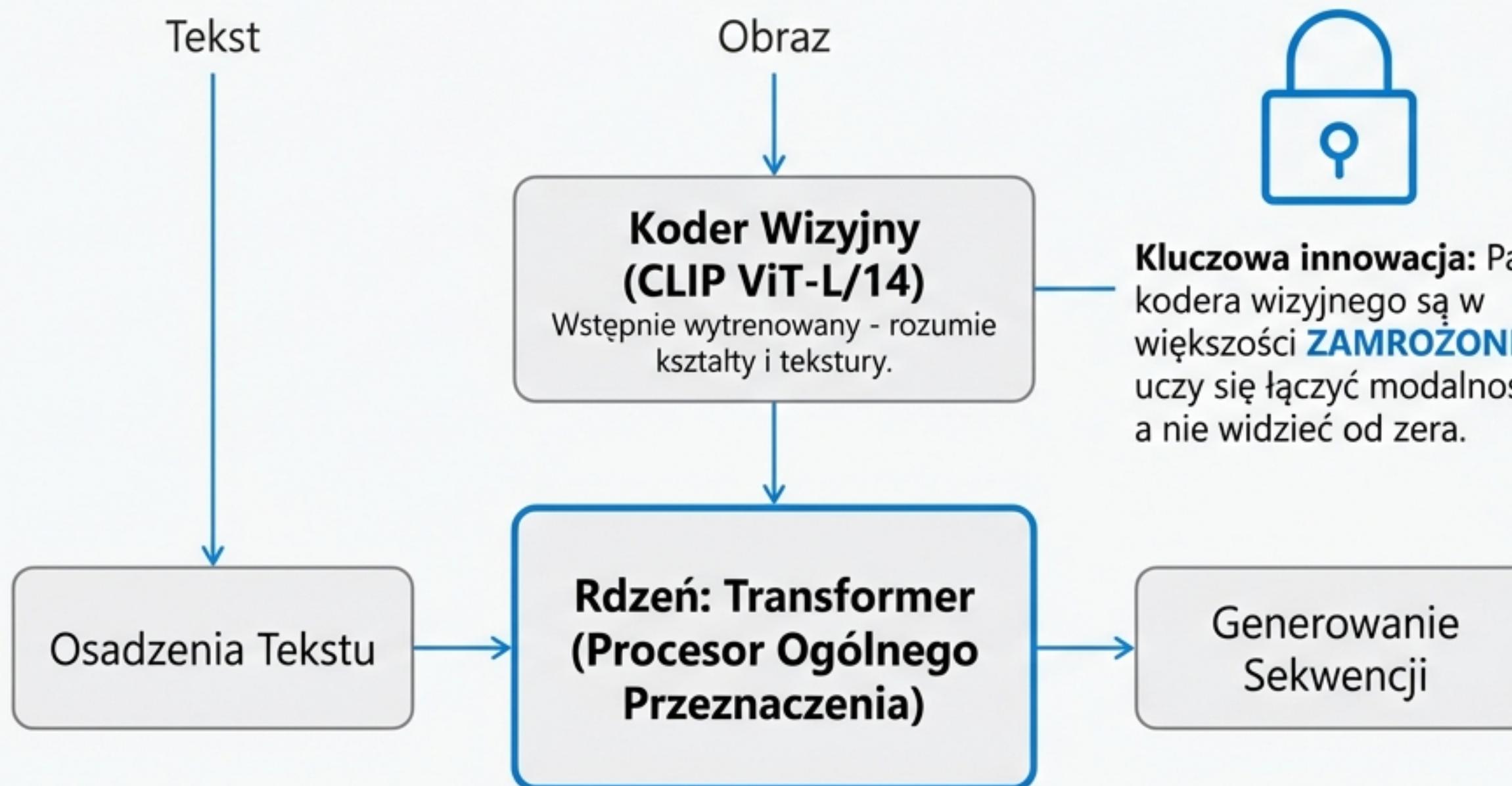
Interakcja

Świat jest z natury multymodalny. Sam tekst to tylko wąski wycinek doświadczenia.

Główna teza:

Nasza misja: **Połączenie percepji z LLM**, tworząc jeden zintegrowany system – Multymodalny Duży Model Językowy (MLLM).

Architektura KOSMOS-1: Język jako uniwersalny interfejs



Dekoder Transformera działa jako uniwersalny interfejs, który przetwarza osadzenia tekstu i obrazu w jednej, ujednoliconej sekwencji.

Skala: 1,6 miliarda parametrów

Stabilność Treningu: Architektura **Magneto** zapewniająca stabilność przy dużej skali.

Długie Sekwencje: Technika **xPos** do efektywnego przetwarzania bardzo długich sekwencji.

Paliwo dla Umysłu: Trening na wielkoskalowych korpusach multimedialnych



Czyste korpusy tekstowe

Źródła: The Pile, Common Crawl

Cel: Opanowanie języka na najwyższym poziomie.



Pary obraz-podpis

Źródła: LAION-2B, COYO-700M

Cel: Nauczenie podstawowych skojarzeń wizualno-tekstowych na masową skalę.



Dane przeplatane (Interleaved Data)

Źródła: Strony internetowe z przemieszczanym tekstem i obrazami.

Cel: Nauka głębokiego **rozumienia kontekstowego**.

Dlaczego dane przeplatane są kluczowe?

Ten sam obraz lodowca ma zupełnie inne znaczenie w artykule zatytułowanym '**Skutki globalnego ocieplenia**' niż w galerii '**Piękne arktyczne krajobrazy**'.

To właśnie 'przeplatanie danych' umożliwia głęboką, kontekstową integrację, a nie tylko powierzchowne łączenie modalności.

Rozumienie Scen: Od prostego opisu do wnioskowania przyczynowego



Pytanie: Dlaczego ten chłopiec płacze?

Starsze modele (opis)

„Na zdjęciu jest chłopiec i hulajnoga.”

KOSMOS-1 (wnioskowanie)

„Ponieważ zepsuła mu się hulajnoga.”

KOSMOS-1 wykazuje zdolność do wnioskowania o **przyczynie i skutku**, rozumienia stanów emocjonalnych i dostrzegania relacji między obiekty.

To jest **rozumienie sceny**, a nie tylko jej opis – fundamentalny krok w kierunku prawdziwej inteligencji wizualnej.

Pojmowanie Abstrakcji: Jak maszyna uczy się, co jest żartem

Pytanie: Dlaczego to zdjęcie jest zabawne?



Odpowiedź KOSMOS-1: „Kot ma na sobie maskę, która nadaje mu uśmiech.”

- **Zrozumienie absurdalności:** Model rozpoznaje konceptualne niedopasowanie leżące u podstaw humoru.
- **Pojmowanie konceptu żartu:** Wykracza poza proste rozpoznawanie obiektów ('kot', 'kartka papieru').
- **Zdolności poznawcze wyższego rzędu:** Dowód na rozumienie subtelnych, ludzkich pojęć.

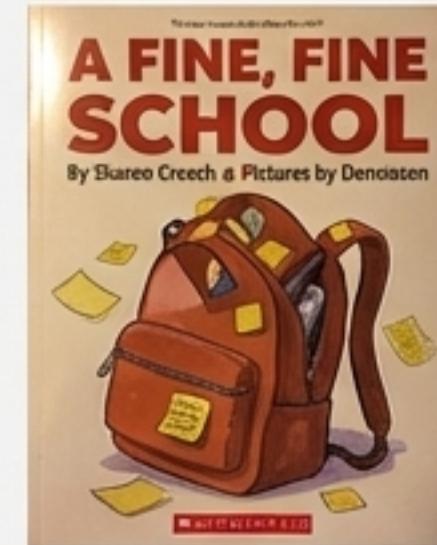
NLP bez OCR: Czytanie tekstu bezpośrednio z obrazu

Top Section - Explanation

-  **Nowe Podejście:** Model czyta tekst **bezpośrednio z obrazu**, traktując go jako część sceny wizualnej, bez potrzeby stosowania zewnętrznego oprogramowania **OCR** (Optical Character Recognition).
-  **Problem z Tradycyjnym OCR:** Stare podejście jest podatne na błędy przy nietypowych czcionkach, cieniach i zagięciach; myli znaki jak 'O' z '0' lub 'I' z '1'.

Bottom Section - Dwa Praktyczne Przykłady

Czytanie z Okładki



Pytanie: Jaki jest tytuł tej książki?

Odpowiedź KOSMOS-1: A Fine, Fine School

Rozumienie Interfejsu Graficznego (GUI)



Pytanie: Chcę ponownie uruchomić komputer. Który przycisk mam kliknąć?

Odpowiedź KOSMOS-1: OK

Ekspert na Żądanie: Klasyfikacja na podstawie dynamicznych opisów



Wizualizacja: Dzięcioł na zdjęciu.



Który opis pasuje do dzięcioła na zdjęciu?

Dzięcioł trójpalczasty

Czarne i białe paski na całym ciele, żółta korona.



Dzięcioł kosmaty

Białe plamki na czarnych skrzydłach i trochę czerwieni na koronie.



Kluczowa Cecha

To demonstruje **rozumowanie w oparciu o dynamicznie dostarczoną wiedzę**, a nie tylko odtwarzanie zapamiętanych faktów.

Analogia

To jak dać AI podręcznik do użycia w czasie rzeczywistym.

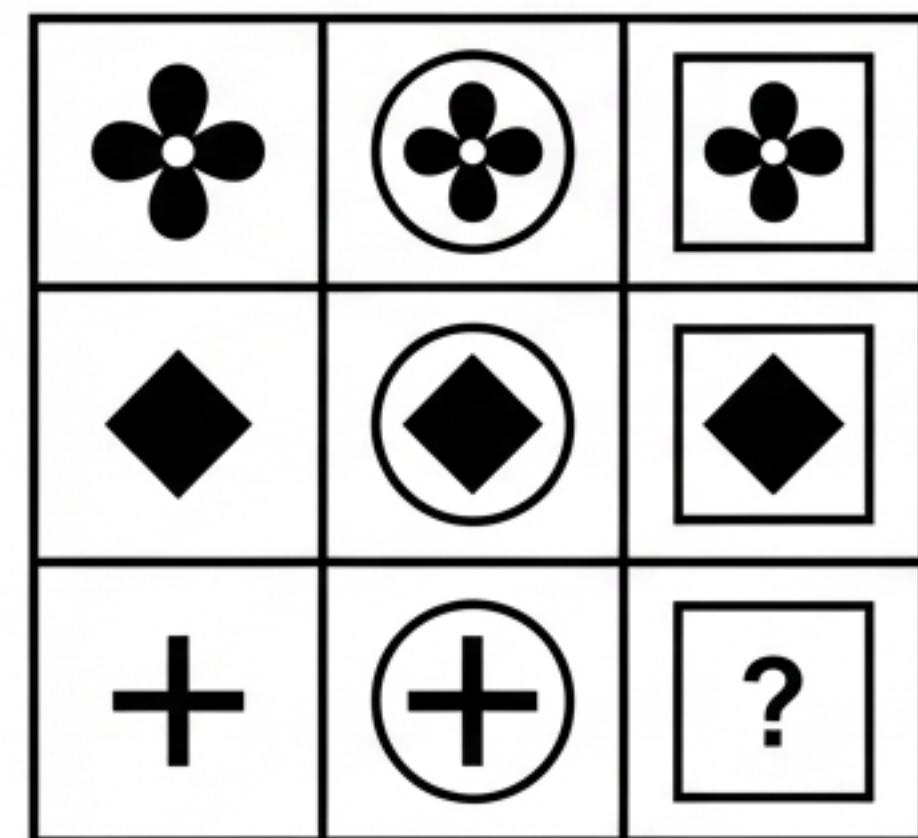
Zastosowania

Potencjalne zastosowania: diagnostyka medyczna, kontrola jakości w fabrykach, identyfikacja dzikiej przyrody.

Załążki Abstrakcji: Zdolność rozumowania nielingwistycznego w teście Matryc Ravena

Wprowadzenie do Testu

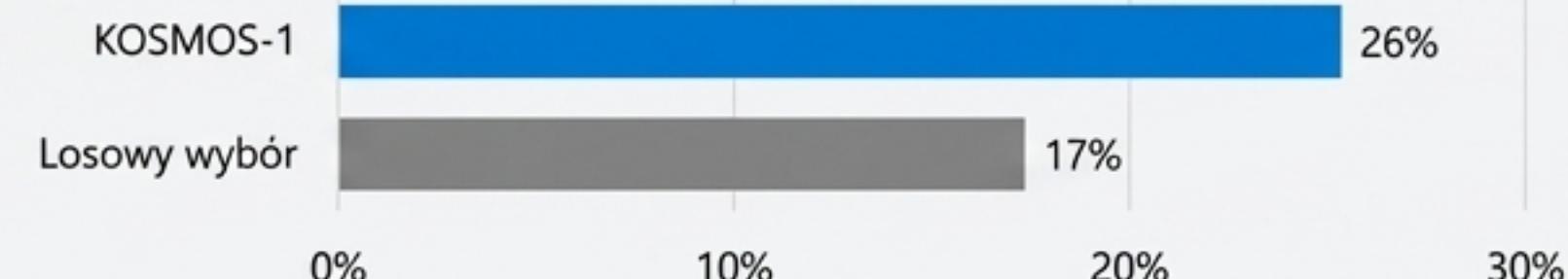
Progresywne Matryce Ravena to test IQ polegający na uzupełnieniu brakującego elementu w siatce abstrakcyjnych wzorów poprzez odkrycie nielingwistycznych reguł (np. rotacja, dodawanie elementów).



Wyniki i Interpretacja

26% Dokładność KOSMOS-1

17% Losowy wybór



To nie jest jeszcze poziom ludzki, ale wynik jest o 9 p.p. wyższy od losowego. To **statystycznie istotny sygnał**, że model uchwycił część logicznej struktury problemu.

Trening na danych wizualnych i tekstowych prowadzi do **wyłaniania się (emergence) ogólnych zdolności poznawczych**, które nie były bezpośrednio programowane ani trenowane.

Multimodalny Tok Myślenia: Poprawa rozumowania przez generowanie uzasadnień

Definicja Techniki (Multimodal Chain-of-Thought)

Technika wzorowana na ludzkim procesie myślowym: model najpierw tworzy szczegółowy opis obrazu (uzasadnienie), a dopiero potem wykorzystuje go do odpowiedzi na konkretne pytanie. Rozbija to złożony problem na mniejsze kroki.



Przykład i Wyniki (Rendered SST-2)

Zadanie: Klasyfikacja sentymentu tekstu przedstawionego jako obraz.

Wynik standardowego promptingu: 67.1% dokładności

Wynik z użyciem Multimodal CoT: **72.9%** dokładności

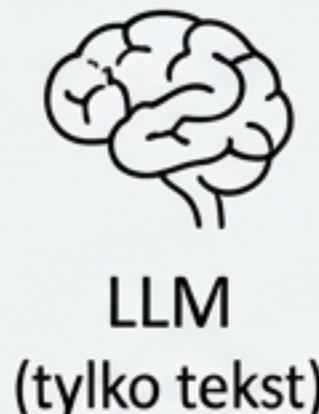
Wniosek: Wzrost o **5.8 punktu procentowego** dowodzi, że generowanie pośrednich kroków rozumowania znacząco poprawia wydajność w złożonych zadaniach.

Transfer Międzymodalny: Jak widzenie świata uczy zdrowego rozsądku

Integracja percepji wizualnej wzbogaca model o wiedzę zdroworozsądkową (np. o kolorach i rozmiarach), która jest trudna lub niemożliwa do zdobycia wyłącznie z tekstu.

Eksperyment Porównawczy

Porównaliśmy KOSMOS-1 (MLLM) z identycznie trenowanym modelem tylko językowym (LLM) na zadaniach wymagających wiedzy o świecie.



Eksperyment
Porównawczy



Wyniki (Zero-Shot) - Visual Commonsense Reasoning

Zadanie	LLM (tylko tekst)	KOSMOS-1 (tekst + wizja)
Rozmiar Względny (np. 'Czy sofa jest większa od kota?')	92.7%	94.2%
Pamięć Kolorów (np. 'Jakiego koloru jest niebo?')	61.4%	76.1%

Analiza

KOSMOS-1 znaczco przewyższa LLM, zwłaszcza w zadaniach dotyczących kolorów. Model "widział" niezliczone obiekty i ich właściwości, zamiast polegać wyłącznie na opisach tekstowych, które często pomijają tak oczywiste dla ludzi informacje.

KOSMOS-1: Konsekwencje i kolejny krok w stronę AGI



Głębsze Rozumienie Świata

MLLM zdobywają wiedzę zdroworozsądkową (o kolorach, rozmiarach, relacjach przestrzennych), która jest nieobecna w tekście. To **uziemienie wiedzy w rzeczywistości**.



Nowe Horyzonty Zastosowań

Integracja percepcji otwiera drzwi do zaawansowanych zastosowań w robotyce, inteligentnym przetwarzaniu dokumentów i tworzeniu **świadomych kontekstu asystentów AI**.



Uniwersalny Interfejs

Zdolność do odczytywania interfejsów graficznych (GUI) to krok w stronę AI, która może wchodzić w interakcję z **dowolnym oprogramowaniem tak, jak człowiek**.

KOSMOS-1 to nie tylko kolejny model. To dowód na to, że prawdziwa inteligencja wymaga zjednoczenia języka z percepcją, co stanowi fundamentalny krok na drodze do Ogólnej Sztucznej Inteligencji (AGI).