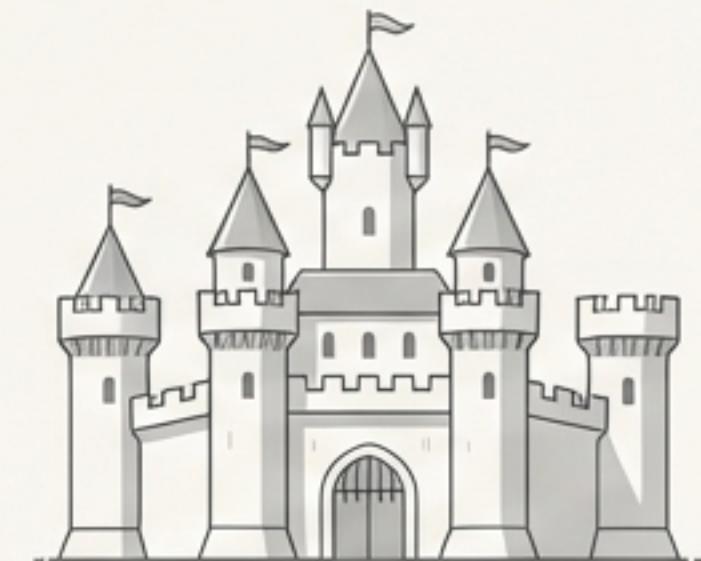


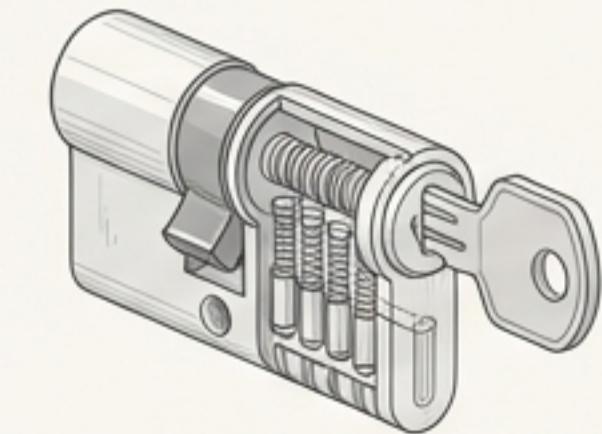
Tytuł: Przed BERT: Maszyny znały słowa, ale nie rozumiały kontekstu

Przez dekady komputery miały trudności ze zrozumieniem niuansów językowych, takich jak polisemantyczność, gdzie to samo słowo ma różne znaczenia w zależności od otoczenia. Modele NLP potrafiły przetwarzanie słowa, ale nie były w stanie uchwycić głębokiego, kontekstowego znaczenia, kluczowego dla zadań takich jak odpowiadanie na pytania czy wnioskowanie.

ZAMEK



"Król mieszka w **zamku**."



"Nie mam **zamku** w drzwiach."

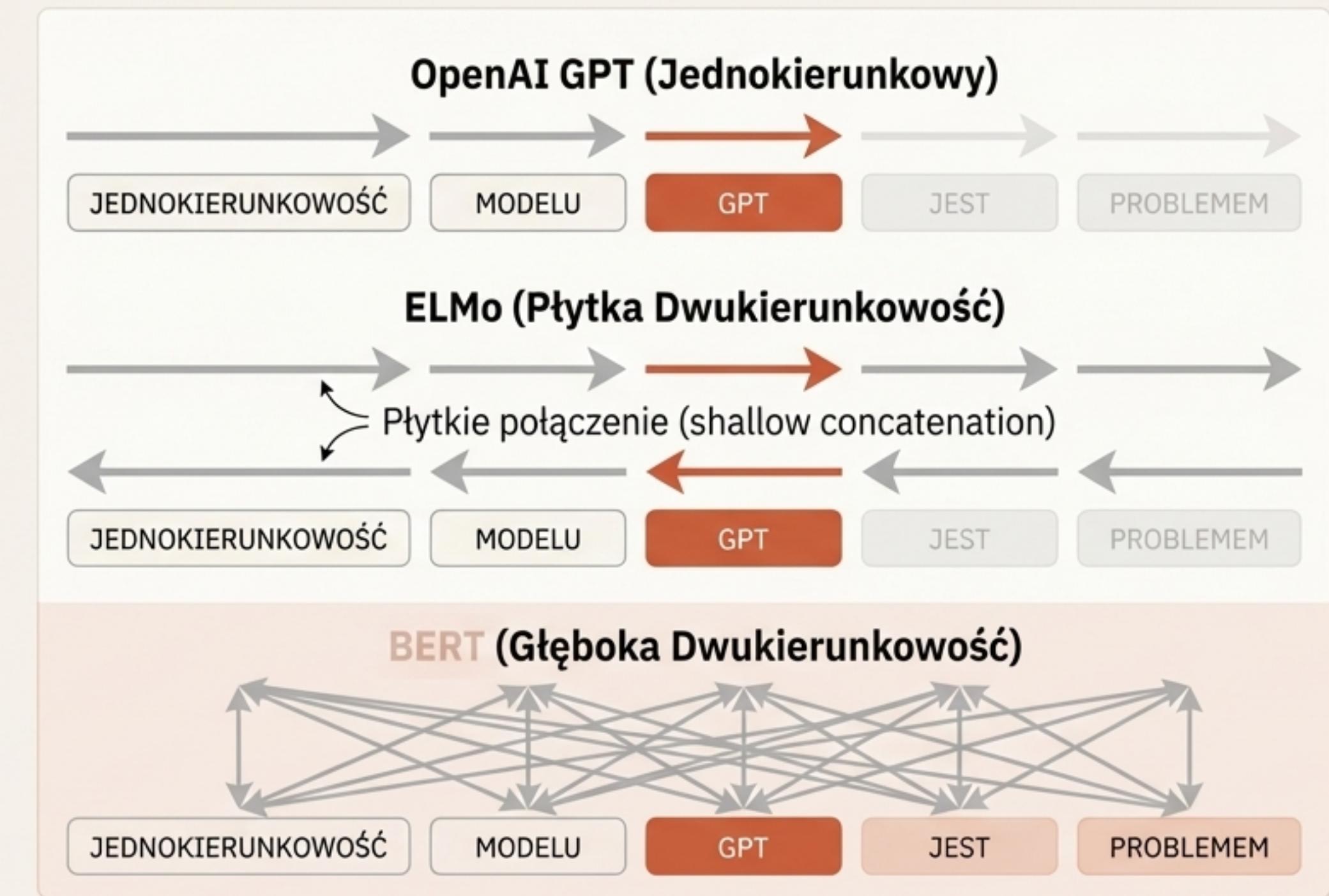
Ograniczenie poprzedników: Jednokierunkowość uniemożliwiała pełne spojrzenie na tekst

Wiodące modele przed BERT, takie jak OpenAI GPT i ELMo, przetwarzają tekst **jednokierunkowo**.

OpenAI GPT widział tylko tokeny po lewej stronie (przesłość).

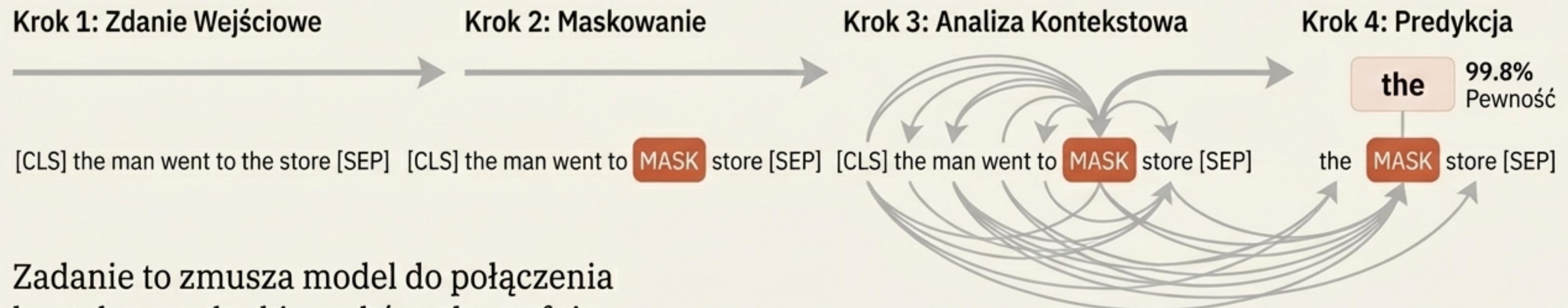
ELMo stosowało „głębokie połączenie” dwóch niezależnie trenowanych modeli (L->P i P->L).

Brakowało **głęboko dwukierunkowej reprezentacji**, gdzie model na każdej warstwie ma jednoczesny dostęp do kontekstu z obu stron.



Innowacja nr 1: Masked Language Model (MLM) wymusił głęboką dwukierunkowość

Aby wytrenować głęboko dwukierunkowy model, BERT wprowadził nowatorskie zadanie pre-treningowe inspirowane testem typu „Cloze”. Zamiast przewidywać następne słowo, model musiał uzupełnić luki w tekście, wykorzystując pełen kontekst.



Zadanie to zmusza model do połączenia kontekstu z obu kierunków, aby trafnie uzupełnić luke. To jest esencja **głębokiej głębokiej dwukierunkowości**.

Strategia maskowania 80/10/10: Klucz do uniknięcia rozbieżności między treningiem a zastosowaniem

Token `[MASK]` pojawia się tylko podczas pre-treningu. Aby zminimalizować tę rozbieżność, dla 15% losowo wybranych tokenów zastosowano wyrafinowaną strategię:

80%

my dog is hairy



my dog is **[MASK]**

Token jest zastępowany przez [MASK]. Uczy to model przewidywania słów na podstawie kontekstu.

10%

my dog is hairy



my dog is **apple**

Token jest zastępowany **losowym** słowem. Uczy to model, że nie każde słowo pasuje, budując odporność.

10%

my dog is hairy



my dog is hairy

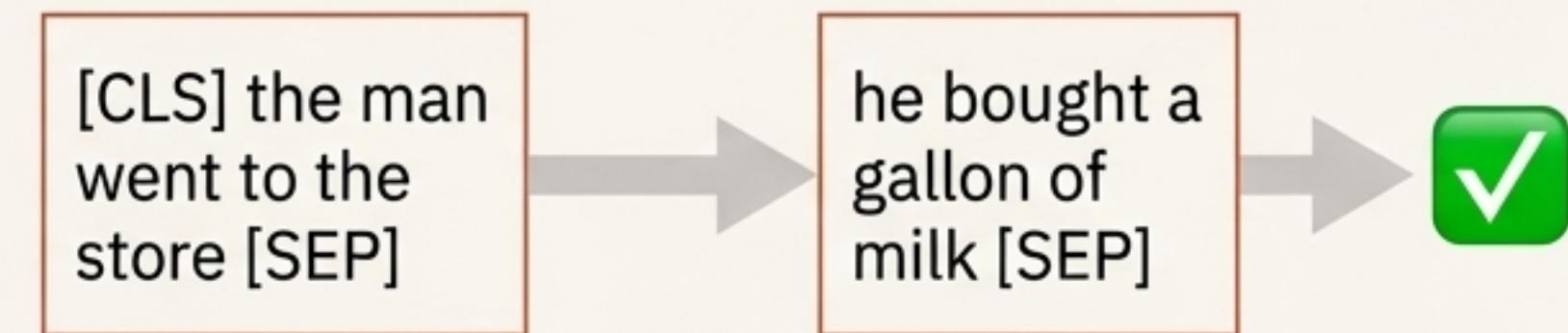
Token pozostaje **niezmieniony**. Uczy to model oceny poprawności danego słowa w oryginalnym kontekście.

Ta procedura zmusza model do utrzymywania dystrybucyjnej, kontekstowej reprezentacji każdego tokena, ponieważ nigdy nie jest pewien, czy dany token jest prawdziwy.

Innowacja nr 2: Next Sentence Prediction (NSP) nauczyło model logiki narracji

- **Problem:** Masked Language Model uczy rozumienia relacji wewnętrz zdań, ale nie między nimi.
- **Zadanie NSP:** Aby nauczyć model rozumienia spójności tekstu, wprowadzono drugie zadanie pre-treningowe: przewidywanie następnego zdania.
- **Proces:** Model otrzymuje parę zdań (A i B) i przewiduje, czy zdanie B jest faktycznym zdaniem następującym po A.
- **Znaczenie:** Kluczowe dla zadań wymagających rozumowania na poziomie wymagających rozumowania na poziomie wielu zdań, jak Question Answering (QA) i Natural Language Inference (NLI).

Przypadek 1: `IsNext` (50% danych)



Przypadek 2: `NotNext` (50% danych)



Rezultat: BERT ustanowił nowe rekordy State-of-the-Art w 11 zadaniach NLP

GLUE Benchmark

80.5%

Absolutna poprawa o +7.7 pkt. proc.
w stosunku do poprzedniego SOTA.

SQuAD 1.1 (QA)

93.2 F1

Jako pierwszy system przewyższył ludzką
wydajność (Human: 91.2 F1).

SQuAD 2.0 (QA with No Answers)

83.1 F1

Poprawa o +5.1 pkt. F1 w stosunku do
najlepszego opublikowanego systemu.

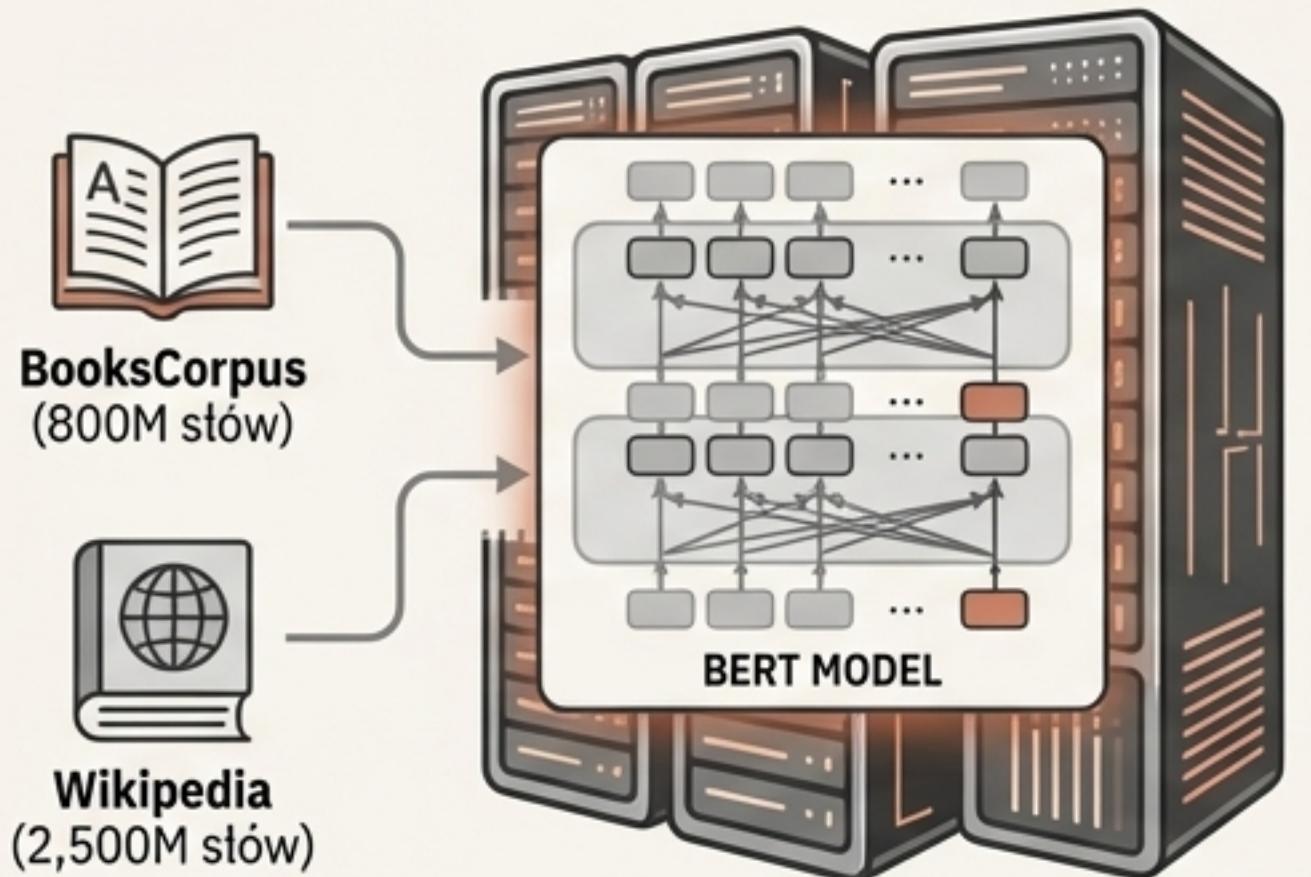
MultiNLI (Inference)

86.7%

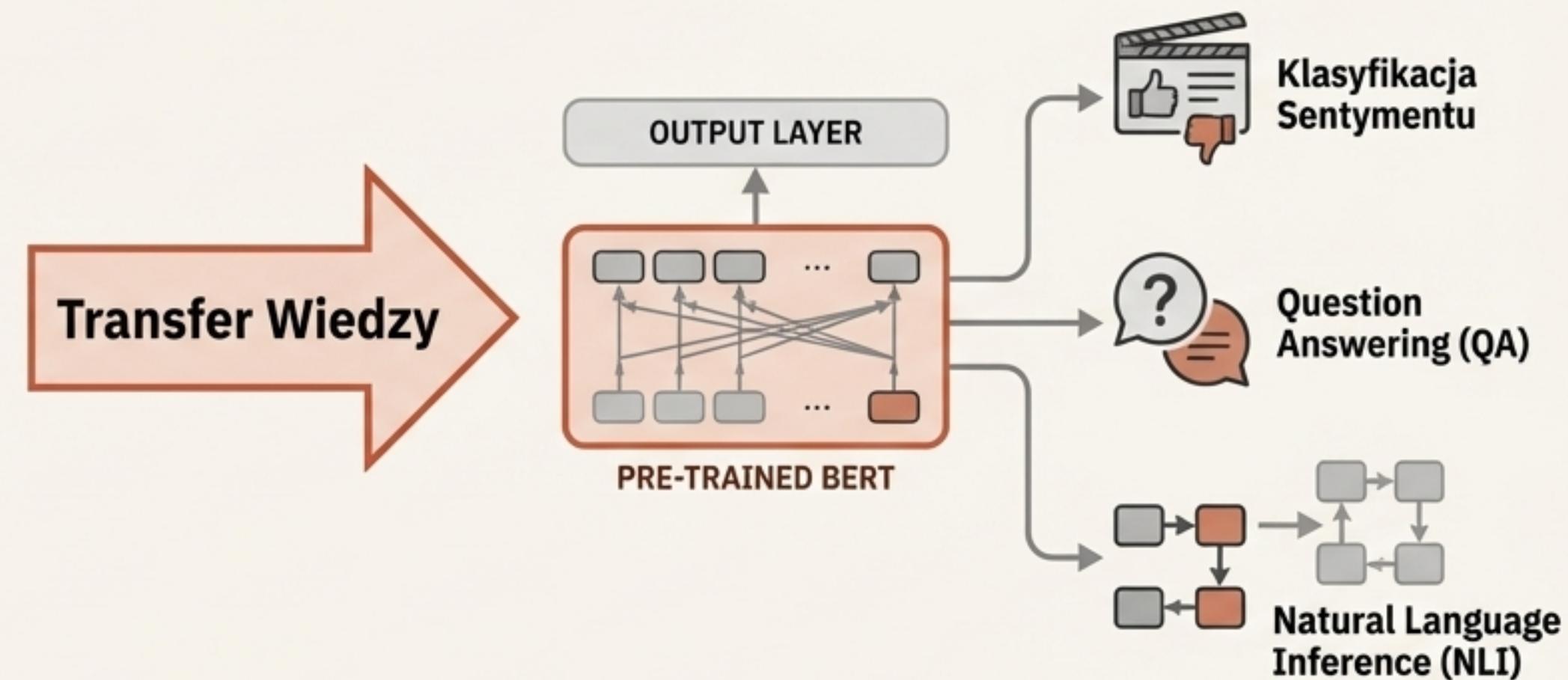
Poprawa dokładności o +4.6 pkt. proc.

Nowy paradymat: Trenuj raz na wielką skalę, dostrajaj tanio do konkretnych zadań

Etap 1: Pre-trening



Etap 2: Fine-tuning



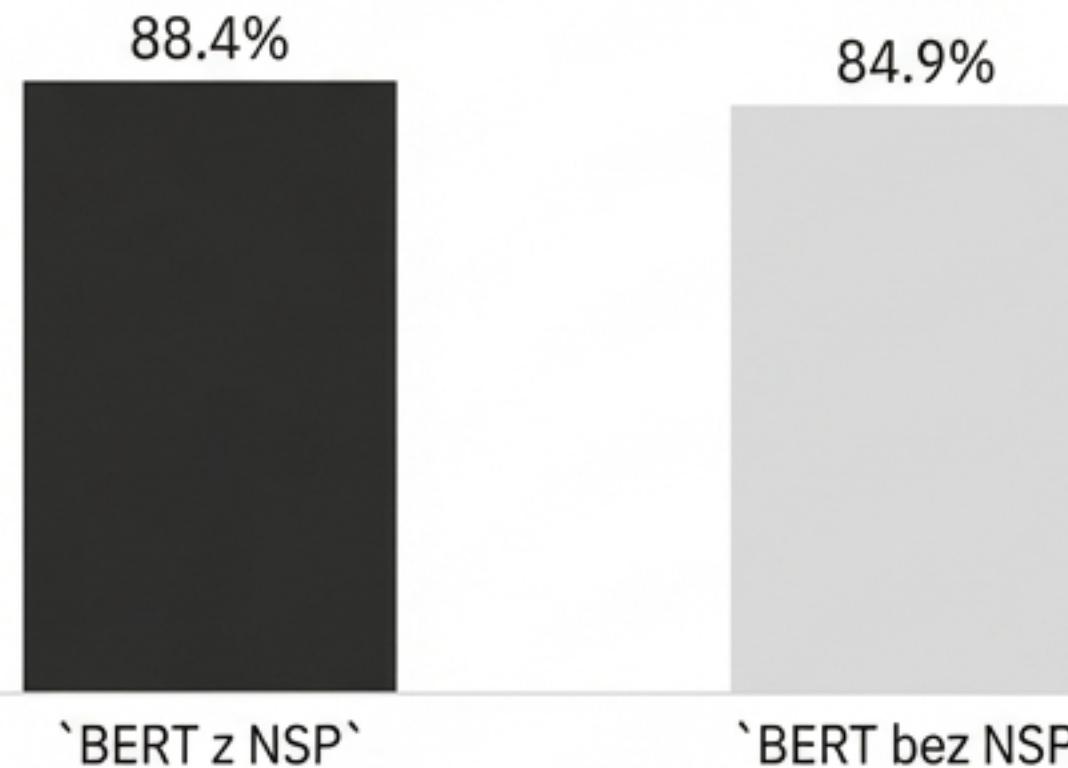
- * **Cel:** Nauczenie modelu ogólnego rozumienia języka.
- * **Proces:** Trening na ogromnym, nieetykietowanym korpusie danych (MLM i NSP).
- * **Koszt:** Bardzo wysoki (4 dni na 64 chipach TPU dla BERT-LARGE).

- * **Cel:** Adaptacja modelu do specyficznego zadania.
- * **Proces:** Dostrajanie wag na małym, etykietowanym zbiorze danych.
- * **Koszt:** Relatywnie niski (~30 minut na jednym Cloud TPU dla SQuAD).

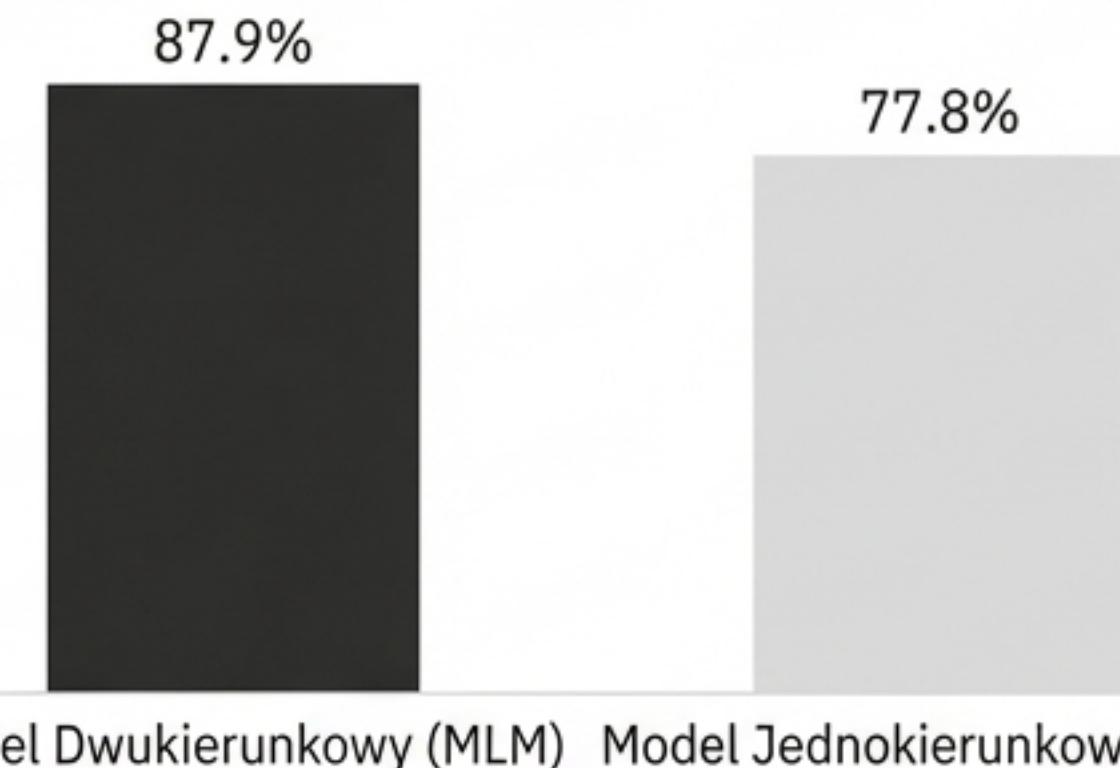
Analiza ablacyjna: Co tak naprawdę stało za sukcesem BERTa?

Aby zrozumieć znaczenie każdego komponentu, autorzy przeprowadzili eksperymenty, usuwając kluczowe elementy z architektury 'BERT BASE'.

Wpływ usunięcia NSP na zadanie QNLI



Wpływ dwukierunkowości na zadanie SQuAD



Wniosek: NSP jest kluczowe dla zadań wymagających rozumowania na poziomie relacji między zdaniami.

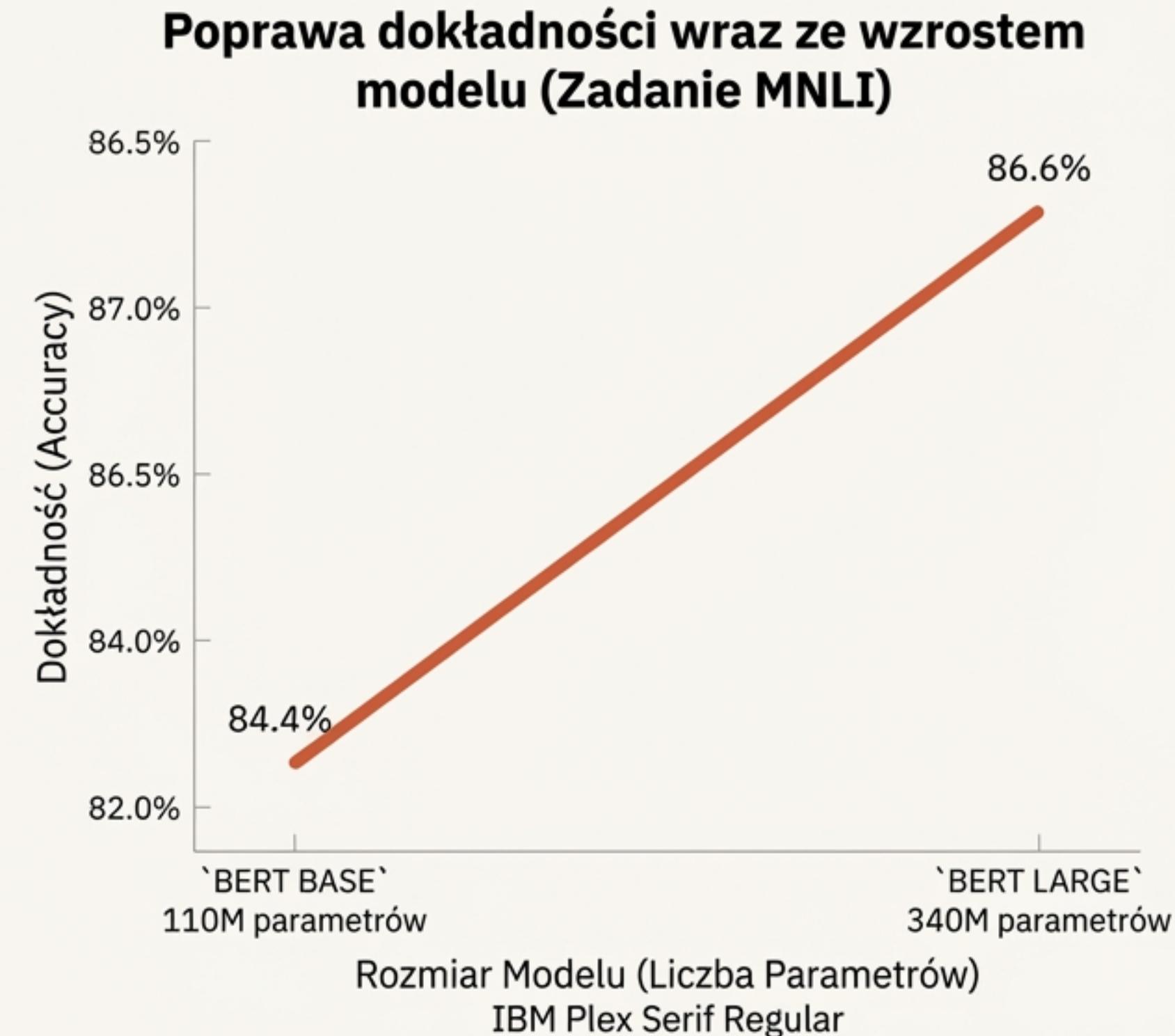
Wniosek: Głęboka dwukierunkowość, którą umożliwia MLM, jest najważniejszym czynnikiem przewagi BERTa.

Efekt skali: Większe modele prowadzą do lepszego rozumienia języka

Autorzy zbadali wpływ rozmiaru modelu na dokładność, porównując dwie wersje:

- `BERT BASE`
 - 110M parametrów
 - 12 warstw, 768 jedn. ukrytych
- `BERT LARGE`
 - 340M parametrów
 - 24 warstwy, 1024 jedn. ukrytych

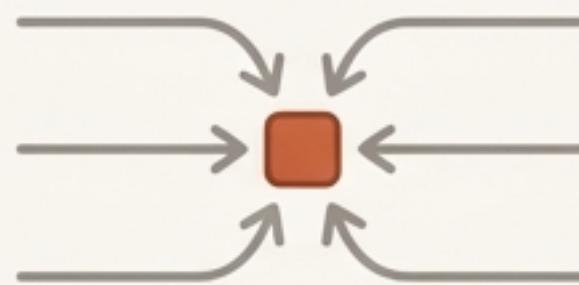
Kluczowe odkrycie: Większe modele prowadziły do ścisłej poprawy dokładności we wszystkich zadaniach, nawet tych z bardzo małą ilością danych treningowych. Był to dowód, że skalowanie, przy odpowiednim pre-treningu, jest kluczem do lepszych wyników.



Dziedzictwo BERTa: Fundamentalna zmiana w sposobie, w jaki maszyny rozumieją język

Techniczny

1. Głęboka Dwukierunkowość



Metodologiczny

2. Pre-trening i Fine-tuning



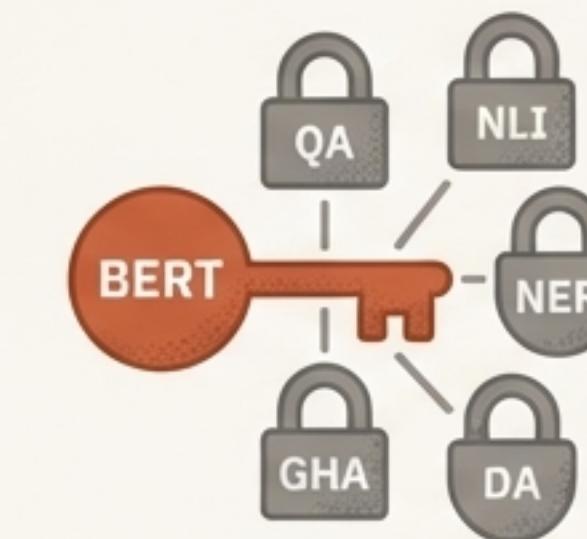
Umożliwiona przez **Masked Language Model (MLM)**, pozwoliła na pełne, kontekstowe rozumienie tekstu i relacji między słowami.

Ustanowił uniwersalny, efektywny proces, który zredukował potrzebę tworzenia dedykowanych architektur dla każdego zadania.

BERT stał się fundamentem dla całej generacji modeli językowych i jest integralną częścią systemów, z których korzystamy na co dzień, od wyszukiwarek po konwersacyjną AI.

Filozoficzny

3. Jedno Rozwiązanie, Wiele Zastosowań



Udowodnił, że rozwiązyując jeden ogólny problem (reprezentacja języka), można skutecznie radzić sobie z szerokim spektrum specyficznych zadań.