Language Models are Unsupervised Multask Learners

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever

OpenAI, 2019

The Prevailing Paradigm: Powerful but Narrow Experts

Status Quo:
Machine learning systems excel at the **single task** they are **trained for**, using large, task-specific, supervised datasets.

The Problem:
These systems are **brittle** and sensitive to slight changes in data distribution or task specification. They are “narrow experts rather than competent generalists.”

The Goal:
Move towards **general** systems that can perform many tasks without needing a custom-labeled dataset for each one.

Supervised Specialist

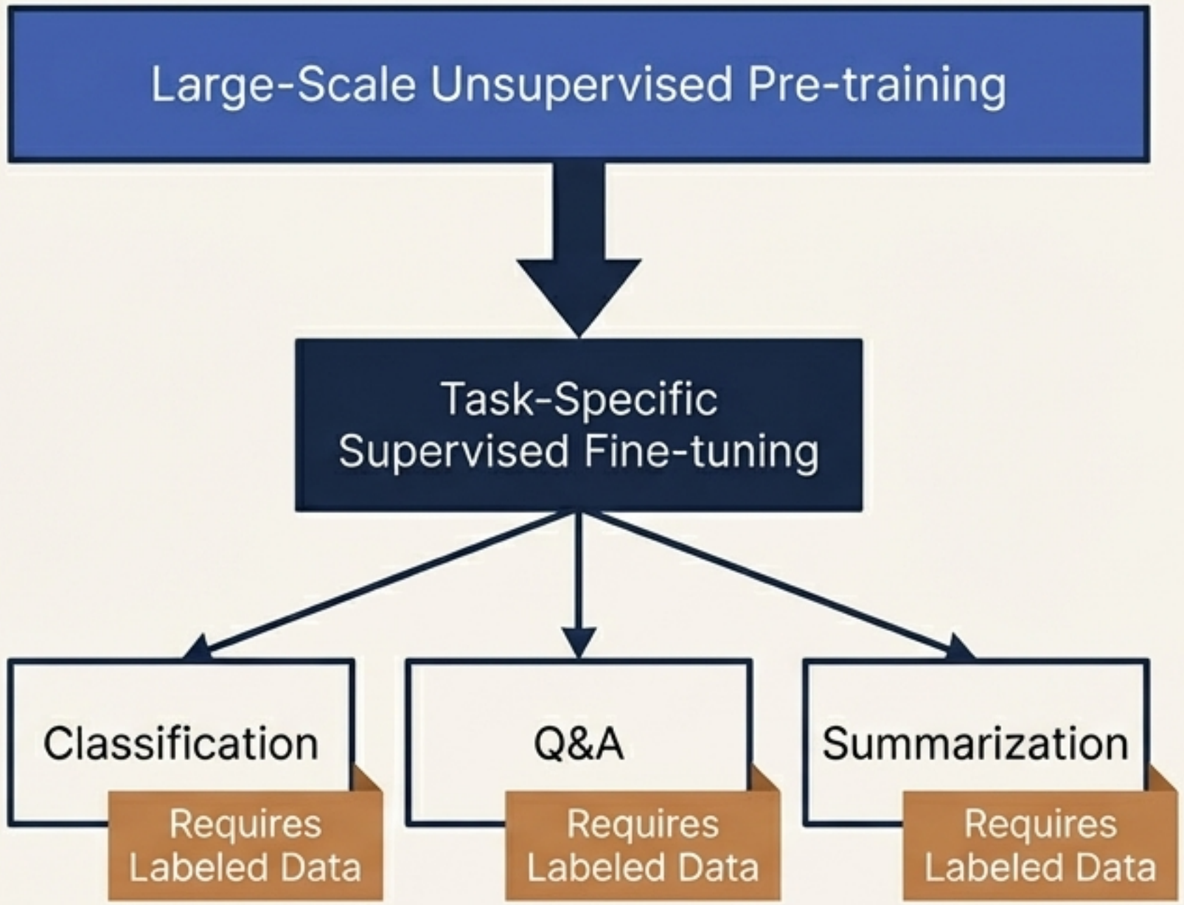
Unsupervised Generalist

The Bottleneck: Supervised Fine-Tuning

Dominant Method: Pre-train a general model on a large text corpus (like GPT or BERT), then fine-tune it with supervised learning for a specific downstream task.

The Dependency: This fine-tuning step requires thousands of labeled examples for every new task, creating a significant bottleneck for scaling.

The Research Question: Can a language model learn to perform these tasks in a **zero-shot setting**—without any parameter updates or fine-tuning?



The Hypothesis: Every Task is a Language Modeling Problem

A powerful language model can infer a task from a natural language prompt, without explicit labels. The model should learn p(output | input, task). The standard language modeling objective—predicting the next word—is a sufficient training signal. The “supervised objective is the same as the unsupervised objective but only evaluated on a subset of the sequence.”

The Bet: A large enough model, trained on a diverse enough dataset, will learn to complete task patterns it sees in the wild.

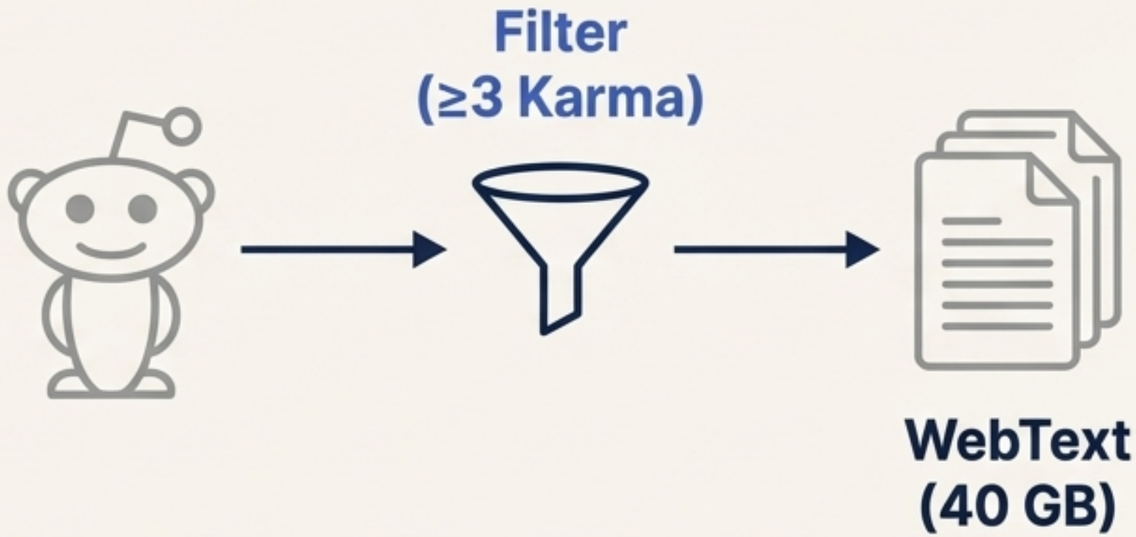
Translation:
(translate to french, english text, french text)

Reading Comprehension:
(answer the question, document, question, answer)

Ingredient 1: A Massive, Diverse Dataset Called WebText

- **Motivation:** To learn diverse tasks, the model needs to see “natural language demonstrations” from varied domains, not just news or Wikipedia.

- **Source:** Scraped all outbound links from Reddit that received at least 3 karma. This acts as a human-driven quality filter.
- **Scale:** 40 GB of text from over 8 million documents.
- **Key Feature:** Designed for high diversity and quality. Wikipedia was explicitly removed to avoid contaminating downstream evaluation tasks.

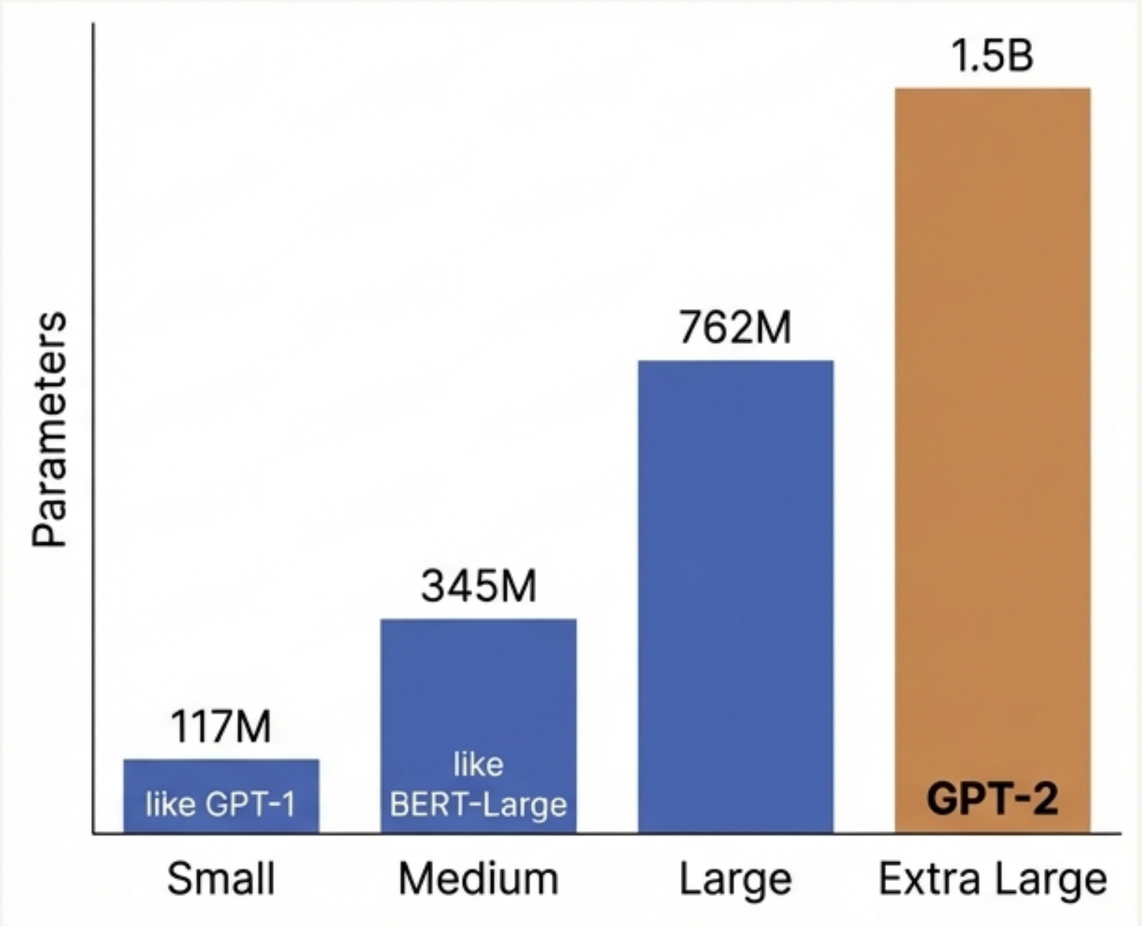


Ingredient 2: A Scaled-Up Transformer Architecture

Architecture: A decoder-only Transformer, largely following the design of the original GPT model.

Key Modification: A larger context window of 1024 tokens (from 512) and minor changes to layer normalization and initialization.

The Main Innovation: Scale. The paper tested four model sizes, with the largest being GPT-2.



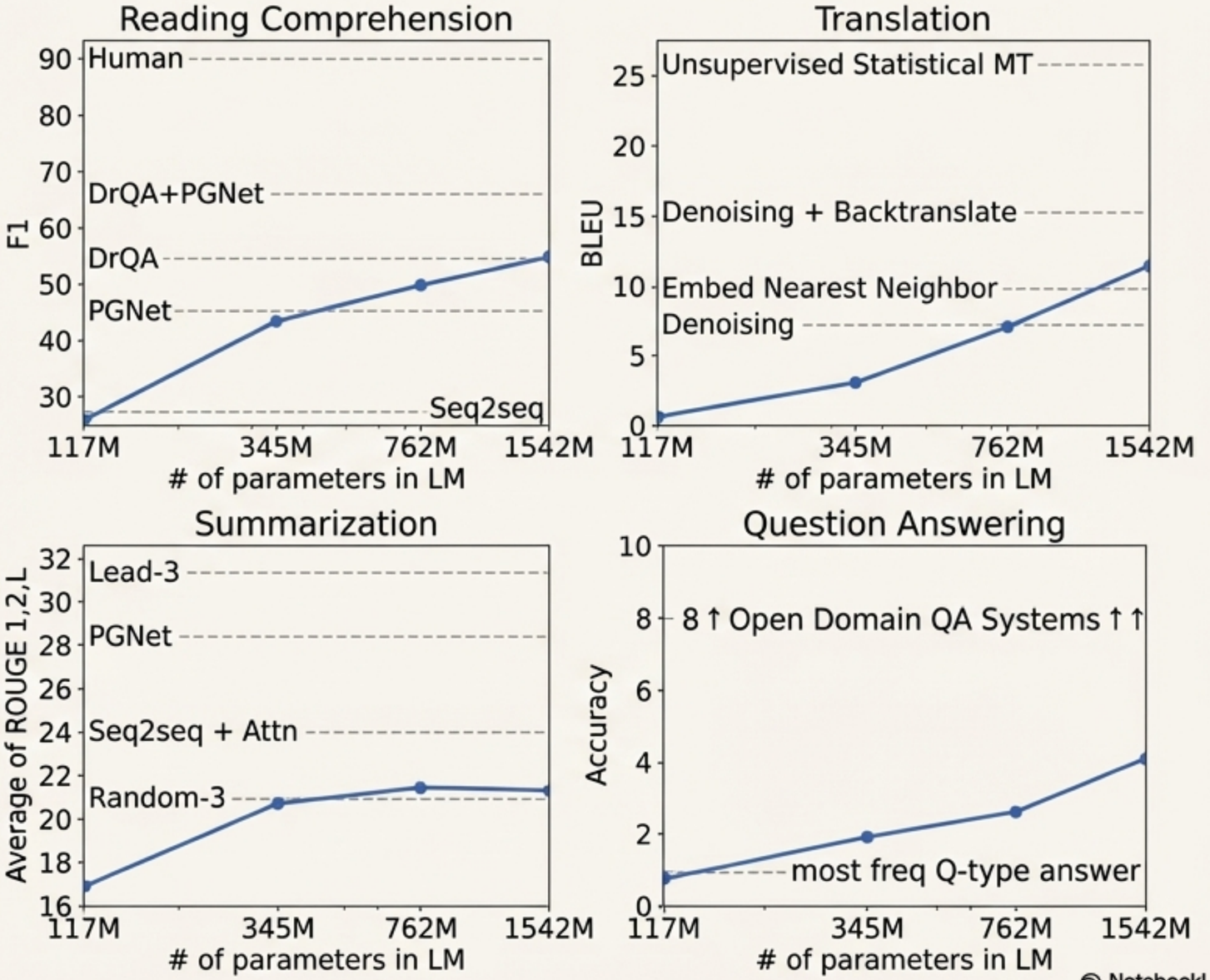
The Breakthrough: Zero-Shot Performance Scales with Model Size

Larger models consistently perform better across a range of tasks, **without any task-specific training (zero-shot)**.

The performance improvement is often log-linear.

Key Results:
Reading Comprehension (CoQA): F1 score improves from ~28 to 55.
Translation (WMT-14 Fr-En): BLEU score improves from ~0 to 11.5.

Summarization (CNN/DM): ROUGE score improves from ~16 to ~22.
Question Answering (Natural Questions): Accuracy improves from ~1% to over 4.1%.



The Foundation: State-of-the-Art in Zero-Shot Language Modeling

GPT-2's core strength is its ability to model language. It achieved new state-of-the-art results on 7 out of 8 benchmark language modeling datasets. This was achieved in a **zero-shot setting**, demonstrating strong domain transfer.

Notable improvements:

- **LAMBADA (Accuracy)**: 63.24% (vs. 59.23% SOTA)
- **Children's Book Test (Accuracy)**: 89.05% on Named Entities (vs. 82.3% SOTA)
- **Penn Treebank (Perplexity)**: 35.76 (vs. 46.54 SOTA)

Dataset	Metric	Previous SOTA	GPT-2 (1.5B)
LAMBADA	Accuracy	59.23%	**63.24%**
Children's Book Test	Accuracy (NE)	82.3%	**89.05%**
WikiText-103	Perplexity	18.3	**17.48**
Penn Treebank	Perplexity	46.54	**35.76**

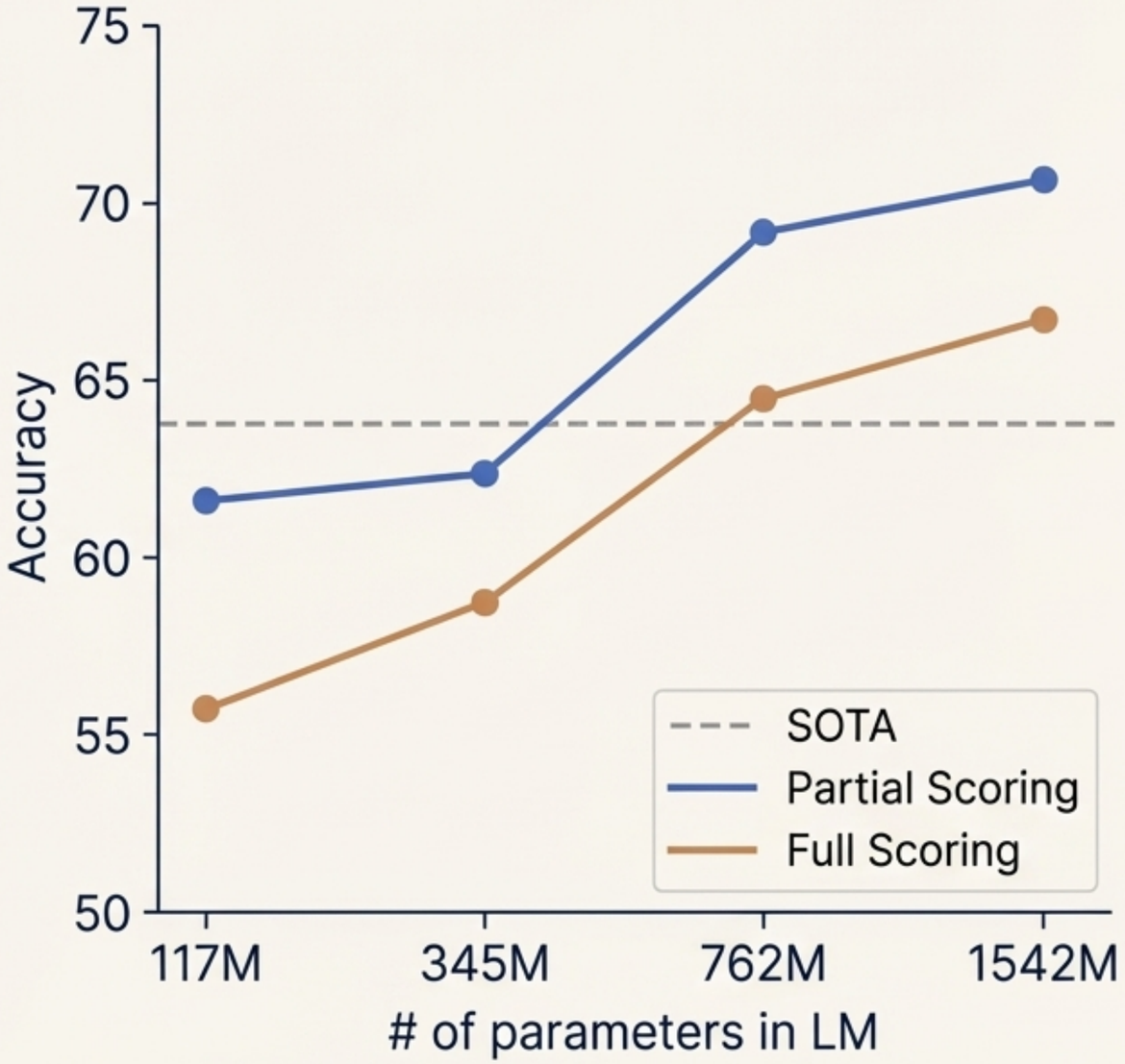
Glimpses of Emergent Commonsense Reasoning

The **Winograd Schema Challenge** tests a system's ability to resolve ambiguities that require commonsense knowledge.

Example: “The trophy would not fit in the suitcase because **it** was too big.” (What was “it”?)

Result: GPT-2 improved the state-of-the-art accuracy by 7%, achieving **70.70%**.

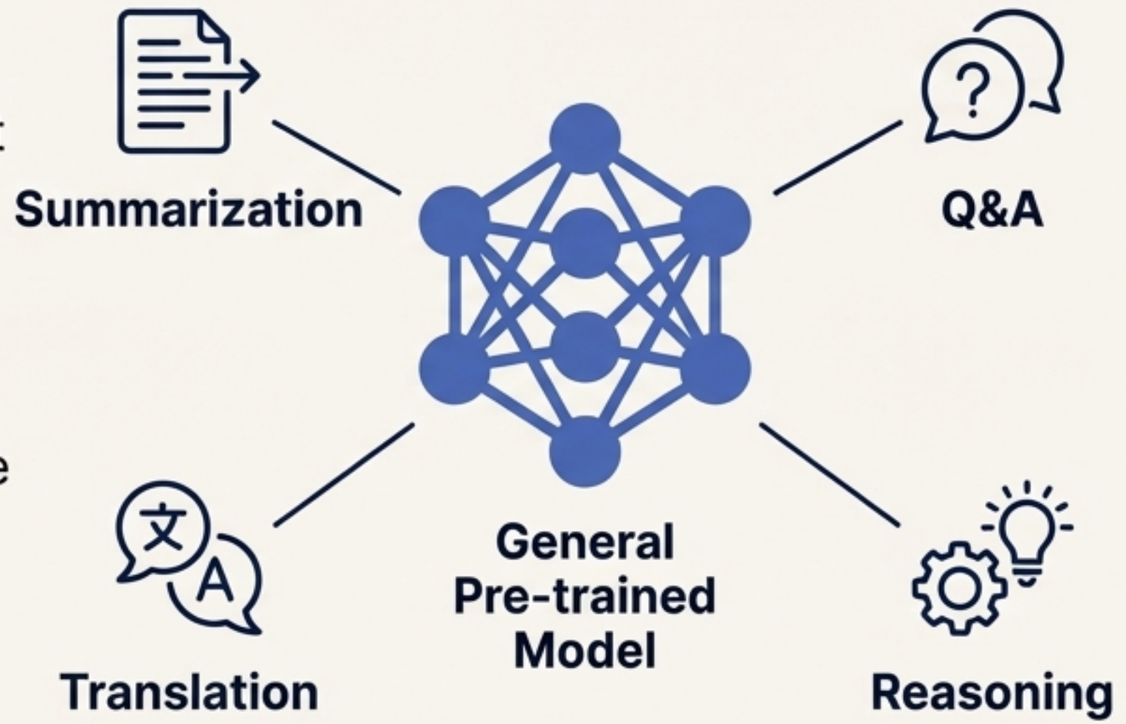
Significance: This suggests the model is acquiring world knowledge embedded within the training text, not just learning surface-level linguistic patterns.



The Implication: A New Paradigm for General-Purpose AI

Viability of Unsupervised Learning: High-capacity models trained on vast, diverse data can learn tasks without any explicit supervision.

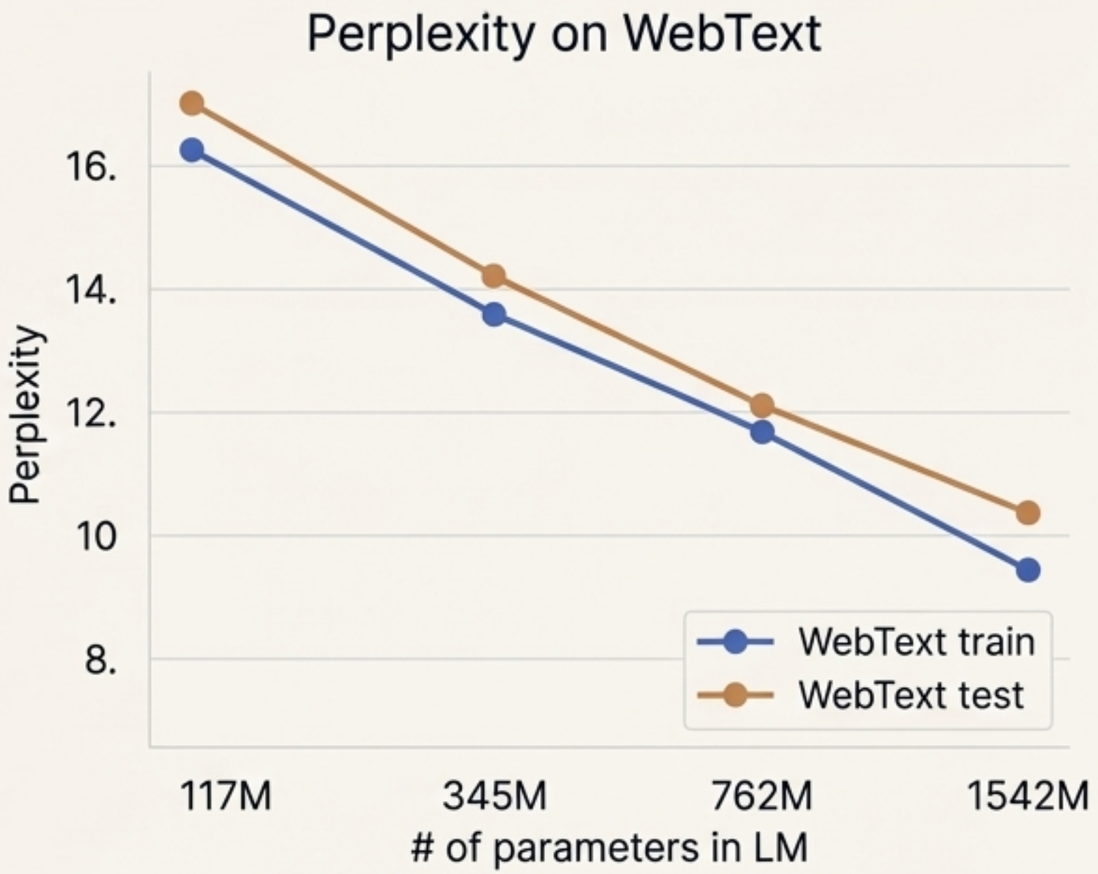
A New Research Direction: This work provided strong evidence for scaling up existing architectures as a primary method for unlocking emergent, general-purpose capabilities.



Towards 'Foundation Models': It demonstrated a path toward building single, large models that can be adapted to many reducing the reliance on expensive, human-labeled datasets.

A Reality Check: The Limitations of Zero-Shot Performance

- **Performance on Complex Tasks**: On tasks like summarization and translation, zero-shot performance was “rudimentary” and far below supervised state-of-the-art systems.
- **Reliance on Heuristics**: For question answering, the model often used “simple retrieval based heuristics” (e.g., answering “who” with a name from the text) rather than deep comprehension.
- **Still Underfitting**: Even the 1.5B parameter model was still underfitting the WebText dataset, suggesting performance could continue to improve with even more scale.



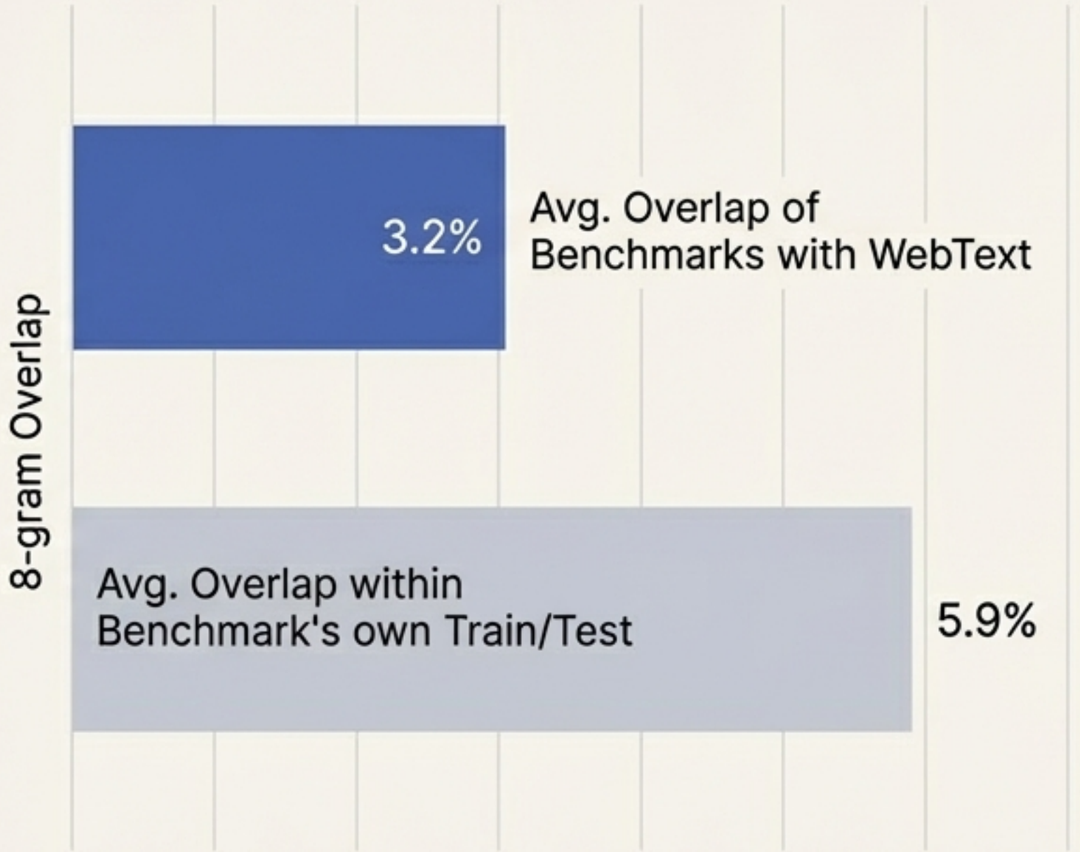
Generalization, Not Just Memorization

To test for memorization, an 8-gram overlap analysis was performed between the WebText training set and the test sets of various benchmarks.

Finding: The average overlap was 3.2%. This is comparable to, and in some cases less than, the overlap already present between the standard training and test sets of those benchmarks (avg. 5.9%).

LAMBADA Example: Recalculating metrics after excluding all overlapping examples barely changed the results (Accuracy shifted from 63.2% to 62.9%).

Conclusion: While data overlap provides a small benefit, it does not account for the significant performance gains.



Conclusion: A Scalable Path Toward General Models

The Finding: Scaling language models on large, diverse datasets enables **unsupervised multitask learning.**

The Mechanism: High-capacity models learn to perform tasks by inferring them from “naturally occurring demonstrations” in text, without needing explicit supervised signals.

The Impact: This work established a viable and scalable path toward more general-purpose AI systems, setting the stage for the next generation of large language models.

Q&A

Thank You