BERT: A New Era for Language Understanding

Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova

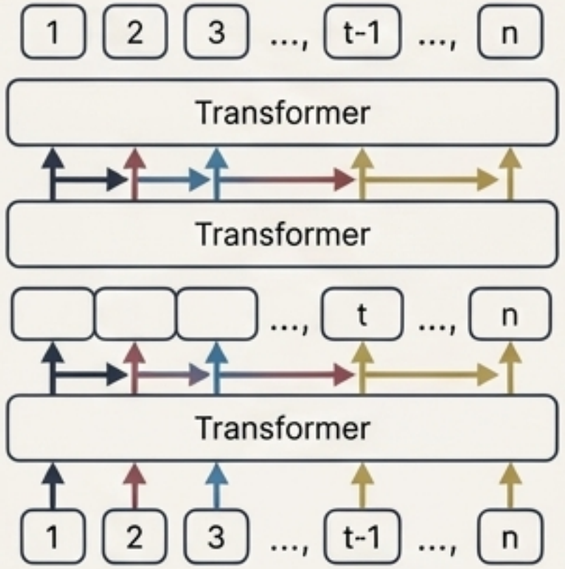
Google AI Language

NAACL 2019

Previous Models Lacked True Contextual Understanding

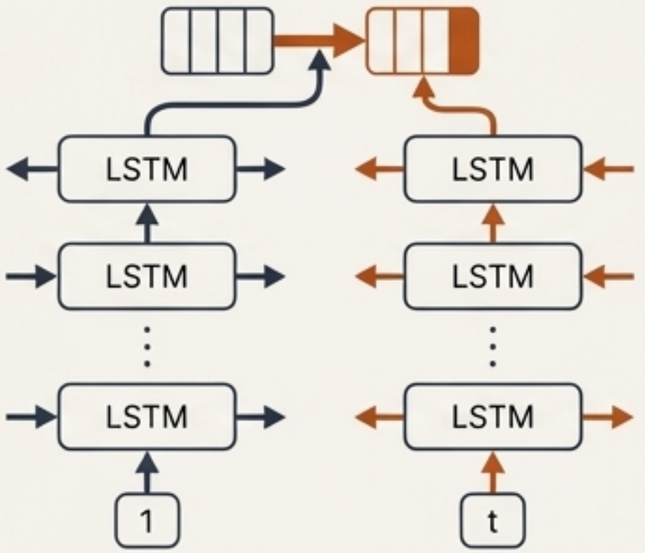
Pre-trained models were powerful but handicapped by their architecture. They could not consider both left and right context simultaneously in all layers.

OpenAI GPT (Unidirectional)



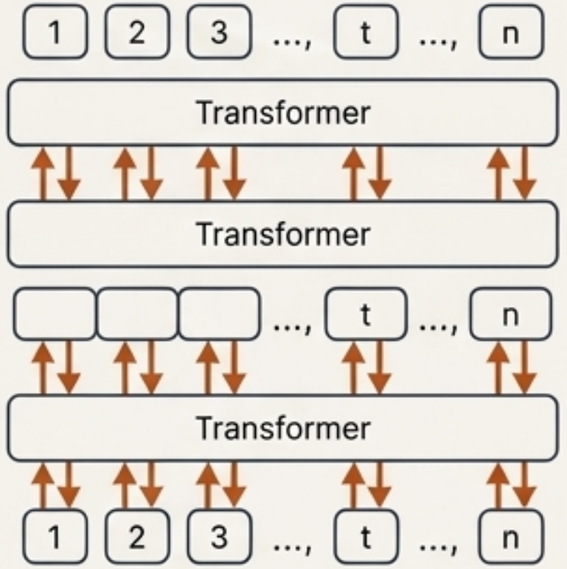
A strict left-to-right architecture. Each token can only attend to previous tokens.

ELMo ("Shallow" Bidirectional)



A concatenation of independently trained left-to-right and right-to-left models. The two streams are not deeply integrated.

BERT (Deeply Bidirectional)

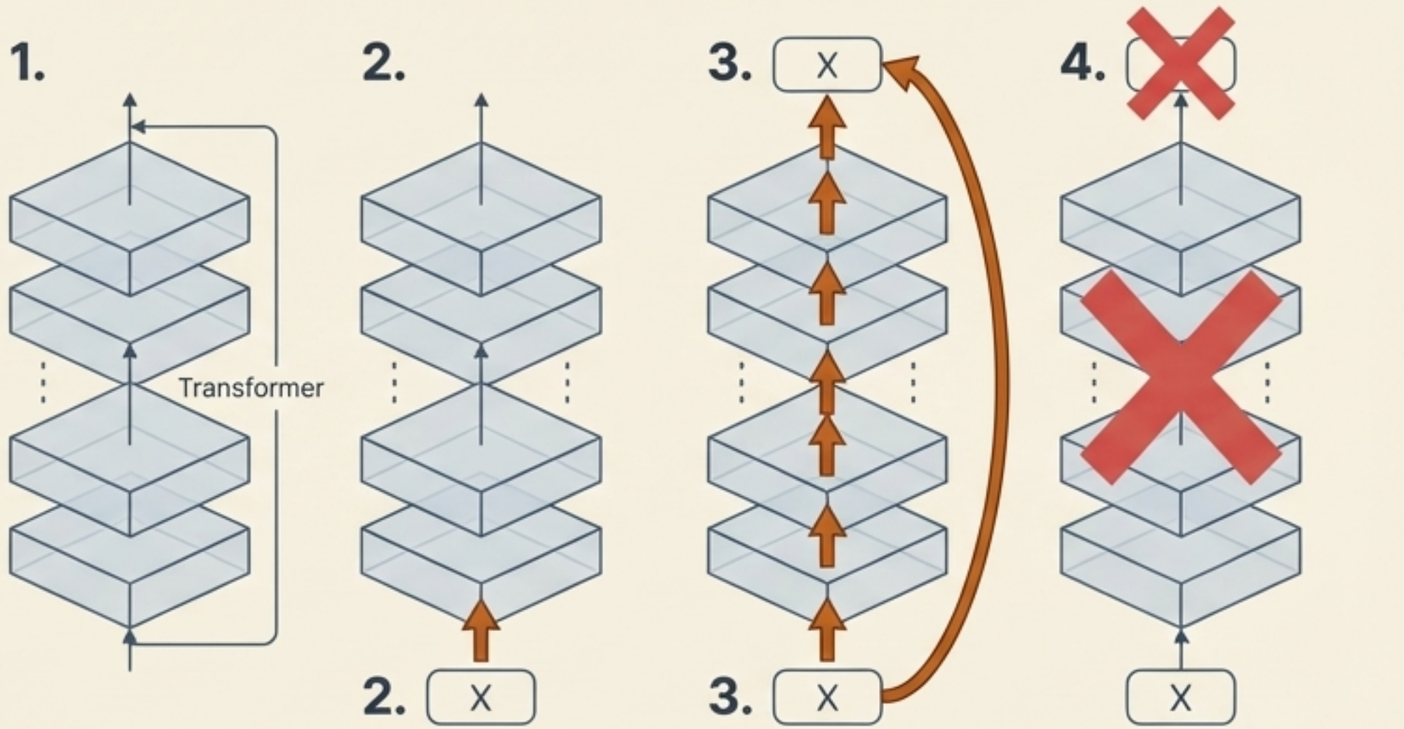


Representations are jointly conditioned on both left and right context in all layers.

Key Insight: The paper argues these restrictions are "sub-optimal" and "could be very harmful" for tasks where incorporating context from both directions is crucial.

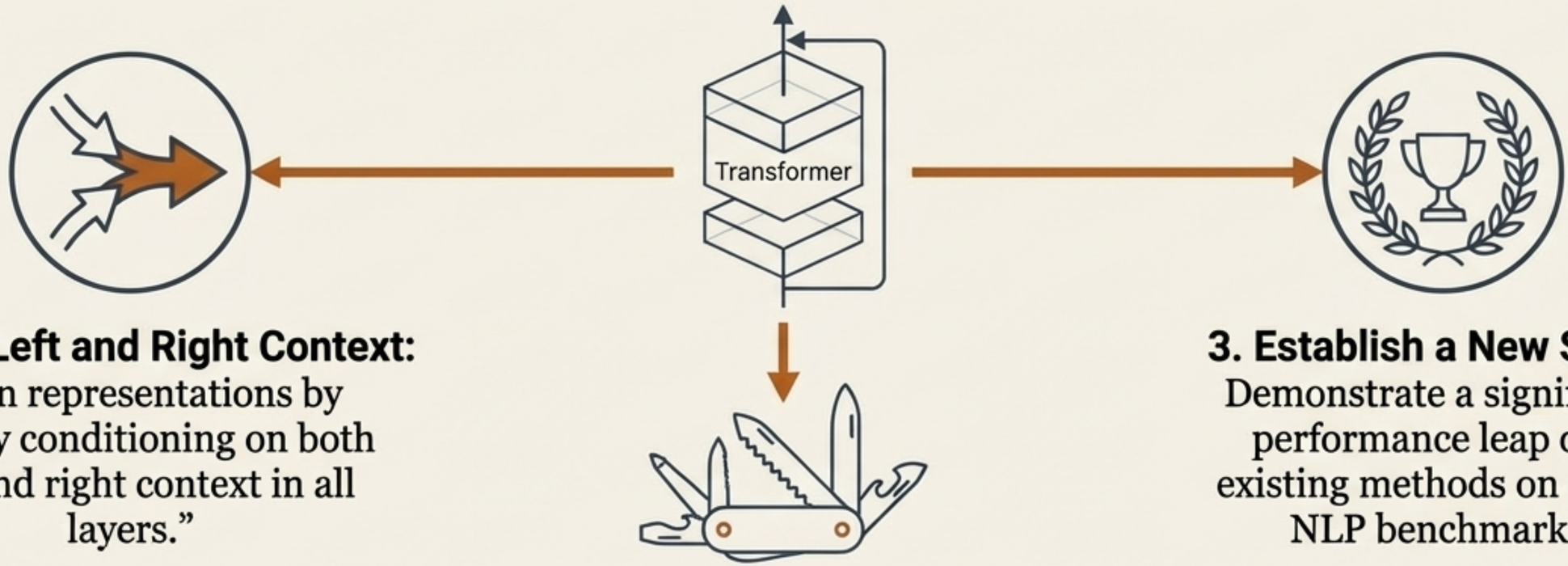
Why Not Just Look Both Ways? The “Cheating” Problem

- Standard language models could not be trained with deep bidirectional conditioning.
- In such a model, a word could indirectly “see itself”, allowing it to trivially predict the target word in a multi-layered context.
- This architectural limitation forced the field into unidirectional or shallowly-combined models.



The Gap: A new pre-training objective was needed to enable deep bidirectionality without allowing the model to cheat.

The Research Goal: Pre-training a Truly Deep Bidirectional Representation



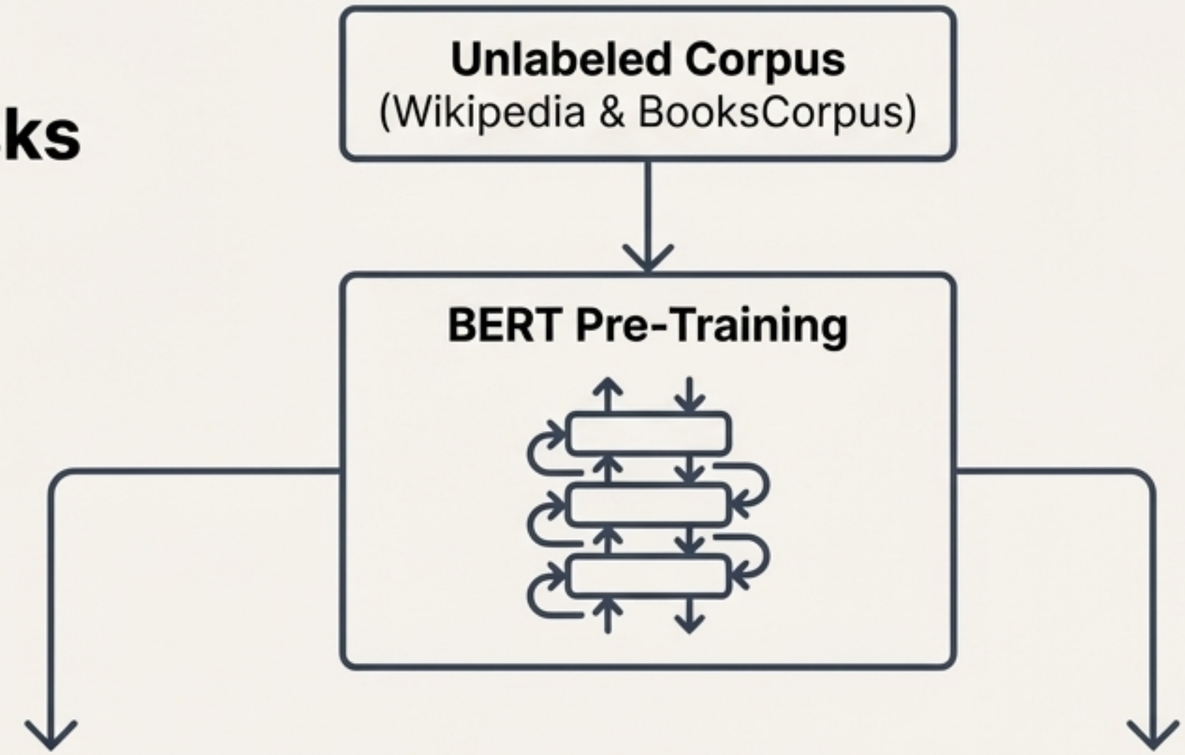
1. Fuse Left and Right Context:
Learn representations by "jointly conditioning on both left and right context in all layers."

2. Be Easily Fine-Tuned:
Create state-of-the-art models for a wide range of tasks with "just one additional output layer" and no substantial task-specific modifications."

3. Establish a New SOTA:
Demonstrate a significant performance leap over existing methods on major NLP benchmarks.

BERT's Solution: Two Novel Pre-Training Tasks

BERT is pre-trained on unlabeled text using two unsupervised tasks simultaneously, inspired by the Cloze task.



Masked Language Model (MLM)

the quick brown [MASK] jumps over...

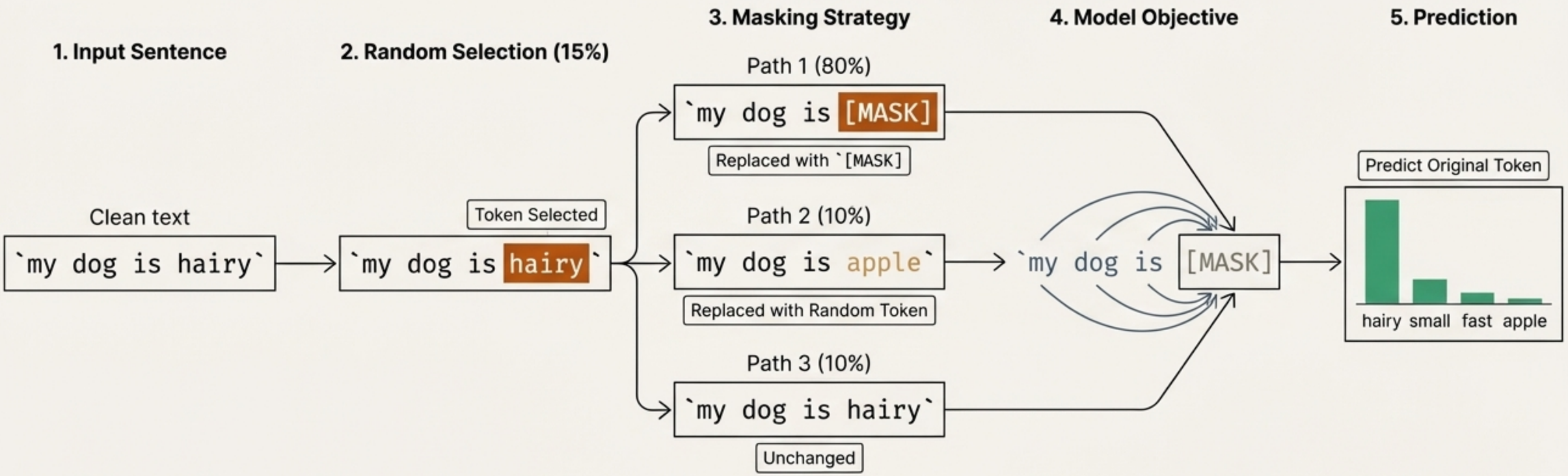
Solves the bidirectionality problem by forcing the model to predict randomly masked tokens.

Next Sentence Prediction (NSP)

Teaches sentence relationships, crucial for downstream tasks like Q&A.

How Masked LM Unlocks Bidirectionality

Instead of predicting the *next* word, BERT predicts randomly masked words using their full context. 15% of all WordPiece tokens are chosen at random for this task.



Result: This forces the model to maintain a rich, distributional representation of every token, enabling deep bidirectionality.

Teaching Sentence Relationships with with Next Sentence Prediction (NSP)

Many downstream tasks like Question Answering (QA) and Natural Language Inference (NLI) depend on understanding inter-sentence relationships. The model is trained on a simple binary classification task using the `[CLS]` token.

Sentence A: `the man went to the store`

Sentence B: `he bought a gallon of milk`

IsNext

50% of training pairs are actual consecutive sentences.

Sentence A: `the man went to the store`

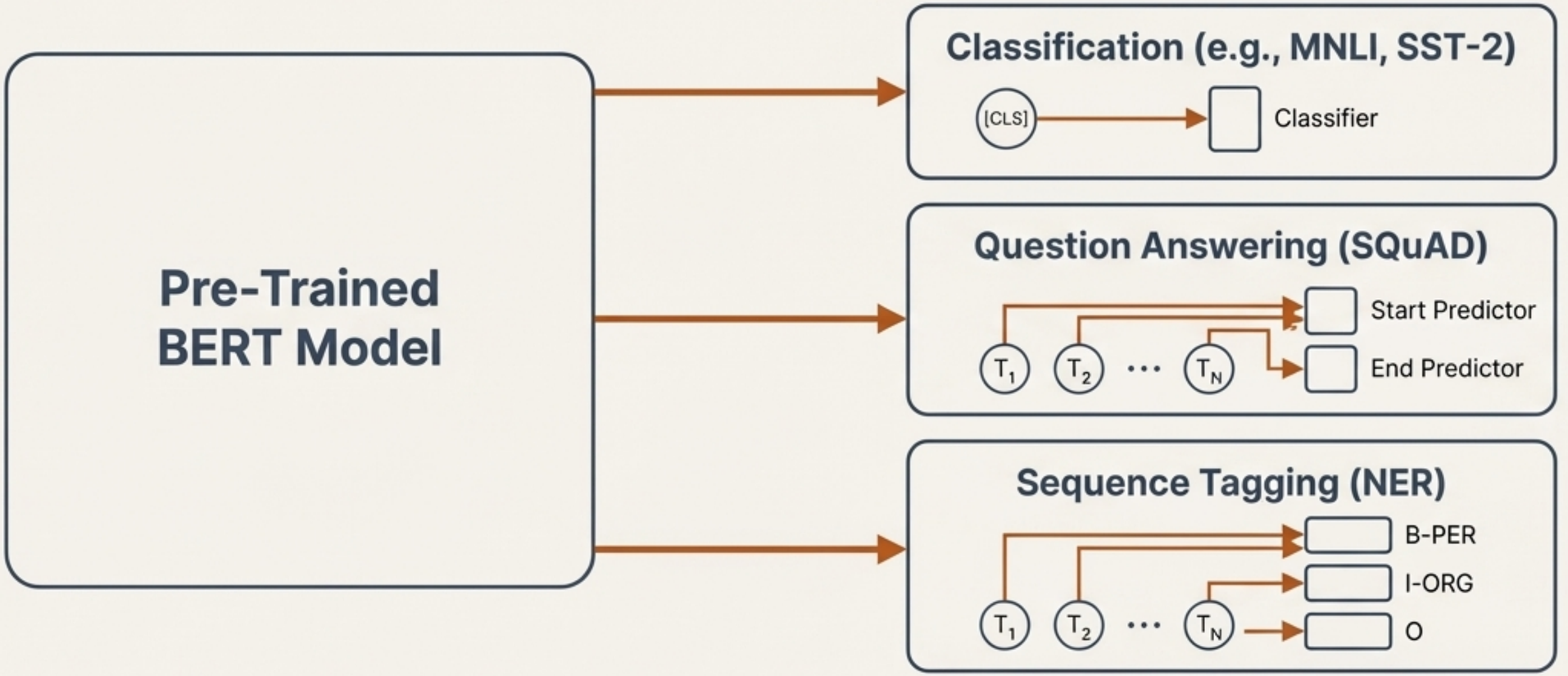
Sentence B: `penguins are flightless birds`

NotNext

50% of training pairs are random pairings.

One Pre-Trained Model, Many Downstream Tasks

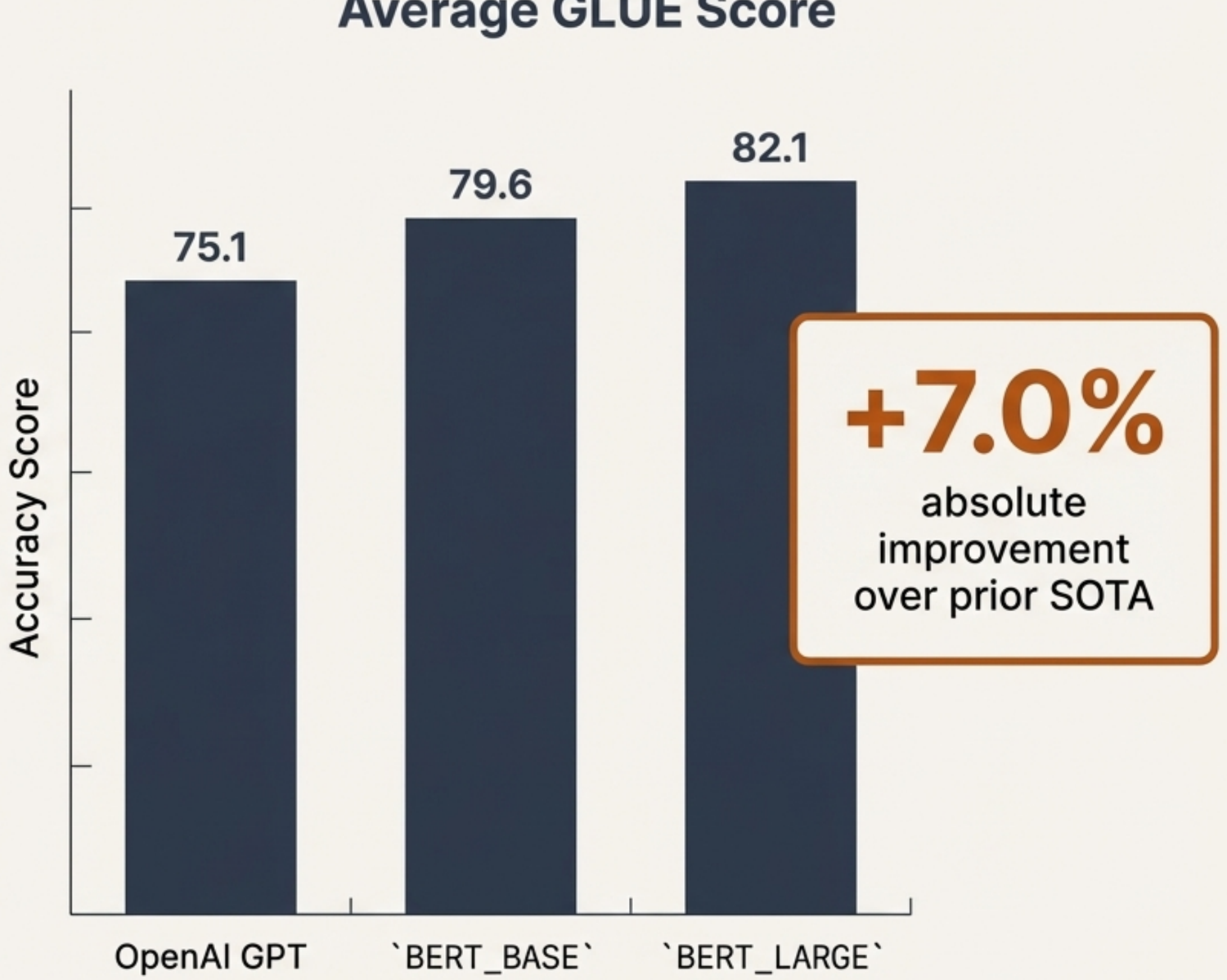
Fine-tuning is straightforward and computationally inexpensive. The same pre-trained model is adapted with minimal additions for a variety of tasks.



Fine-tuning can be done in “at most 1 hour on a single Cloud TPU, or a few hours on a GPU.”

BERT Achieved a New State-of-the-Art on 11 NLP Tasks

On the General Language Understanding Evaluation (GLUE) benchmark, BERT_LARGE obtained a score of 80.5, a 7.7% absolute improvement over the previous state-of-the-art.



Dominance in Question Answering and Commonsense Inference

BERT's deep bidirectional context proved exceptionally powerful for complex reasoning.

SQuAD v1.1
(Question Answering)

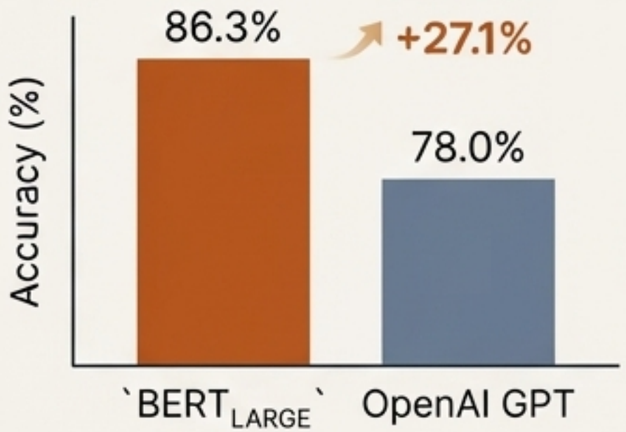
Surpassed human-level performance.

SQuAD v2.0
(Unanswerable Questions)

+5.1 F1

Massive improvement over prior state-of-the-art.

SWAG
(Commonsense Inference)

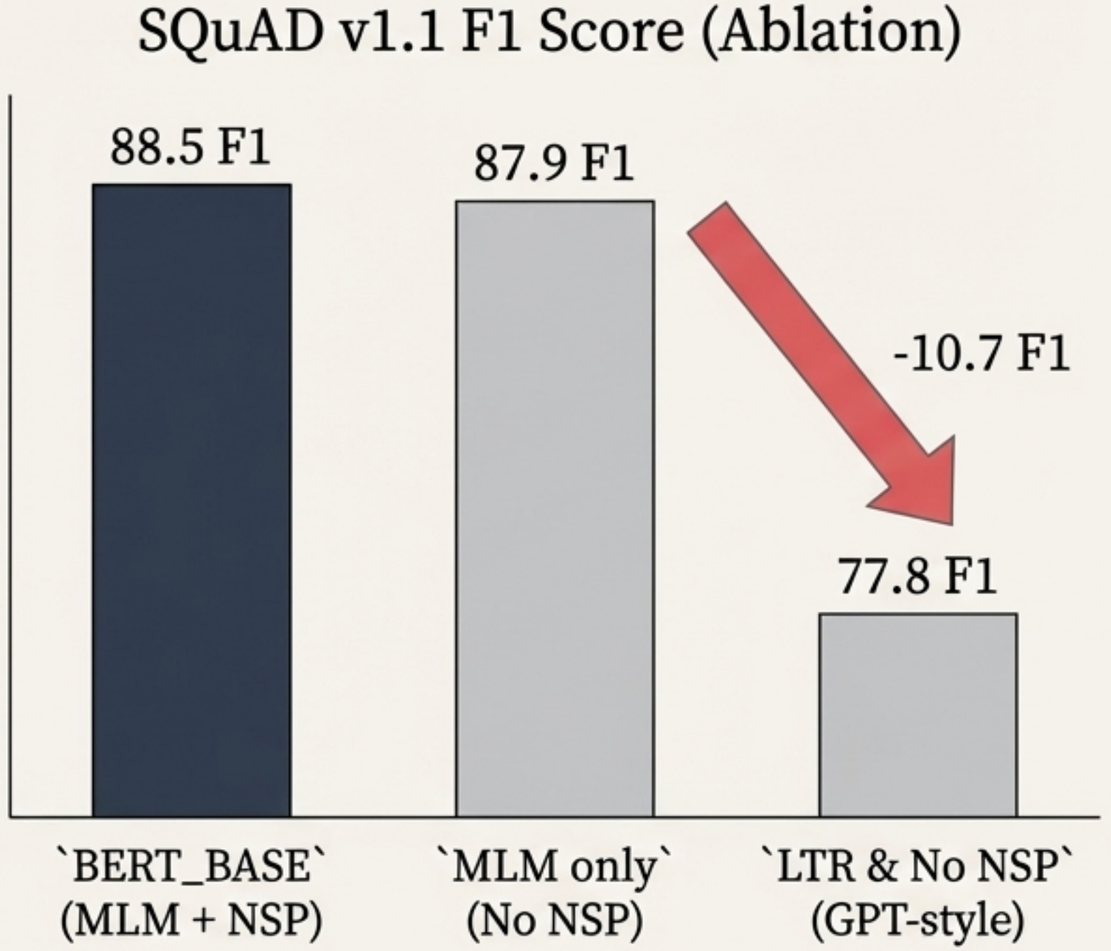


Outperformed the prior SOTA by +27.1%.

Ablation Studies Prove Bidirectionality is the Key

Experiments confirmed the importance of BERT's core design choices by removing them and observing the performance.

- **Removing NSP:** Hurt performance significantly on QNLI, MNLI, and SQuAD.
- **Using Left-to-Right (LTR) Model:** Performance dropped on all tasks, with a massive drop on SQuAD.



Conclusion: The MLM approach enabling deep bidirectionality is "strictly more powerful" than left-to-right or shallowly concatenated models.

BERT Redefined the NLP Playbook

The New Standard

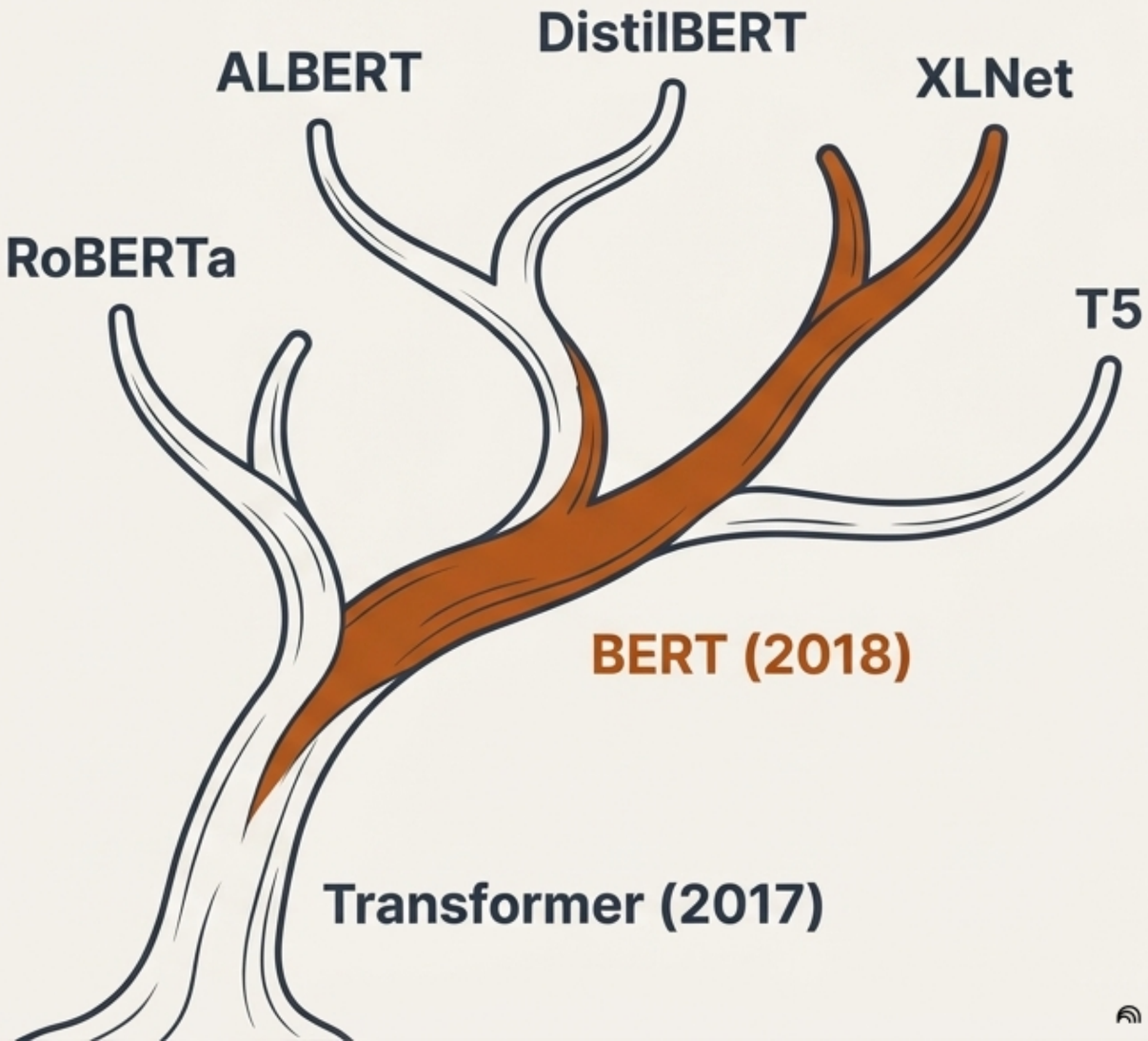
The ‘pre-train and fine-tune’ approach with a deep bidirectional Transformer became the dominant methodology.

Reduced Engineering

The model’s power “reduce[d] the need for many heavily-engineered task-specific architectures.”

Foundation for a New Generation

BERT’s core insights directly inspired a subsequent ‘Cambrian explosion’ of more advanced Transformer-based models.



Limitations and Future Directions

BERT was a massive leap, but it also pointed the way toward future challenges, inspiring the next wave of research.

Computational Cost

Pre-training BERT is extremely expensive, requiring massive datasets and hardware, which spurred research into more efficient models.

[MASK] Token Mismatch

The `[MASK]` token used in pre-training doesn't appear during fine-tuning, creating a discrepancy that later work sought to resolve.

Encoder-Only Architecture

BERT is ideal for understanding tasks but not for free-form text generation, leading to the development of encoder-decoder and decoder-only models.

The Bidirectional Revolution

The Problem

Language models were fundamentally constrained by a unidirectional view of text, limiting their contextual awareness.

The Solution

The novel Masked Language Model (MLM) objective unlocked the ability to pre-train a truly deep bidirectional Transformer.

The Impact

BERT established a new state-of-the-art across a wide suite of NLP tasks and created the foundational paradigm for modern large language models.

Final Takeaway: BERT proved that how a model learns from context is as important as its architecture, making deep, simultaneous context the new standard for language understanding.