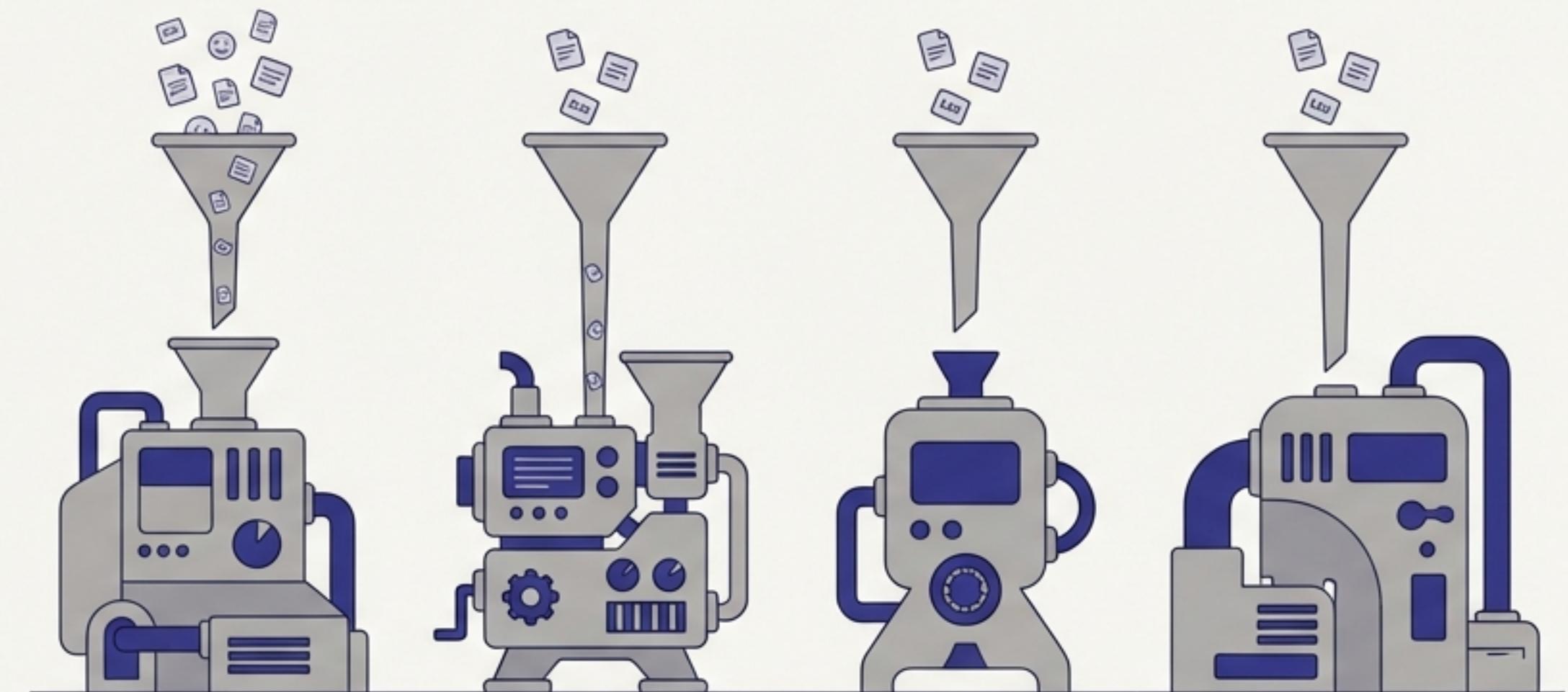


Krajobraz NLP przed 2018: Świat Rzemieślniczych Modeli

Przed GPT, każda dziedzina NLP była odizolowaną wyspą, wymagającą budowy modeli od zera.

- **"Fragmentacja Zadań"**: Każde zadanie (analiza sentymetu, Q&A) wymagało unikalnej, specjalnie zaprojektowanej architektury.
- **"Głód Danych"**: Modele wymagały "substantial amounts of manually labeled data", co było największym ograniczeniem.
- **"Ograniczony Transfer Wiedzy"**: Uczenie transferowe ograniczało się głównie do embeddingów słów – "transfer word-level information".
- **"Problem Fundamentalny"**: Brakowało uniwersalnej metody efektywnej nauki reprezentacji języka z danych nieoznaczonych.



Wąskie gardło: drogie, ręcznie oznaczane zbiory danych

Rewolucyjne Pytanie: Co Gdyby Istniał Jeden Uniwersalny Model?

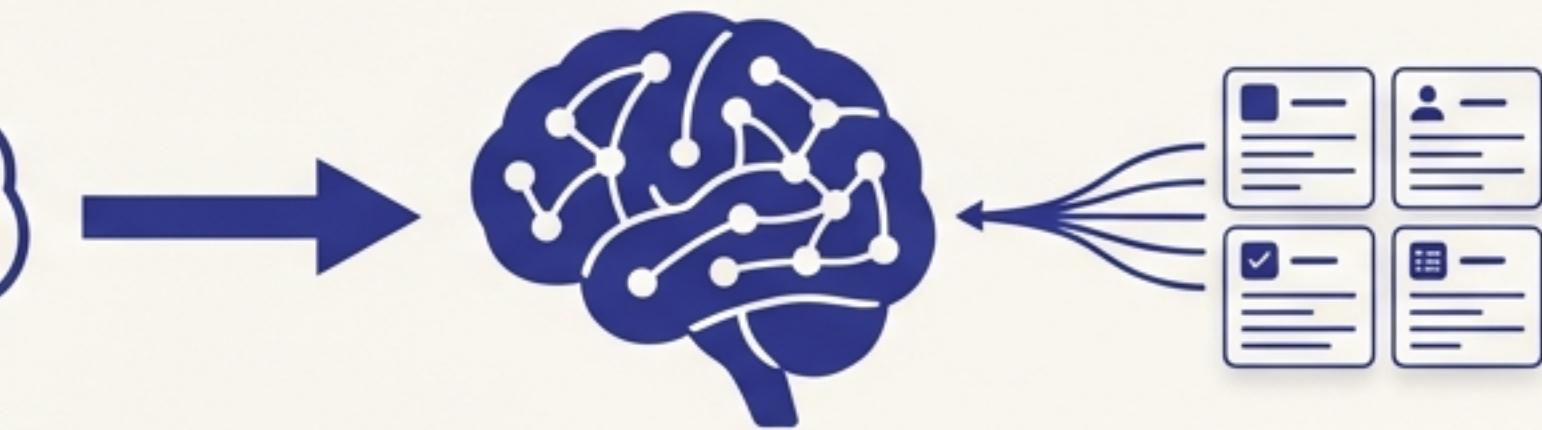
Zamiast tworzyć wiele wyspecjalizowanych narzędzi, stwórzmy jeden, potężny rdzeń, który rozumie język uniwersalnie.

1. Generatywny Pre-trening (Nienadzorowany)



Uczenie się "uniwersalnej reprezentacji" na
ogromnym korpusie nieoznaczonego tekstu.

2. Dyskryminacyjne Dostrajanie (Nadzorowany)



Adaptacja do konkretnego zadania przy
użyciu małego zbioru oznaczonych danych.

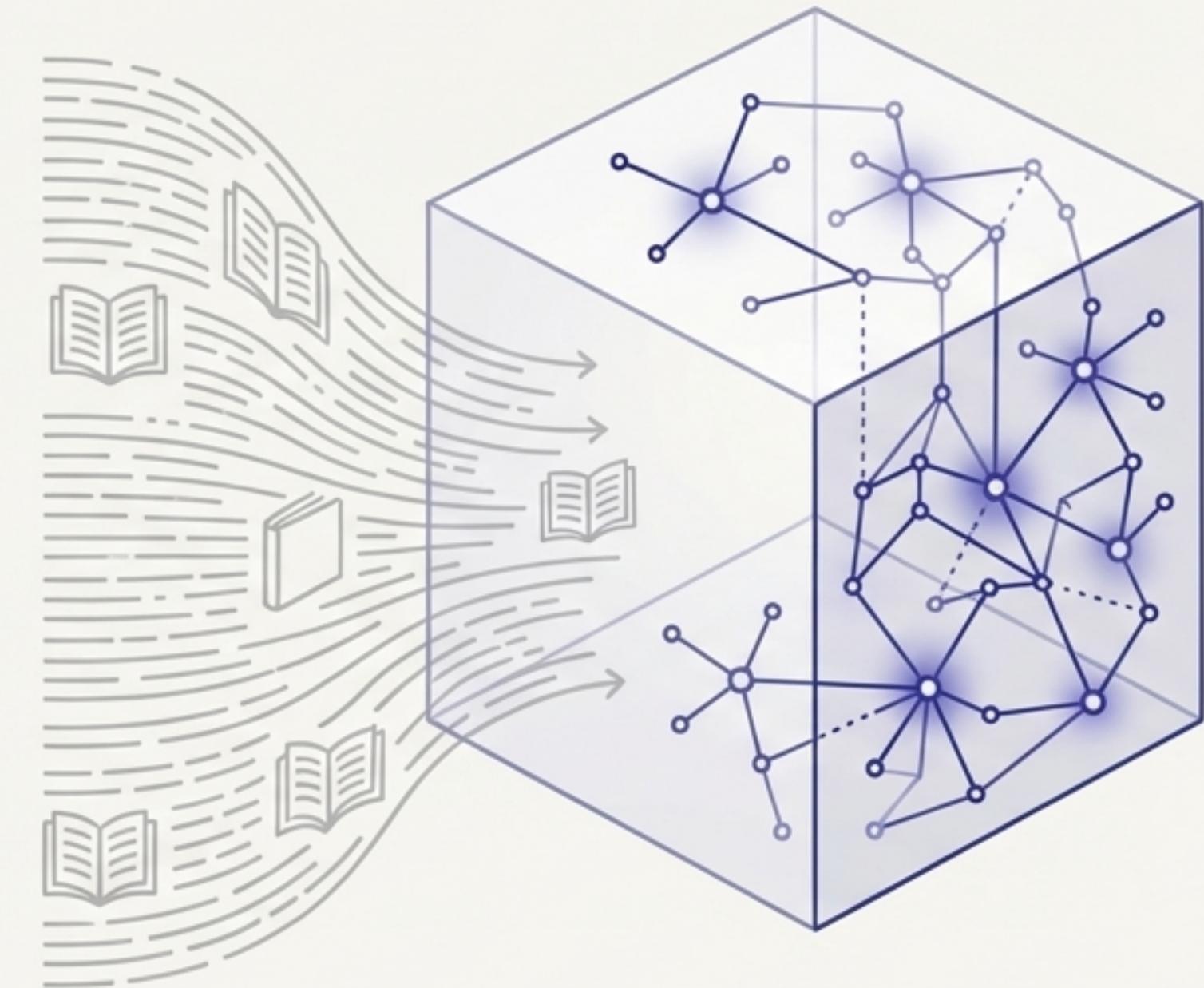
Zmiana Filozofii

Fundamentalne odejście od modeli specyficznych dla zadania ("task-specific") na rzecz
jednego, agnostycznego modelu ("task-agnostic").

Etap 1: Generatywny Pre-trening – Nauka Języka przez Immersję

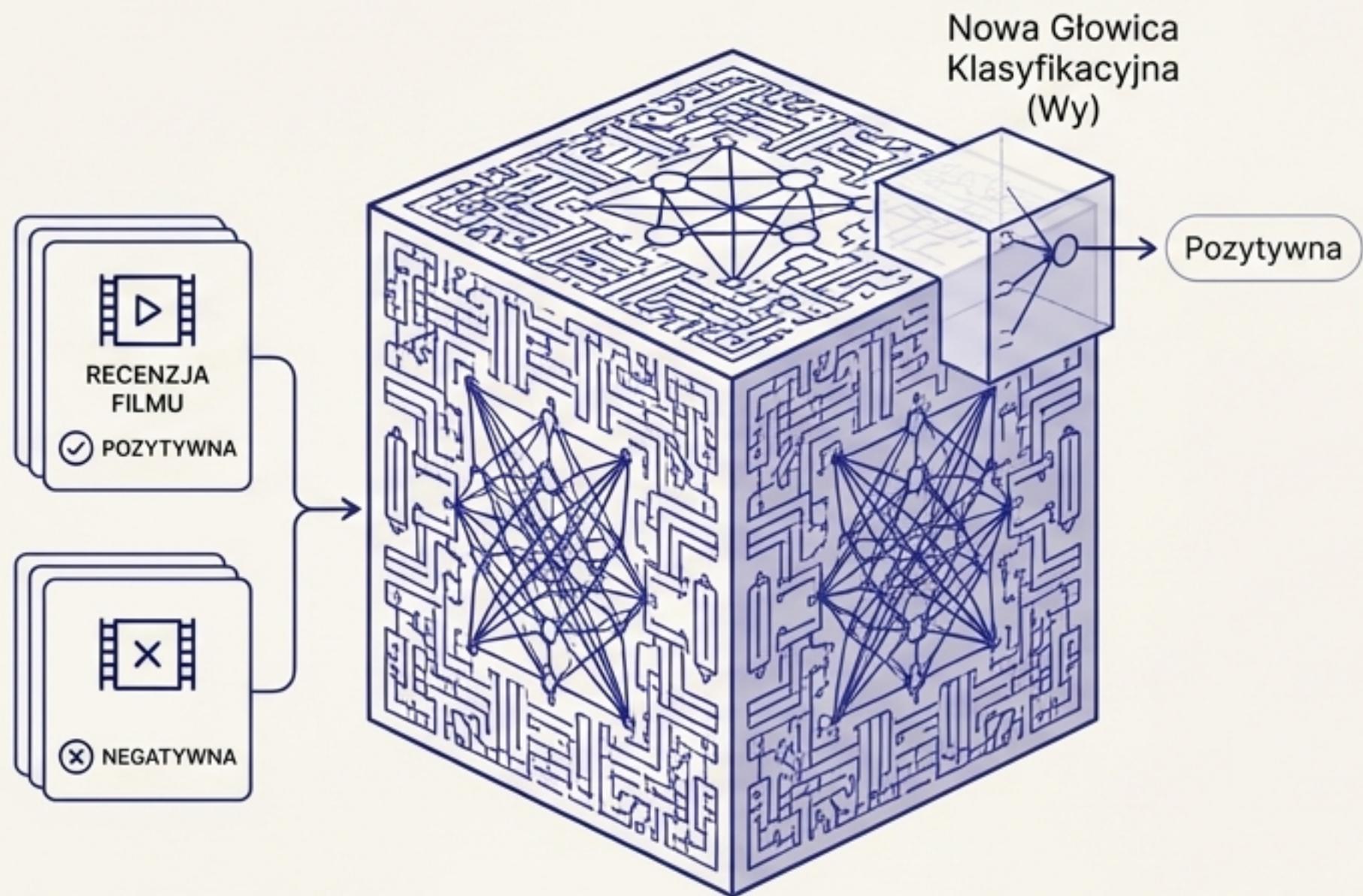
Model, aby przewidzieć następne słowo, musi ***samodzielnie* nauczyć się gramatyki, faktów i logiki.**

- **Cel Treningu:** Maksymalizacja prawdopodobieństwa następnego słowa w sekwencji ('standard language modeling objective').
- **Zbiór Danych:** BooksCorpus – ponad 7000 unikalnych książek, kluczowe dla nauki zależności dalekiego zasięgu dzięki "long stretches of contiguous text".
- **Wiedza Emergentna:** Zdolności (gramatyka, wiedza o świecie) wyłoniły się jako produkt uboczny, nie były celem samym w sobie.
- **Wynik:** 'Uniwersalna reprezentacja' języka, która osiągnęła bardzo niski perplexity (18.4).



Etap 2: Dyskryminacyjne Dostrajanie – Od Absolwenta do Specjalisty

Dostrajanie nie uczy od zera, a jedynie adaptuje istniejącą, głęboką wiedzę językową do nowego zadania.



- **Proces:** Aktywacje z ostatniej warstwy pre-trenowanego modelu (`hml``) są podawane na nową, prostą warstwę liniową (`'Wy'`) przewidującą etykietę.
- **Minimalne Zmiany:** Dodawane są tylko wagi dla warstwy klasyfikacyjnej i tokenów-separatorów.
- **Pomocniczy Cel Treningowy:** Dodanie celu modelowania języka ($L3 = L2 + \lambda * L1$) poprawia generalizację i przyspiesza konwergencję.
- **Efektywność:** Szybkie dostrajanie, często wystarczają 3 epoki treningu.

Serce Modelu: Dlaczego Transformer, a nie LSTM?

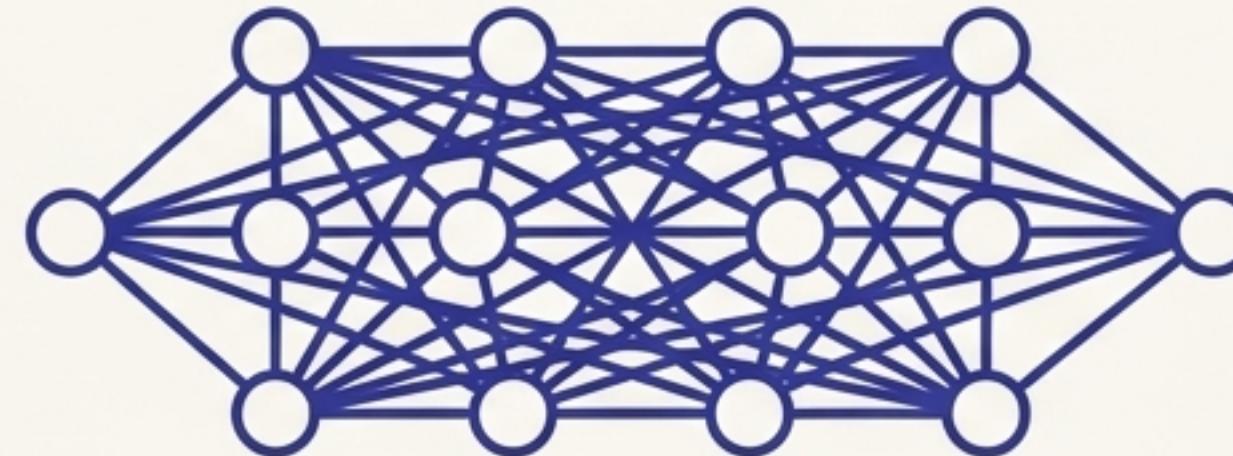
Architektura Transformer posiada 'pamięć strukturalną' niezbędną do przetwarzania zależności dalekiego zasięgu, co było kluczowe dla czytania całych książek.

LSTM: Pamięć Liniowa, Krótkiego Zasięgu



Gubi kontekst w długich tekstach.
‘prediction ability restricted to a short range’.

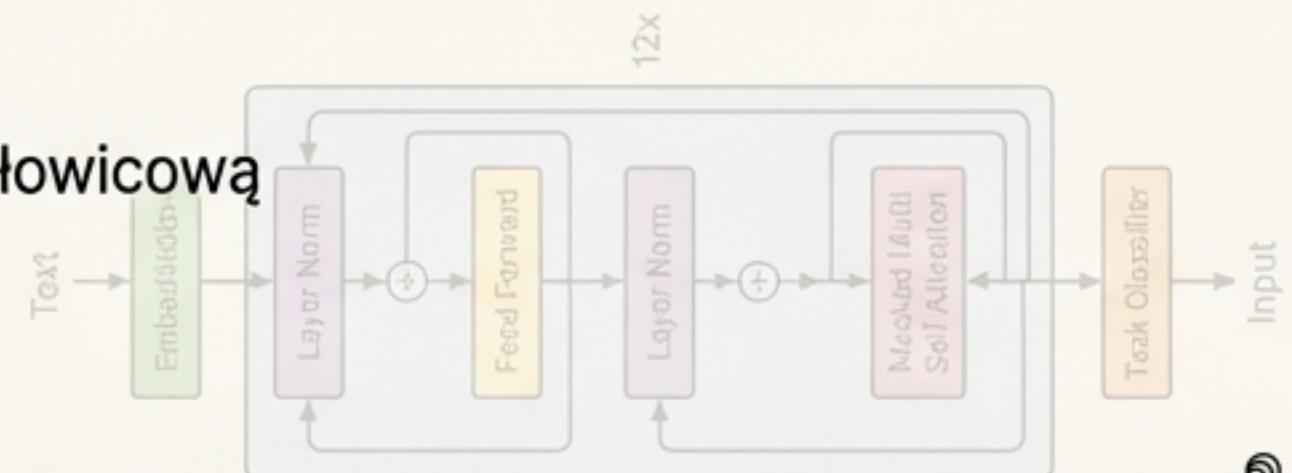
Transformer: Pamięć Globalna, Długiego Zasięgu



Mechanizm uwagi (self-attention) waży znaczenie wszystkich słów, niezależnie od odległości. ‘more structured memory for handling long-term dependencies’.

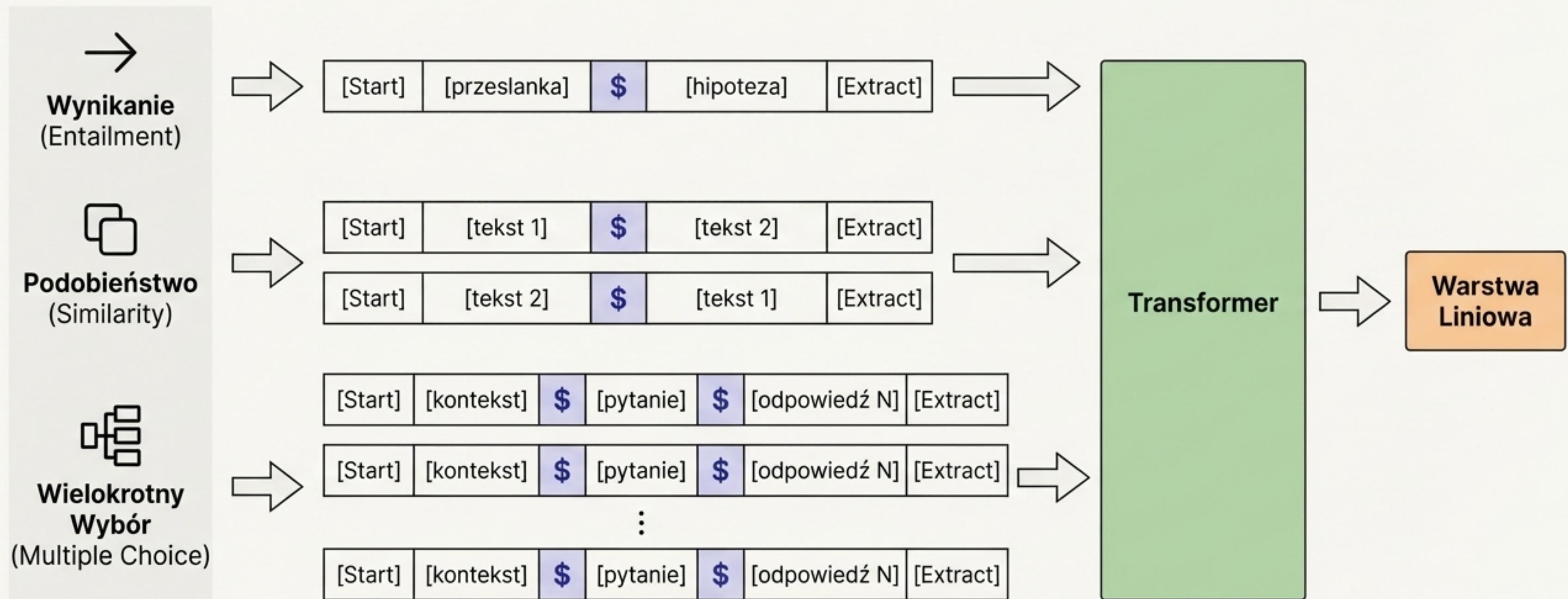
Specyfikacja Architektury GPT-1

- 12-warstwowy dekoder Transformera z zamaskowaną uwagą wielogłowicową
- Stany ukryte: 768 wymiarów
- Głowice uwagi: 12
- Warstwy feed-forward: 3072 wymiary



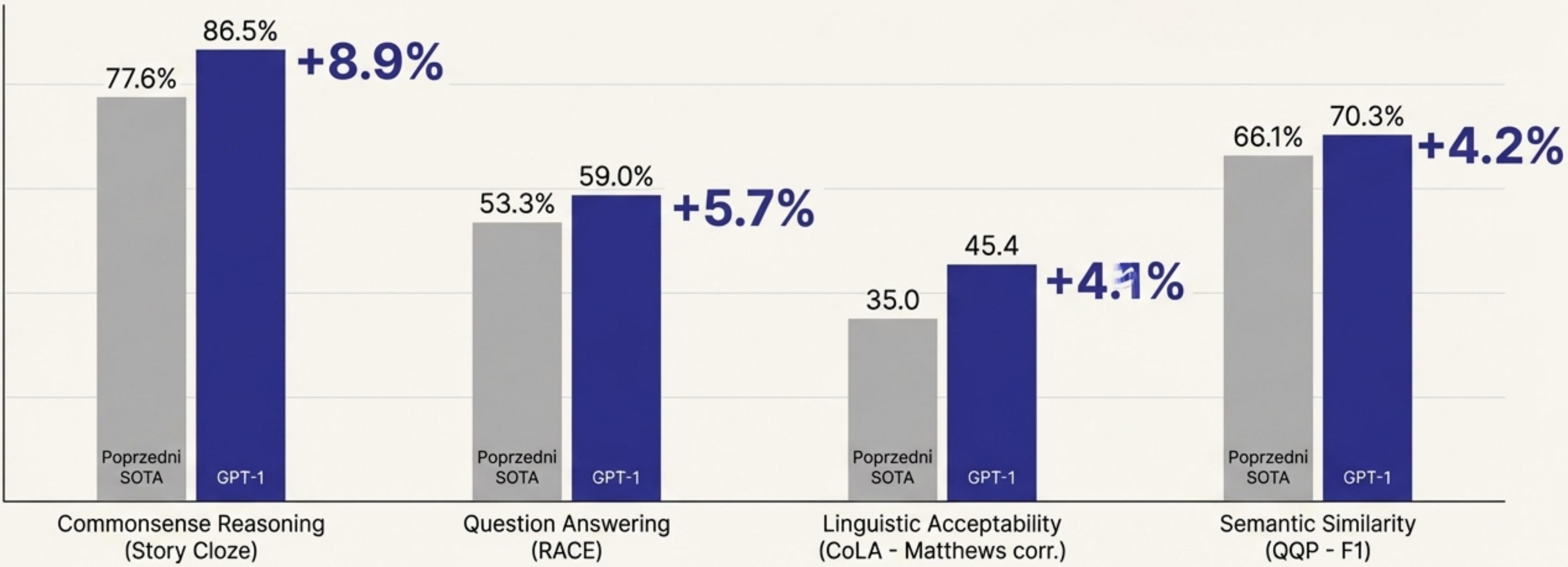
Genialna Sztuczka: Zamiast Zmieniać Model, Zmień Format Danych

Wszystkie zróżnicowane zadania NLP zostały "przetłumaczone" na prosty format sekwencji, eliminując potrzebę tworzenia specyficznych architektur.



Wyniki: Pobito Modele 'Szyte na Miarę' przez Lata

GPT-1, jako model agnostyczny, zdeklasował wyspecjalizowane modele na 9 z 12 zadań, ustanawiając nowy stan najnowszej techniki (SOTA).



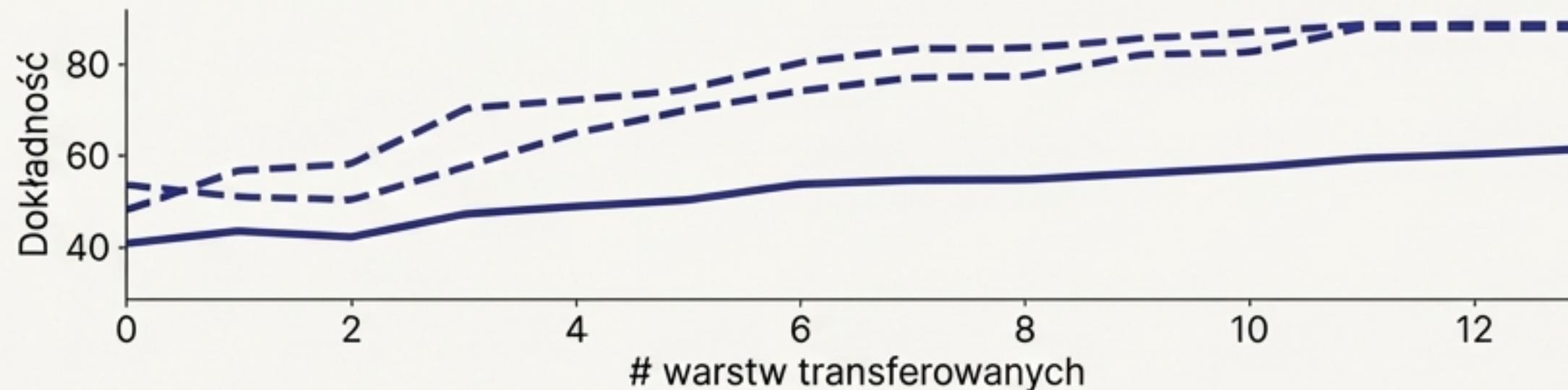
Sukces na tak zróżnicowanych zadaniach dowódł, że podejście `pre-train + fine-tune` jest niezwykle potężne i uniwersalne.

Co Kryje się w Środku? Transfer Wiedzy i Zdolności 'Zero-Shot'

Wiedza jest rozproszona we wszystkich warstwach modelu, a pre-trening prowadzi do powstawania nieoczekiwanych, emergentnych zdolności.

Analiza Transferu Warstw

Każda z 12 warstw wnosi unikalną wiedzę



Obalono tezę, że wiedza jest skumulowana tylko w niższych warstwach.
‘Each transformer layer provides further benefits up to 9%’.

Zdolności 'Zero-Shot'

Produkt uboczny pre-treningu: model nauczył się rzeczy, których go nie uczono

Model z wyższym prawdopodobieństwem wybiera "positive", klasyfikując sentyment bez dostrajania.

Ten film był ___. →

very

positive

negative

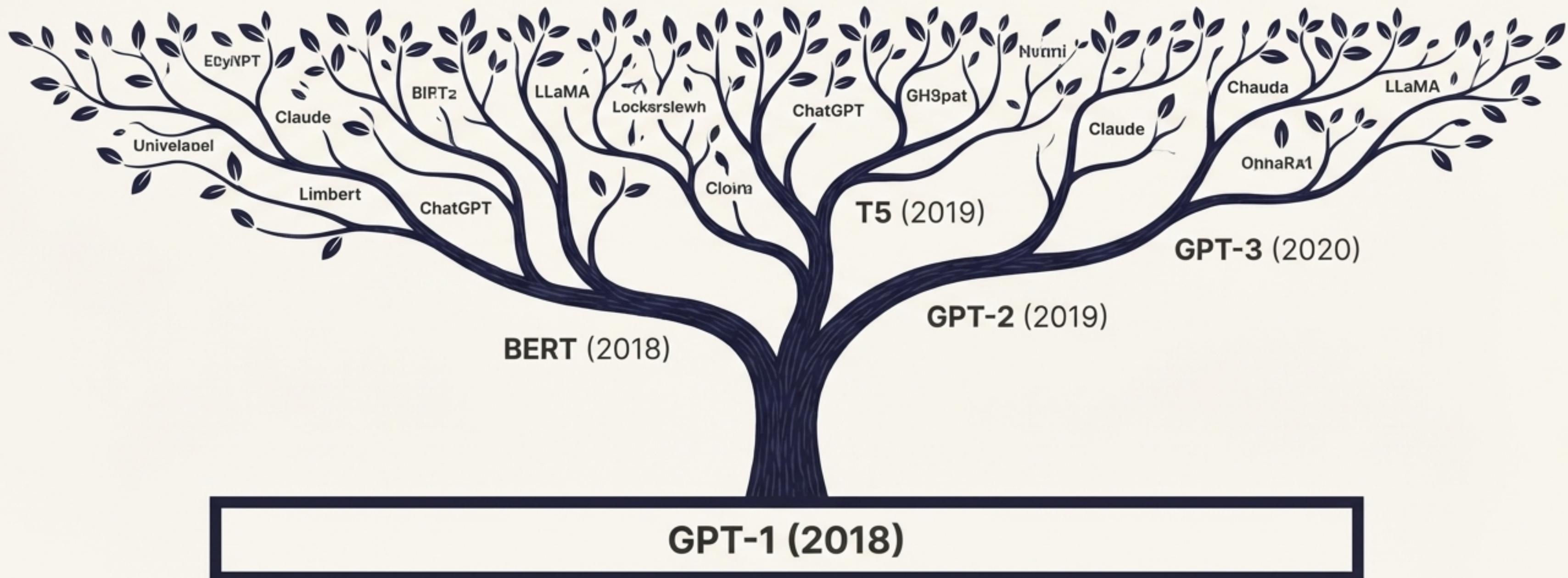
Dowód przez Eliminację: Co Naprawdę Miało Znaczenie?

Badania ablacyjne jednoznacznie wykazały, że zarówno pre-trening, jak i architektura Transformera były niezbędnymi składnikami sukcesu.



Dziedzictwo GPT-1: Fundament Nowej Ery w AI

GPT-1 był dowodem koncepcji, który ustanowił paradygmat "Pre-train, Fine-tune" jako dominującą siłę w NLP i utorował drogę dla całej rewolucji Generatywnej AI.



"Our work suggests that achieving significant performance gains is indeed possible, and offers hints as to what models (Transformers) and data sets (text with long range dependencies) work best with this approach." – Radford et al., 2018