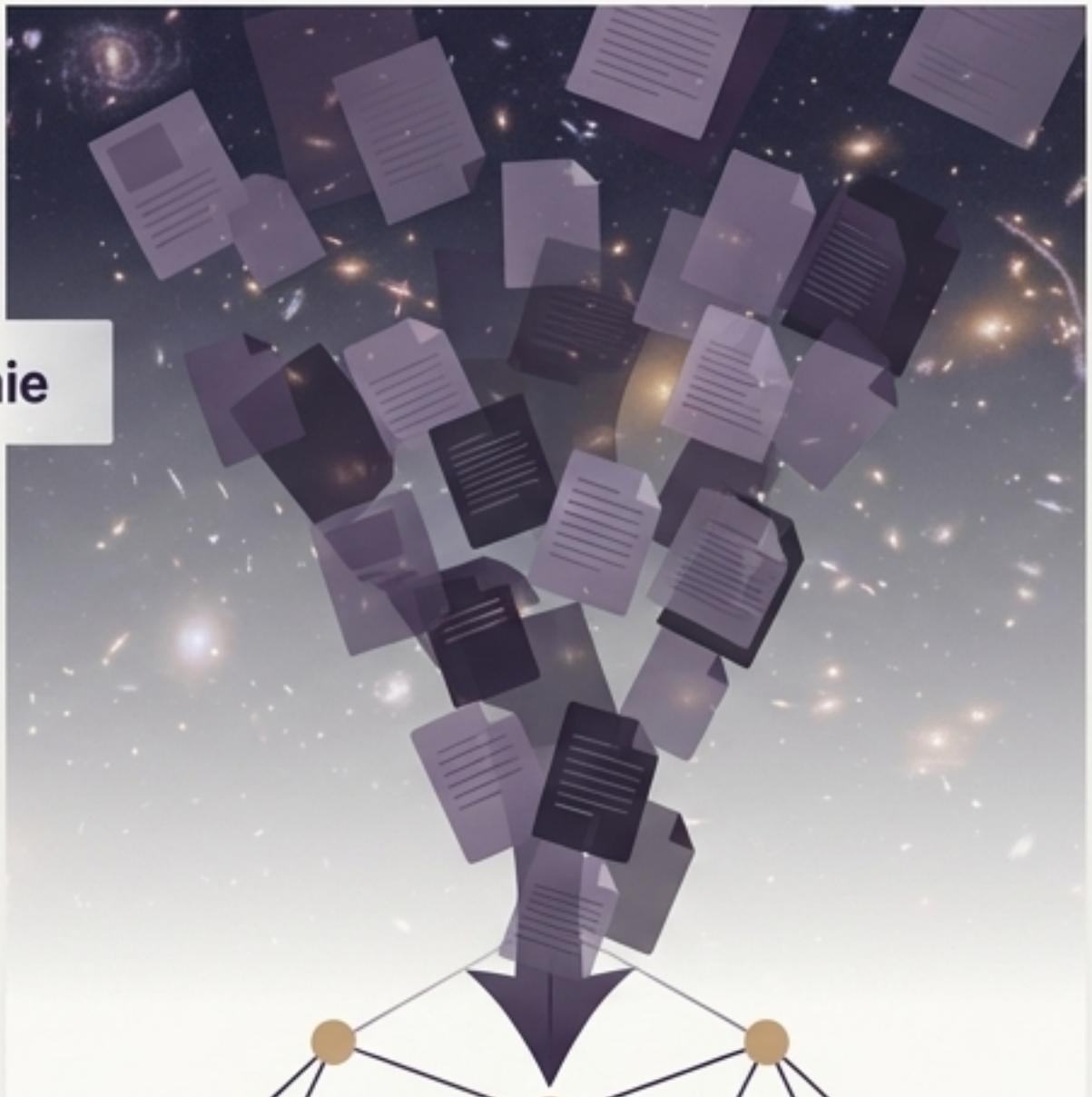


# Tytuł: Kryzys Informacyjny w Nauce

## Podtytuł: Od Przeciążenia do Zrozumienia

- **Wizja z 1945 roku:** Już Vannevar Bush w eseju „As We May Think” ostrzegał, że „publikacje wykroczyły daleko poza naszą obecną zdolność do rzeczywistego wykorzystania zapisu”.
- **Dzisiejsza Rzeczywistość:** Problem eksplodował. W maju 2022 roku do samego serwisu arXiv trafiało średnio 516 prac naukowych dziennie.
- **Paradygmat „Wskaż i Przechowaj”:** Obecne narzędzia, jak wyszukiwarki, nie organizują wiedzy. Wskazują jedynie na wtórne źródła (np. Wikipedia), których tworzenie wymaga kosztownego wkładu ludzkiego.
- **Teza Galactiki:** Wielki model językowy (LLM) może stać się pojedynczym, neuronowym interfejsem do wiedzy naukowej, zdolnym do przechowywania, łączenia i rozumowania na temat tej wiedzy.
- **Zmiana Paradygmatu:** Zamiast przeszukiwać zewnętrzną bibliotekę, możemy zapytać bibliotekarza, który przeczytał wszystko i rozumie wzajemne powiązania, potrafiąc syntetyzować wiedzę i odkrywać ukryte połączenia.

Przeciążenie

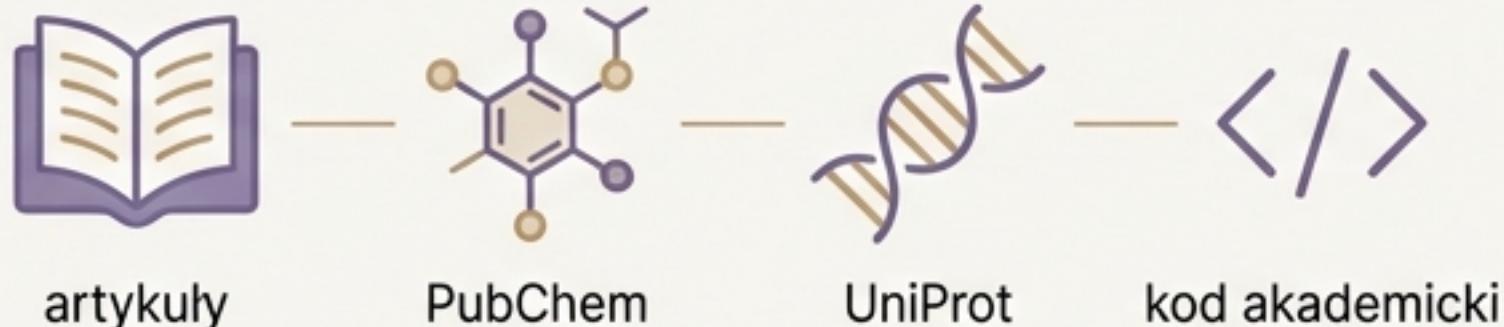


Zrozumienie

# Tytuł: Kuratorowany Korpus – Jakość ponad Ilość

Podtytuł: Filozofia Archiwisty, nie Odkurzacza Sieciowego

## Korpus Galactiki



- **Precyzyjny Zbiór Danych:** Galactica została wytrenowana na **106 miliardach tokenów** starannie wyselekcjonowanych, wysokiej jakości danych naukowych.
- **48+ milionów** artykułów naukowych, podręczników i notatek z wykładów.
- Specjalistyczne bazy danych, jak **PubChem** (związki chemiczne) i **UniProt** (sekwencje białek).
- Referencje, encyklopedie i kod akademicki.
- „**Podejście Normatywne**”: Świadoma decyzja o tym, co stanowi wartościową wiedzę.

## Typowy Korpus LLM

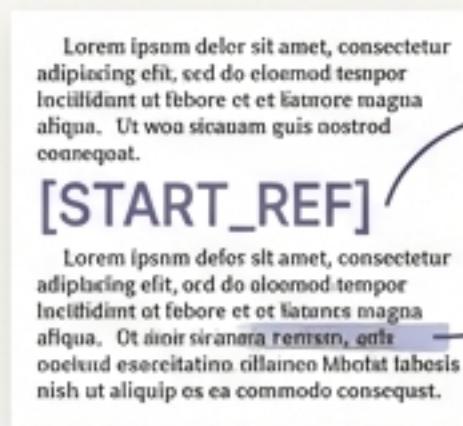


- **Kontrast z Innymi Modelami:** Dla porównania, 50% korpusu modelu PaLM (Google) to konwersacje z mediów społecznościowych, mające ograniczoną wartość transferu wiedzy do zadań naukowych.
- **Paradygmat Niekuratorowany:** Bezkrityczne indeksowanie sieci, które pochłania wszystko bezrefleksyjnie.

# Tytuł: Nauka Języków Nauki – Specjalne Tokeny

## Podtytuł: Od Tekstu po Chemicę i Biologię w Jednym Modelu

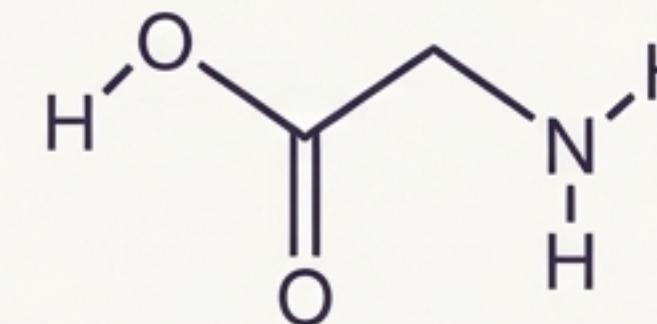
**Ponad Słowami:** Galactica uczy się różnych „języków” nauki, traktując je jako ustrukturyzowane objekty, a nie losowe ciągi znaków, dzięki specjalnym tokenom:



**Cytowania '[START\_REF]':**  
Pozwalają modelowi uczyć się **niejawnego grafu cytowań** i relacji między pracami.



**Biologia '[START\_AMINO]':**  
Traktuje sekwencje aminokwasów jako fundamentalne jednostki biologiczne.



**Chemia '[START\_SMILES]':**  
Uczy modela składni molekularnej z tokenizacją na poziomie znaków, rozumiejąc strukturę chemiczną.

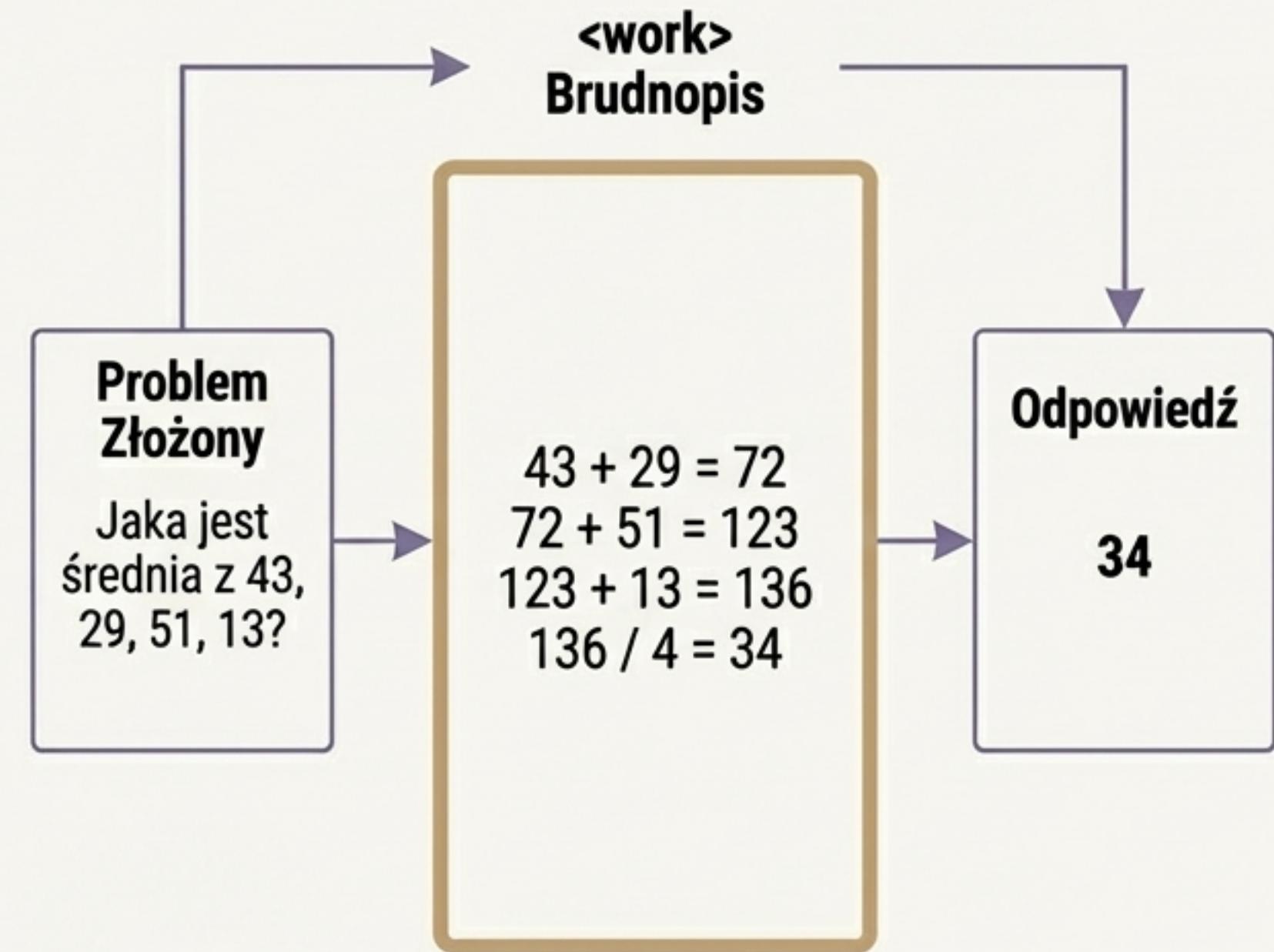
**Matematyka i Liczby:** Cyfry i operatory arytmetyczne są dzielone na indywidualne tokeny dla precyzyjnego przetwarzania.

**Interdyscyplinarne Wnioskowanie:** Umożliwia to modelowi płynne przechodzenie i łączenie wiedzy między tekstem, chemią, biologią i matematyką w ramach jednego, spójnego kontekstu.

# Tytuł: Token `<work>` – Myślenie Krok po Kroku

## Podtytuł: Cyfrowy Brudnopsis na Złożone Problemy

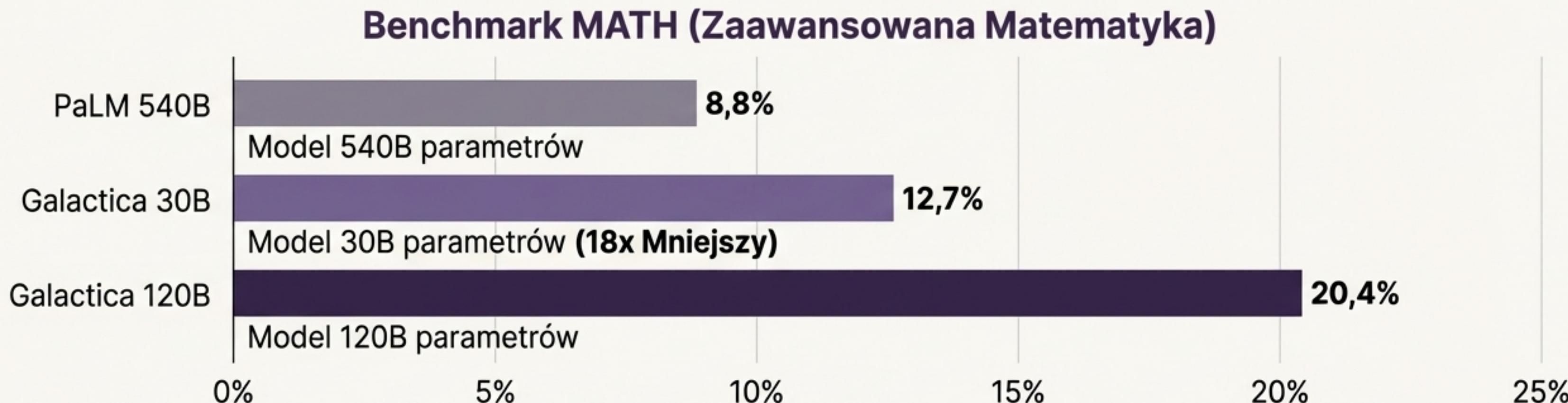
- **Problem:** Standardowe modele Transformer nie posiadają jawnej pamięci roboczej, co ogranicza ich zdolność do wieloetapowego rozumowania w jednym przebiegu.
- **Inspiracja:** Ludzie nie rozwiązują złożonych problemów w głowie natychmiast – używają notatek do rozpisania kolejnych kroków.
- **Rozwiążanie:** Token **<work>** działa jak „cyfrowy brudnopsis”. Model generuje w jego ramach pośrednie etapy rozumowania przed podaniem ostatecznej odpowiedzi.
- **Kluczowa Innowacja:** To fundamentalne ulepszenie dla zdolności rozumowania matematycznego i logicznego, naśladujące ludzki proces myślowy i tworzące bardziej wiarygodne, transparentne wyniki w stylu „chain-of-thought”.
- **Trening:** Model uczy się tej zdolności dzięki specjalnie przygotowanym zbiorom danych (np. GSM8k), w których rozumowanie krok-po-kroku jest opakowane w tokeny **<work>**.



# Tytuł: Benchmarkki Matematyczne – Siła Rozumowania

## Podtytuł: Mniejszy Model, Lepsze Wyniki Dzięki Specjalizacji

- **MMLU (Egzaminy Akademickie z Matematyki):** Zastosowanie tokenu `<work>` znacznie podnosi wydajność.
  - **Galactica 120B (`<work>`): 41,3%**
  - Chinchilla 70B (5-shot): 35,7%
- **Zaskakująca Skuteczność:** Nawet **Galactica 30B**, model **18x mniejszy** od PaLM 540B, osiąga na benchmarku MATH lepszy wynik.



**Kluczowy Wniosek:** Jakość danych treningowych i specjalizowana architektura (tokeny `<work>`) mają fundamentalne znaczenie dla zdolności **rozumowania** – większe nie zawsze znaczy lepsze.

# Tytuł: Sondowanie Wiedzy – Praktyczne Testy Dziedzinowe

## Podtytuł: Od Równań LaTeX po Interpretowalną Chemię

### Generowanie Równań LaTeX

Na prośbę o nazwę równania (np. „równanie różniczkowe Bessela”), model generuje poprawny kod LaTeX.

**Galactica 120B: 68,2%** dokładności

GPT-3 (text-davinci-002): 49,0%

Prompt: Równanie różniczkowe Bessela

$$x^2 \frac{d^2y}{dx^2} + x \frac{dy}{dx} + (x^2 - \alpha^2)y = 0$$

### Przewidywanie Reakcji Chemicznych

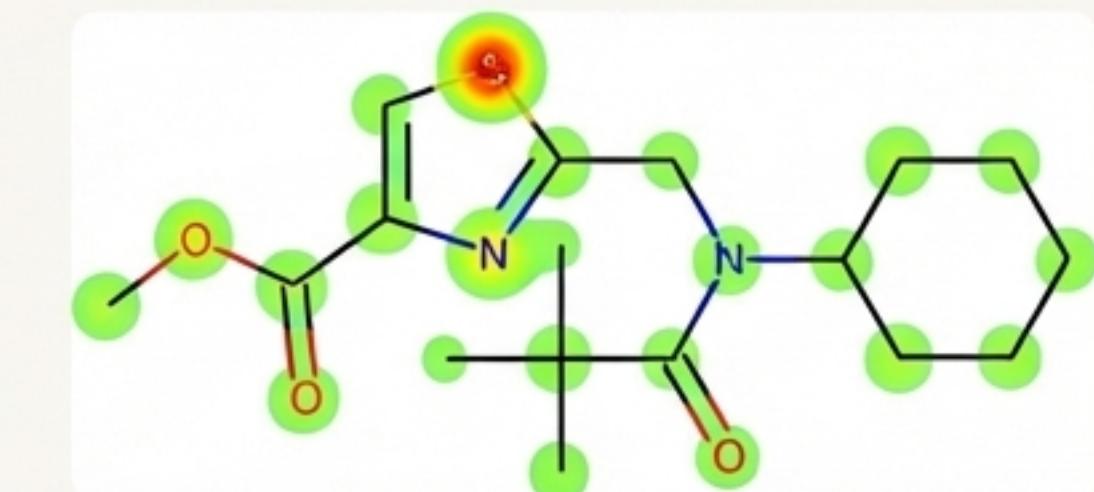
Dokładność **43,1%** w przewidywaniu produktów na podstawie reagentów.

Prompt: NaCl + H2SO4 →



### Interpretowalny Mechanizm Uwagi

Model wykazuje chemiczne „rozumienie”. Generując słowo „tiazol”, uwaga koncentruje się na **atomie siarki** w pierścieniu. To dowód, że model uczy się fundamentalnych zasad, a nie tylko wzorców.

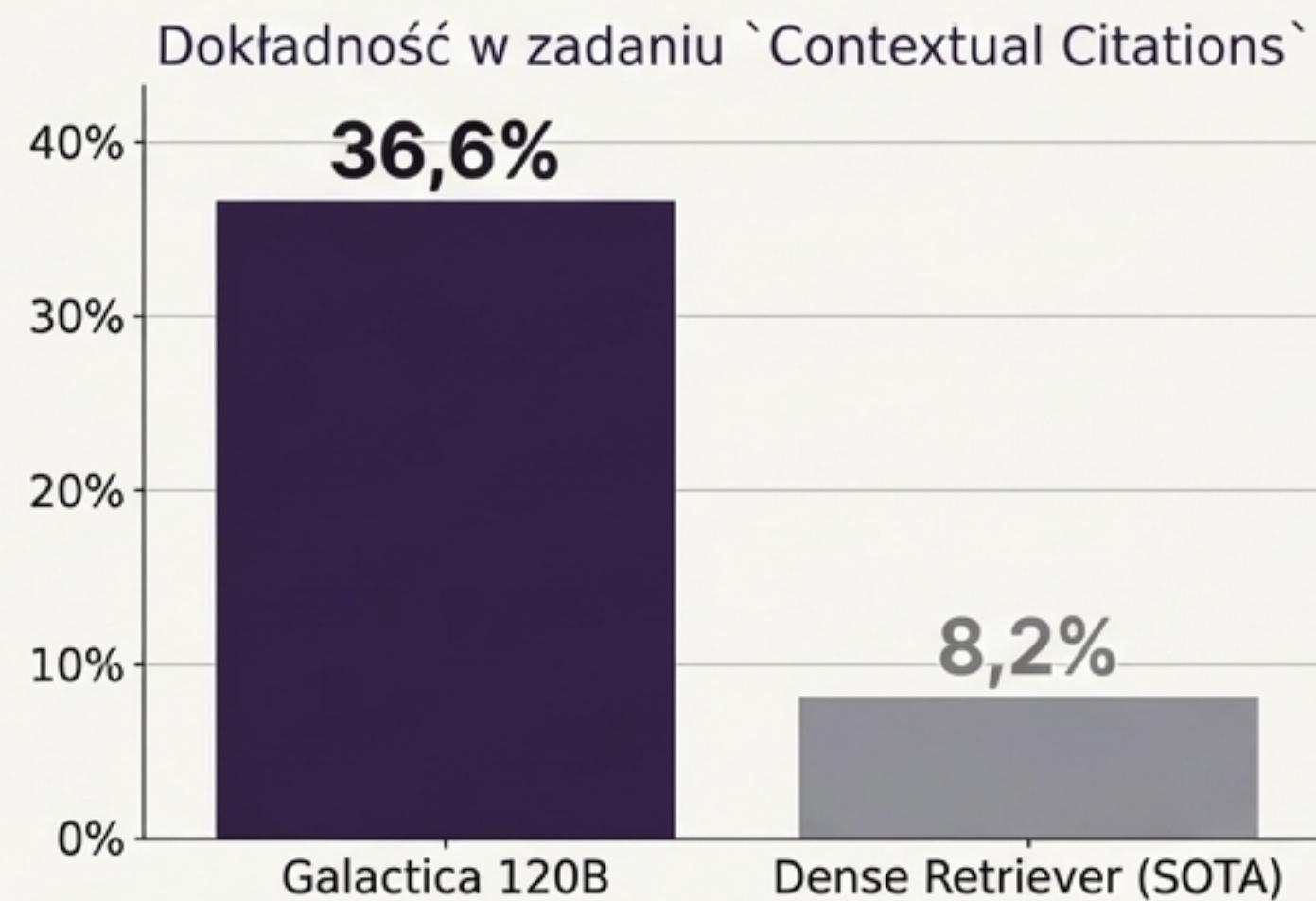


# Tytuł: Przewidywanie Cytowań – Nowy Interfejs do Literatury

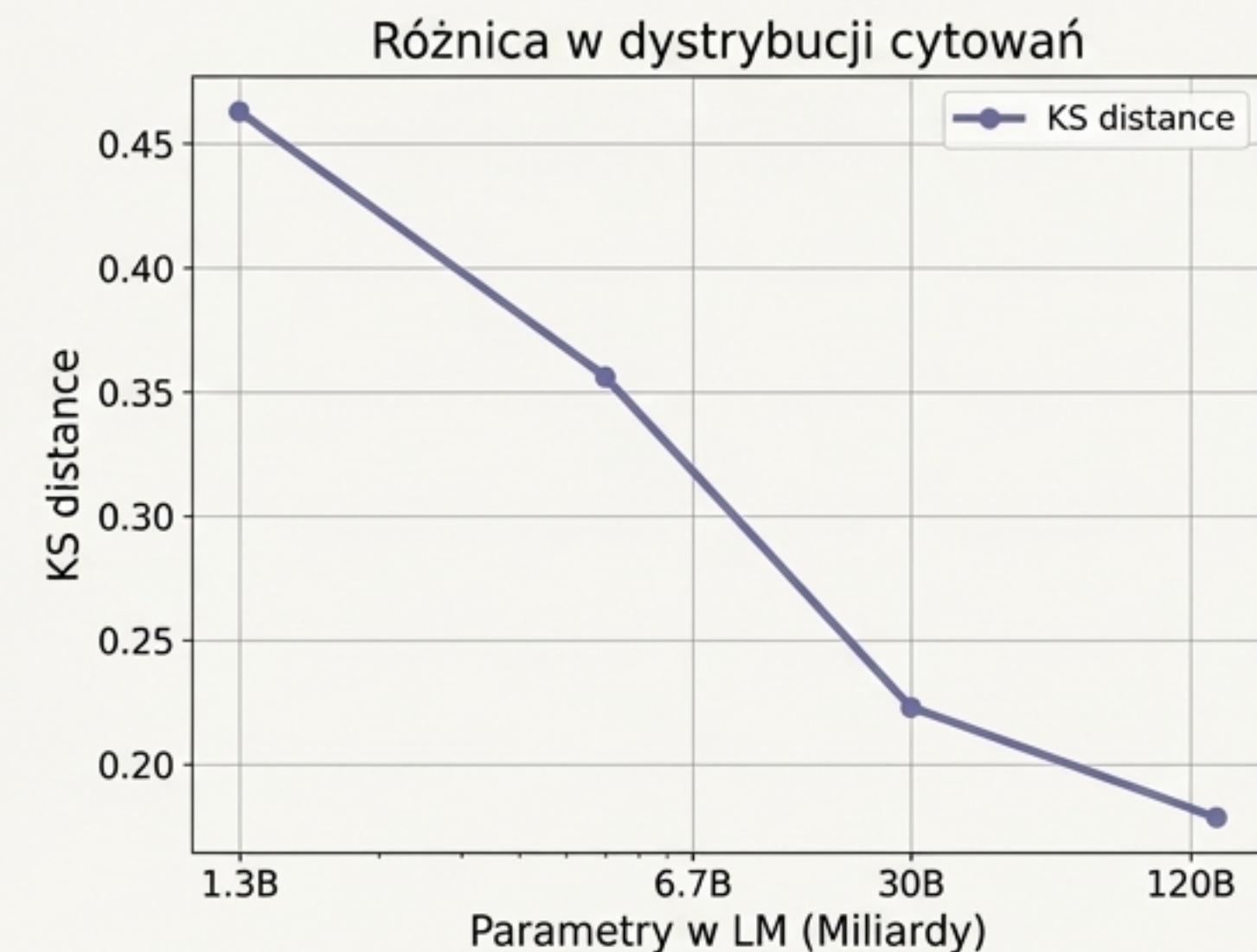
## Podtytuł: 4x Lepszy niż Dedykowane Narzędzia Wyszukiwawcze

**Zadanie:** Na podstawie fragmentu tekstu naukowego, przewidzieć, która praca powinna być w tym miejscu zacytowana.

**Implikacje:** Model językowy jest **ponad 4 razy skuteczniejszy** w kontekstowym rozumieniu potrzeby cytowania niż narzędzie zaprojektowane specjalnie do tego celu.



**Jakość ponad Popularność:** Wraz ze wzrostem modelu, maleje jego skłonność do cytowania tylko najpopularniejszych prac. Zamiast tego, sugeruje bardziej niszowe, ale trafniejsze pozycje.



# Tytuł: Odkrycie #1: Powtarzanie Danych Działa

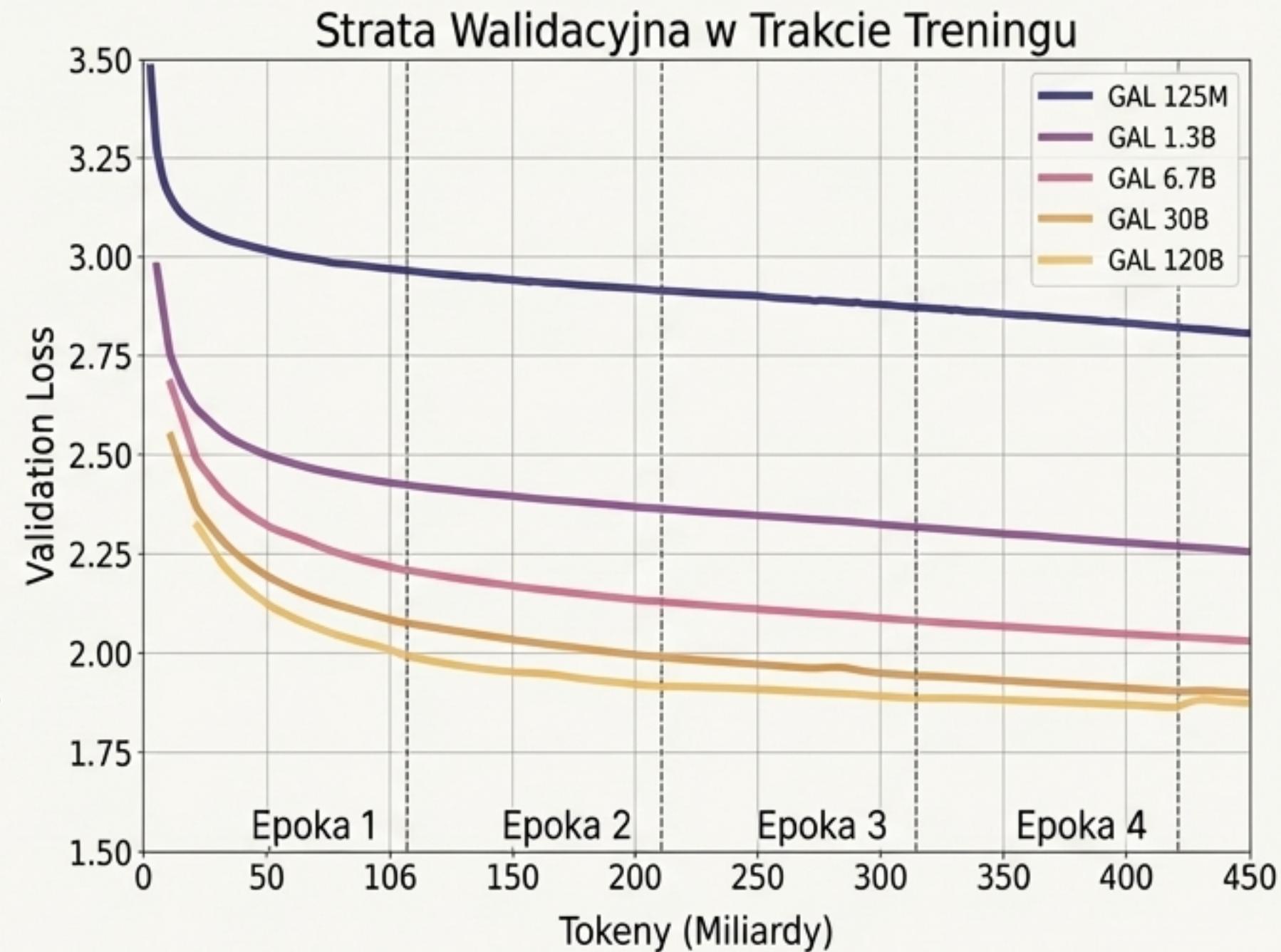
## Podtytuł: Podważenie Dogmatu o Prawach Skalowania

**Dogmat (Prawa Skalowania Chinchilla):** Panowało przekonanie, że model nigdy nie powinien być trenowany na powtórzonych danych. Każdy token miał być unikalny, aby uniknąć przeuczenia (overfittingu) i zapamiętywania zamiast generalizacji.

**Eksperyment Galactiki:** Model był trenowany przez **ponad 4 epoki**, co oznacza, że każdy fragment tekstu widział wielokrotnie.

**Zaskakujący Rezultat:** Zaobserwowano **ciągłą poprawę wyników** – spadek straty walidacyjnej trwał przez 4 epoki. Co więcej, wydajność rosła nie tylko na zadaniach z domeną, ale także na zadaniach ogólnych, spoza domeny (out-of-domain).

**Nowa Hipoteza:** Prawa skalowania Chinchilla prawdopodobnie dotyczą **zaszumionych danych z internetu**. Gęste, wysokiej jakości korpusy naukowe zyskują na głębszym przetwarzaniu poprzez wielokrotne iteracje.



# Tytuł: Odkrycie #2: Wąski Trening, Szeroka Poprawa

## Podtytuł: Naukowe Fundamenty Wzmacniają Ogólne Zdolności

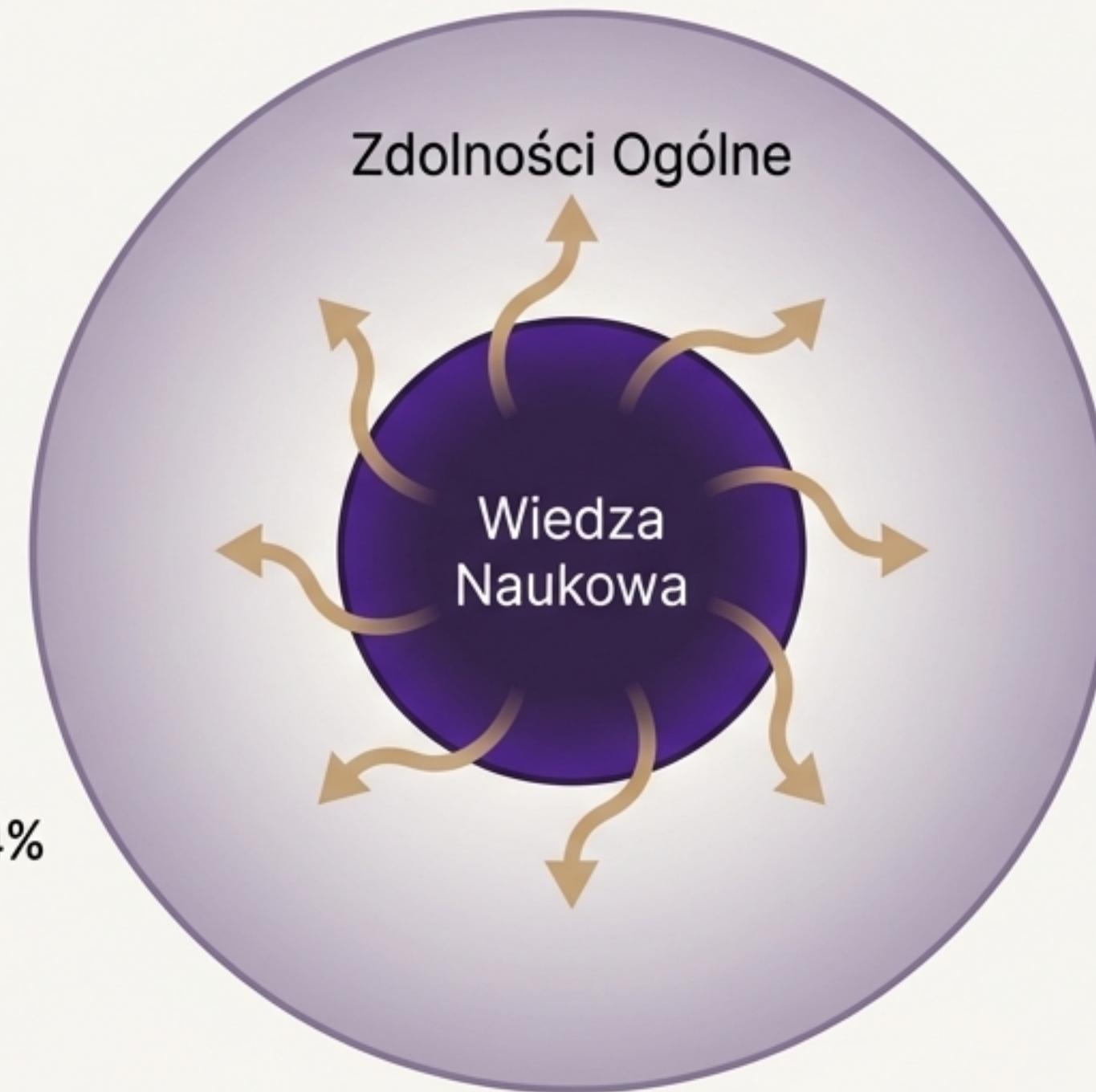
### Nieintuicyjne Przeniesienie

#### Wiedzy:

Trening wyłącznie na wąskim, naukowym korpusie poprawił wyniki na ogólnych (nienaukowych) zadaniach.

### Benchmark BIG-bench (głównie zadania nienaukowe):

- **Galactica 120B: 48,7%**
- OPT 175B (model ogólny): 43,4%
- BLOOM 176B (model ogólny): 42,6%



#### Wniosek:

Wiedza naukowa buduje silniejsze fundamenty do rozumowania niż przypadkowe teksty z internetu. To jak nauka łaciny i greki, która poprawia zrozumienie współczesnych języków.

### Wyzwanie dla Filozofii „Odkurzania Internetu”:

Wynik ten kwestionuje podejście, w którym jedynym sposobem na budowę potężnego modelu jest przetworzenie jak największej części publicznej sieci.

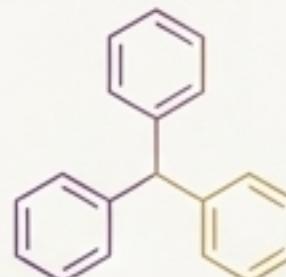
# Tytuł: Realne Zastosowania: Chemia i Biologia

## Podtytuł: Od Nazewnictwa po Opis Funkcji Białek

### Chemia

#### 1. Nazewnictwo IUPAC

Ucząc się w trybie samonadzorowanym, model osiągnął **39,2%** dokładności w generowaniu złożonych, systematycznych nazw chemicznych – zadania trudnego do automatyzacji.



1,1,1-Triphenylethane

[START\_SMILES]C1=CC=C(C=C1)C(C2=CC=CC=C2)C3=CC=CC=C3[END\_SMILES]

#### 2. Odkrywanie Leków

Zadania klasyfikacyjne z Zadania klasyfikacyjne z MoleculeNet (np. czy związek przeniknie barierę krew-mózg) można formułować jako prompt w języku naturalnym, łącząc sekwencje SMILES i tekst.

[START\_SMILES]CN1CCN(CC1)C(=O)c2ccc(cc2)S(=O)(=O)N3CCCCC3...[END\_SMILES]

Will it penetrate the blood-brain barrier?



### Biologia

#### 1. Opis Funkcji Białek

Na podstawie samej sekwencji aminokwasów, Galactica jest w stanie generować trafne, swobodne opisy funkcji białka, osiągając wysoką zgodność z bazami danych jak UniProt.

[START\_AMINO]MQKSPLERAS  
GANVPLAFFKGNDLILHTMVSBLK  
LGOALNAGRGVNDEIVEDGKRDO  
SPRCERRIVPBSJRKYFRFSTLRG  
BKARNLDVRMSVRDRADNGE...  
[END\_AMINO]

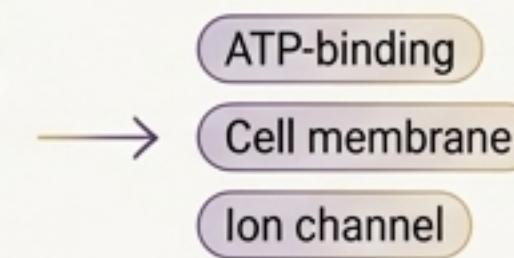


Component of the ubiquinol-cytochrome c reductase complex, which is part of the mitochondrial respiratory chain. Transfers electrons from ubiquinol to cytochrome c.

#### 2. Przewidywanie Słów Kluczowych

Z samej sekwencji model potrafi z dużą dokładnością (F1 do 54.5% dla 120B) przewidzieć słowa kluczowe opisujące funkcję i lokalizację białka.

[START\_AMINO]...MFRNKLVFLLBLV  
RNBLFFJDOKMQVA.YLRLGNVFINFR  
KDRDVLTKCJVRWWLNEGAD,CWNL  
CVDFEJKRCWNRFIVINEVYRVAM...  
[END\_AMINO]

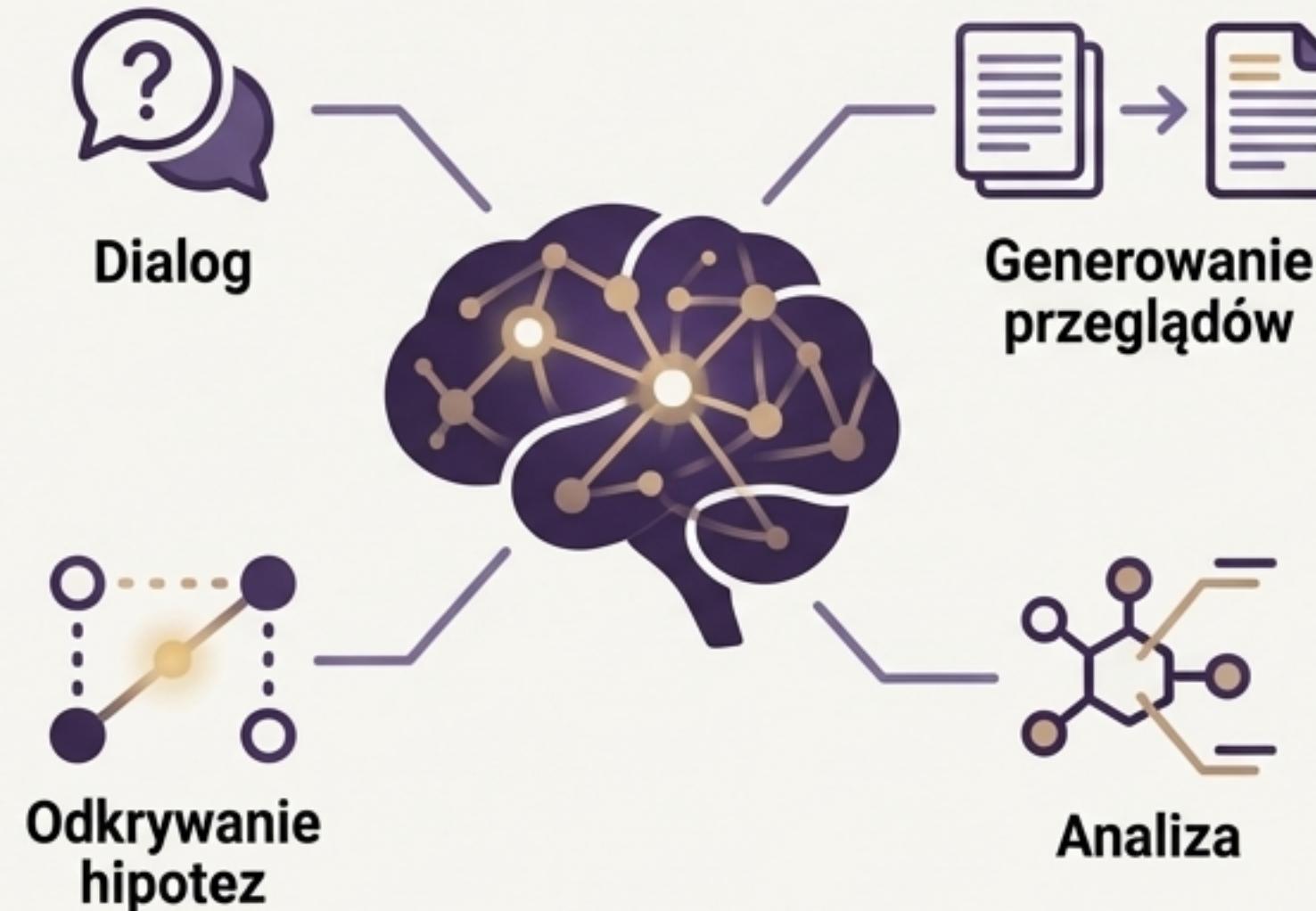


# Tytuł: Nowy Interfejs dla Nauki

Podtytuł: Od Wyszukiwania do Dialogu z Wiedzą

## Podsumowanie Mocy Galactiki:

- **Przechowuje:** Kuratorowany korpus naukowy w wagach sieci.
- **Łączy:** Buduje mosty między tekstem, wzorami, chemią i biologią.
- **Rozumie:** Rozwiązuje problemy krok po kroku, przewyższając przewyższając większe modele.



## Wizja Przyszłości

**Przelamanie Paradygmatu:** Dominujący od ponad 50 lat model „przechowaj i wyszukaj” jest niewydolny. Przetwarzanie i synteza wiedzy wciąż zależą od ludzkiego wysiłku, tworząc wąskie gardło.

**Wizja Przyszłości:** Modele takie jak Galactica to początek nowego paradygmatu – bezpośredniej interakcji z globalną wiedzą naukową.

## Kolejny Krok:

Udostępniamy modele jako open source, aby społeczność naukowa mogła na nich budować przyszłość badań.