Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder,
Melanie Subbiah, Jared Kaplan, et al.

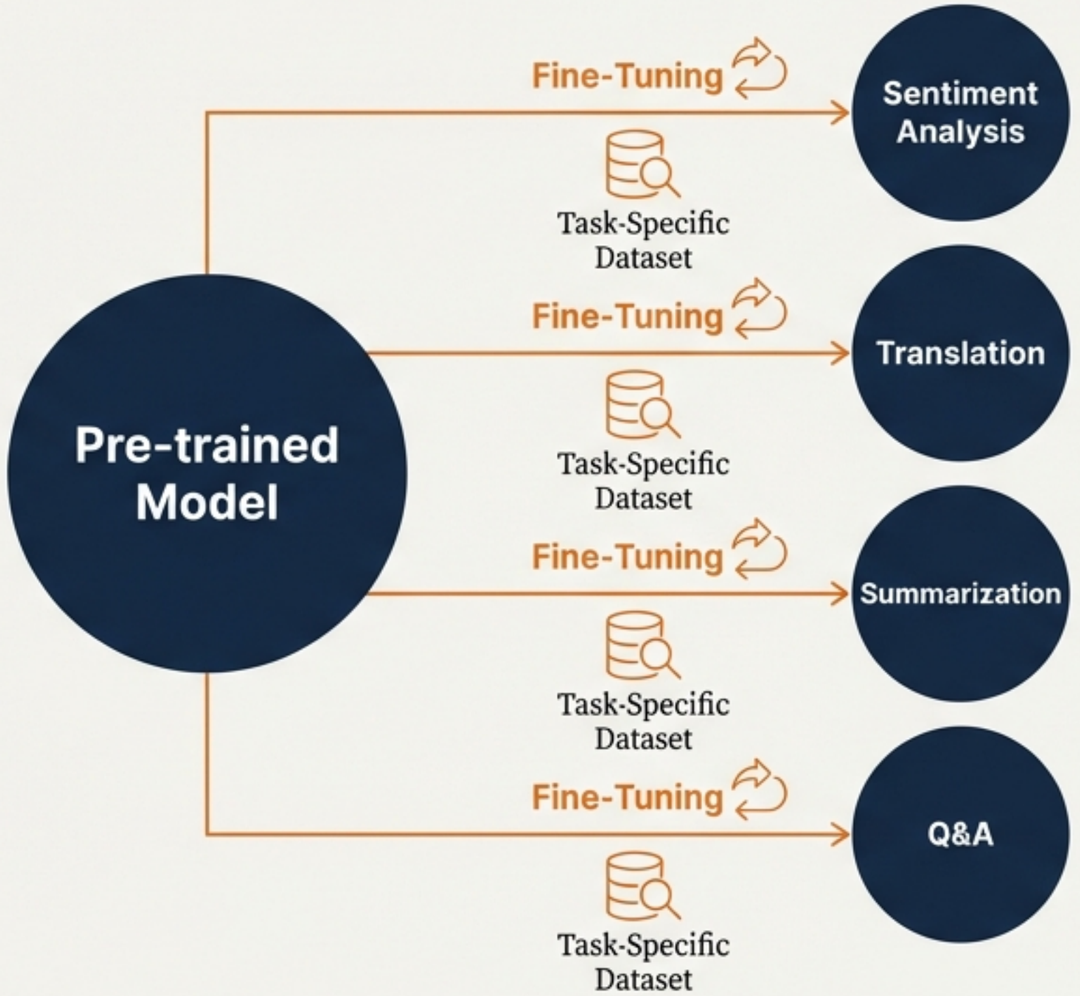
OpenAI

Conference on Neural Information Processing Systems (NeurIPS 2020)

The NLP Paradigm Was Powerful, But Inflexible

The dominant approach required creating specialized models for every individual task.

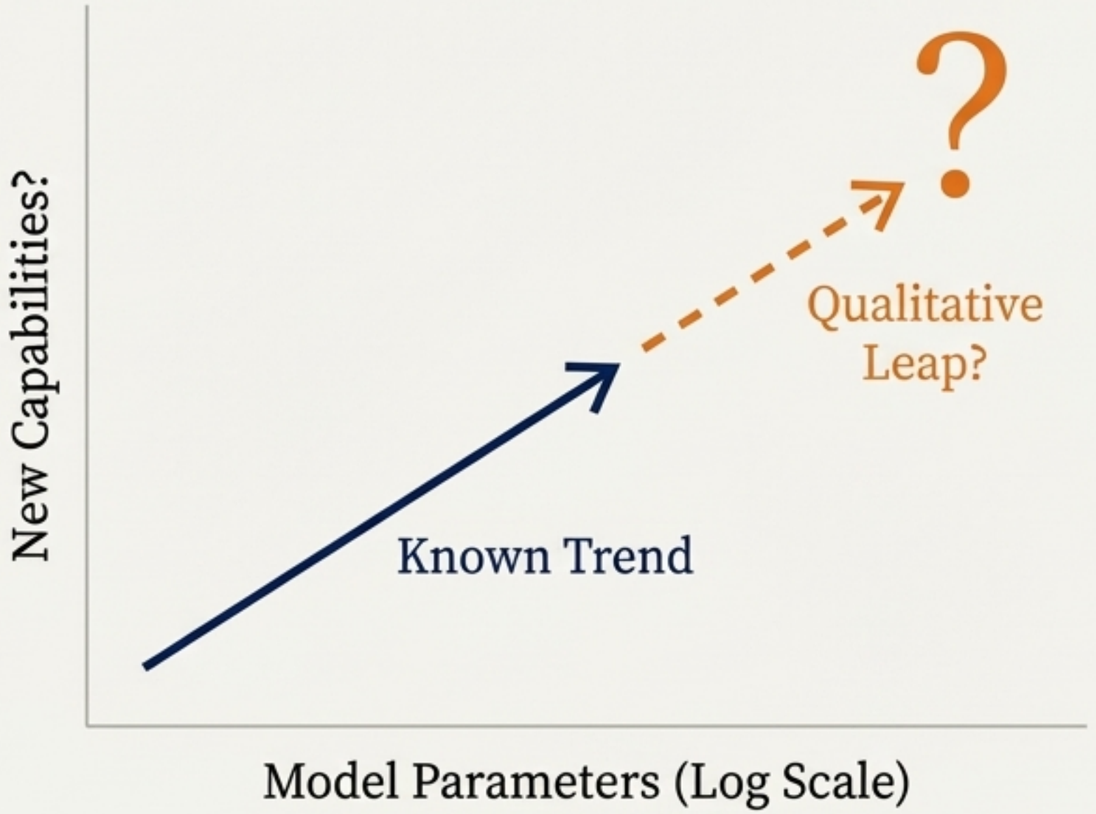
- **Pre-train & Fine-Tune:** The standard method involved taking a large pre-trained model and fine-tuning its weights on thousands of labeled examples for each new task.
- **Major Bottlenecks:** This process was data-hungry, computationally expensive, and limited a model's generality.
- **An Unnecessary Step?:** Recent work suggested this final fine-tuning step might not be necessary, but performance was far from state-of-the-art.



The Research Question: Could Massive Scale Eliminate Fine-Tuning?

Previous work showed performance improved predictably with model size, but it was unclear if this trend could unlock entirely new abilities.

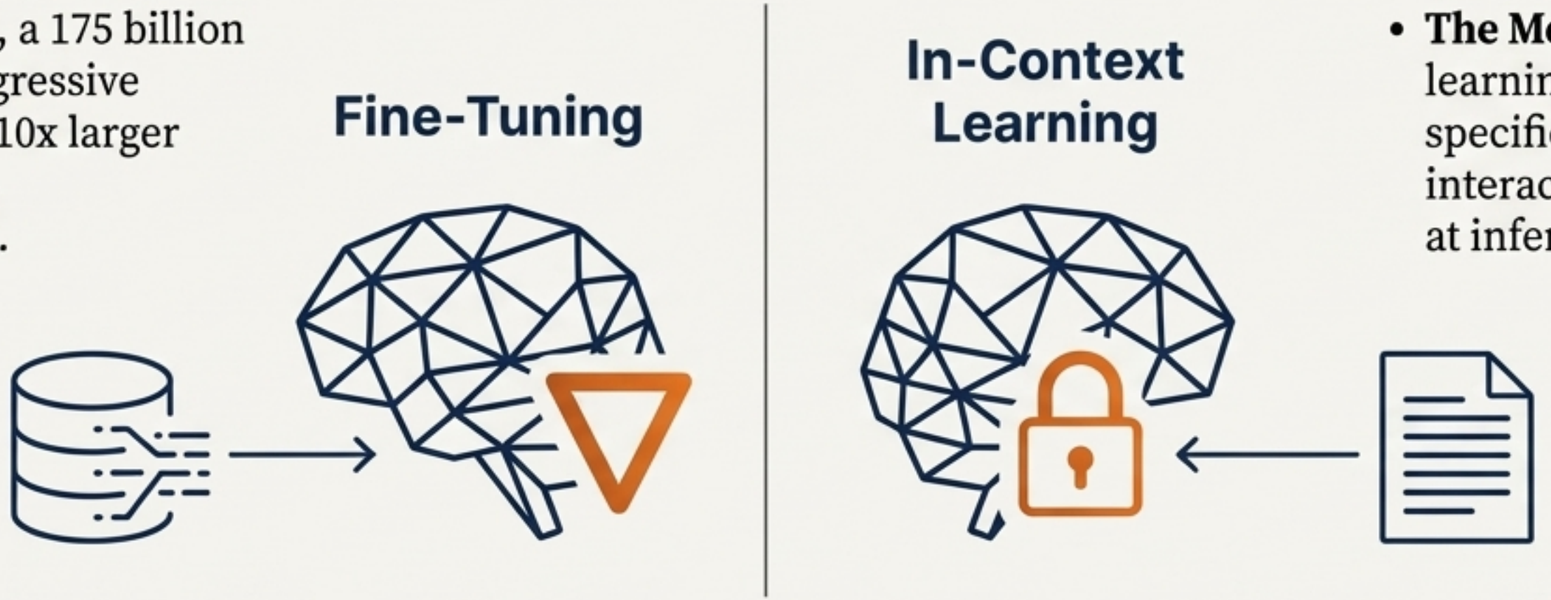
- **Known Trend:** Language model performance showed consistent log-linear improvements with scale.
- **Unproven Hypothesis:** Could scaling a model by another two orders of magnitude allow it to perform new tasks *without* any gradient updates?
- **The Core Question:** Can a sufficiently large language model generalize to new tasks based only on a few examples provided in its prompt?



The Approach: Testing In-Context Learning at 175B Scale

GPT-3 was designed as a massive-scale experiment to test task-agnostic learning.

- **The Model:** GPT-3, a 175 billion parameter autoregressive language model—10x larger than any previous non-sparse model.



- **The Method:** “In-context learning,” where tasks are specified purely via text interaction with the model at inference time.

- **The Key Constraint:** No gradient updates or fine-tuning. The model's weights remain frozen, testing its ability to learn "on the fly".

How It Works: Zero-Shot, One-Shot, and Few-Shot Prompts

The model is “programmed” through the prompt, with varying levels of demonstration.

Zero-Shot (0S)

Provide a natural language description of the task, but no examples.

Translate English to French.

cheese =>

[fromage]

One-Shot (1S)

Provide the instruction and a single complete example of the task.

Translate English to French.

sea otter => loutre de mer

cheese =>

[fromage]

Few-Shot (FS)

Provide the instruction and several examples that fit within the model's 2048-token context window.

Translate English to French.

sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe en peluche

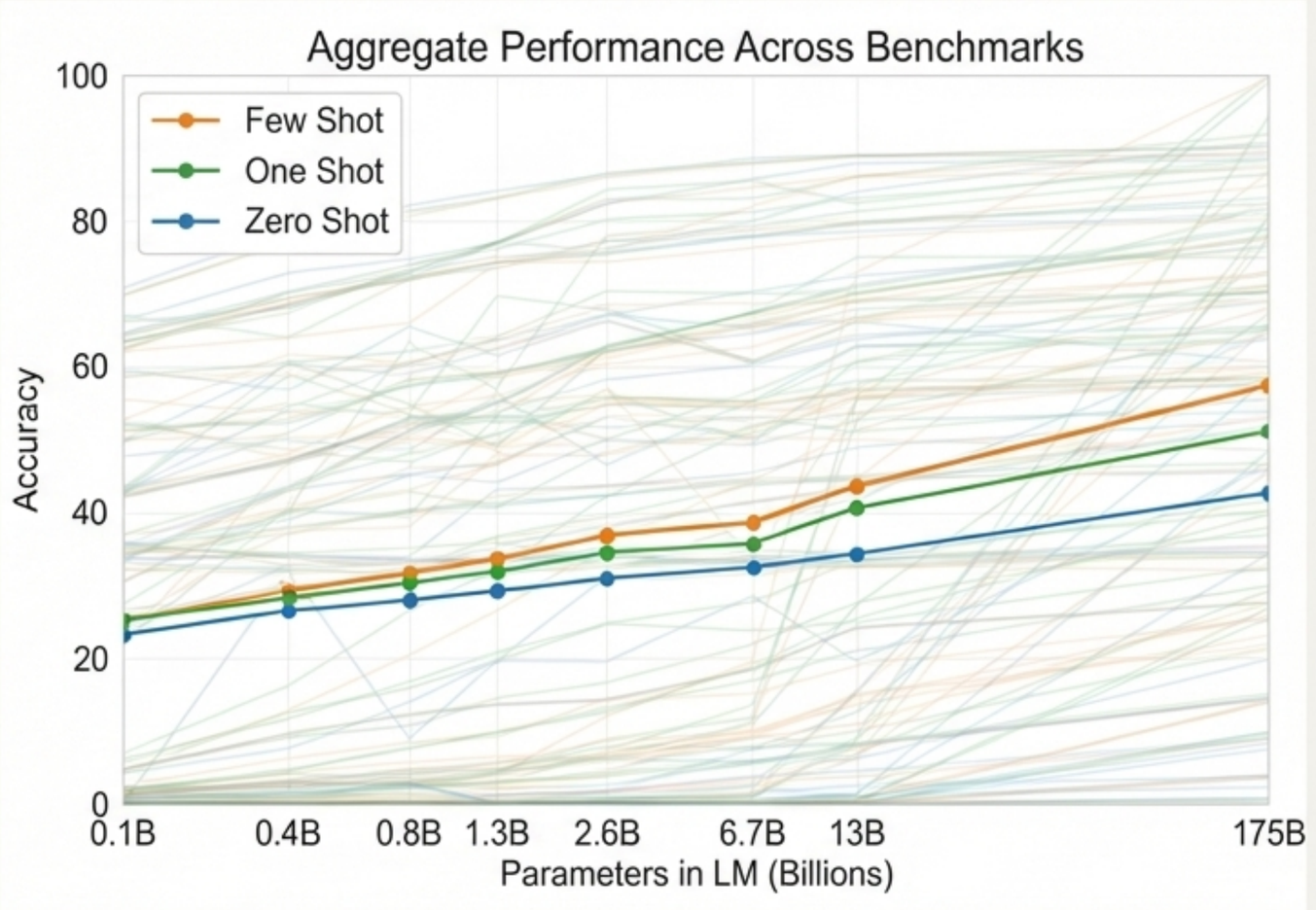
cheese =>

[fromage]

Finding 1: Performance Scales Smoothly and Predictably

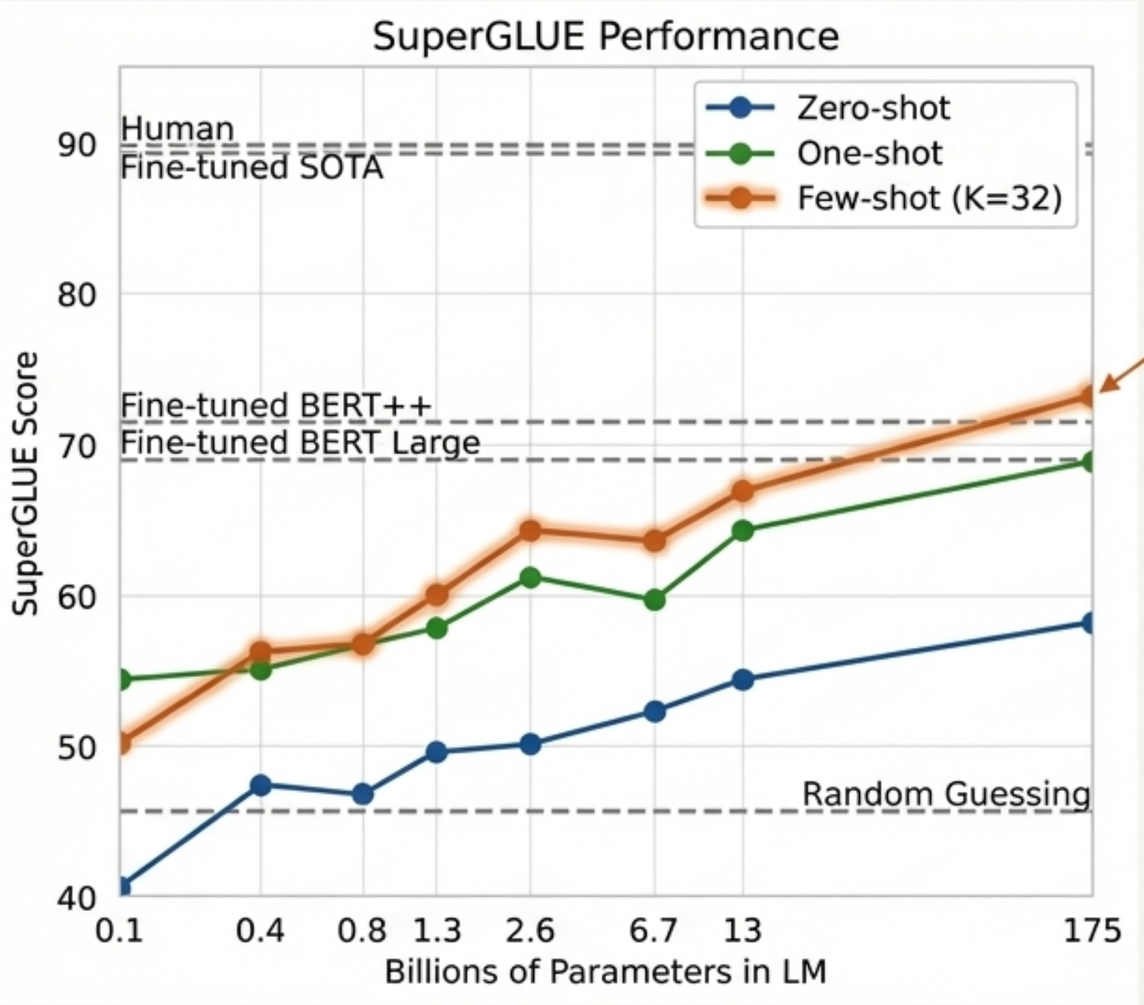
Across 42 benchmarks, larger models are simply better few-shot learners.

- **Consistent Improvement:** Model accuracy improves log-linearly with size for zero-shot, one-shot, and few-shot settings.
- **Few-Shot Scales Fastest:** The performance gap between few-shot and zero-shot learning *widens* as model size increases, demonstrating that larger models are more proficient ‘meta-learners’.
- **No Plateau:** The trend shows no signs of slowing, even at 175 billion parameters.



Finding 2: In-Context Learning Rivals Task-Specific Fine-Tuning

On the challenging SuperGLUE benchmark, few-shot GPT-3 approaches the performance of models specifically trained for the task.



- **Beats a Strong Baseline:** With just 32 in-context examples (K=32), the 175B GPT-3 surpasses a fine-tuned BERT-Large model.
- **More Examples Help:** Performance steadily improves with the number of examples provided in the context.
- **Generalist Power:** This is achieved without any task-specific training, using a single, task-agnostic model.

Finding 3: GPT-3 Achieves State-of-the-Art on Several Tasks

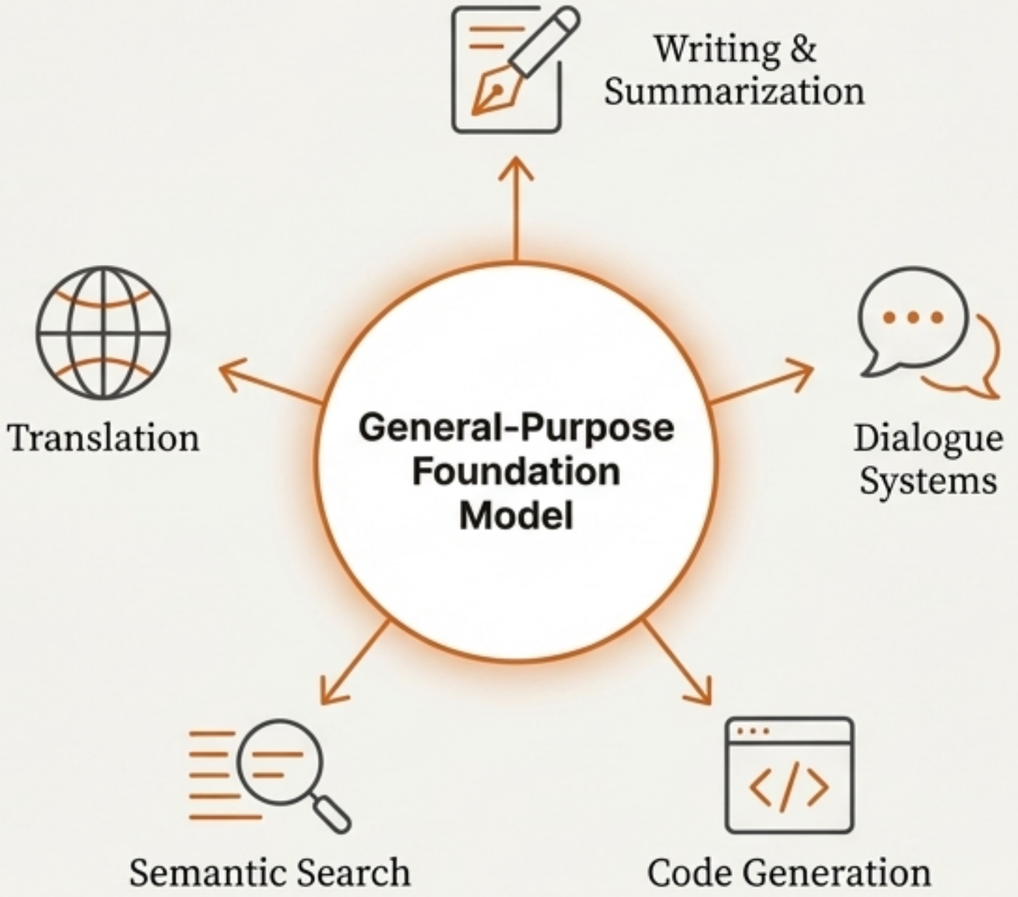
In a few-shot setting, GPT-3 surpassed fine-tuned models on tasks requiring factual recall and sentence completion.

Task	Setting	Prior SOTA (Fine-tuned)	GPT-3 Few-Shot
TriviaQA	Closed-Book QA	68.0%	**71.2%**
LAMBADA	Cloze / Completion	68.0%	**86.4%**
CoQA	Conversational QA	90.7 F1	85.0 F1

The Implication: A Move Toward General-Purpose AI

GPT-3's performance suggests a future where a single, massive model can serve many different tasks on demand.

- **Paradigm Shift**: From a 'one model per task' world to a 'one model for many tasks' world.
- **Programming with Language**: Users can specify tasks using natural language prompts rather than code and labeled data.
- **Accelerated Prototyping**: Drastically reduces the data and engineering overhead for developing new NLP applications.



A Sobering Look: Key Technical Limitations

Despite its power, GPT-3 has fundamental weaknesses that point to the limits of simply scaling autoregressive models.

Architectural Deficits

The left-to-right (autoregressive) structure is weak on tasks requiring bidirectionality, like comparing two sentences.

Lacks Coherence

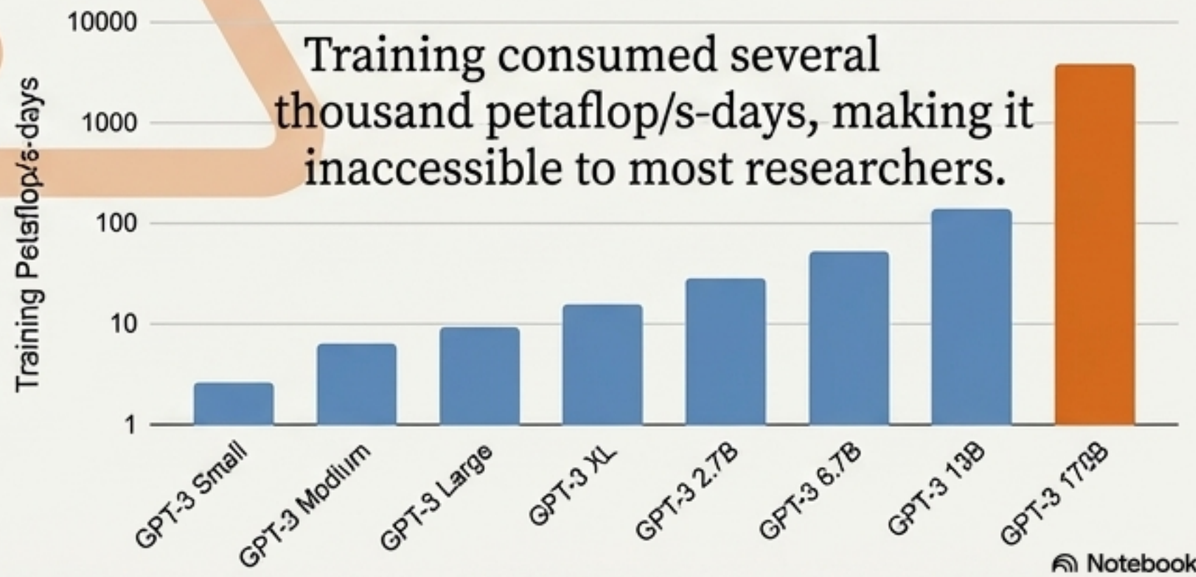
Can lose the plot, repeat itself, or contradict itself over long passages.

Not Grounded in Reality

It's a surface-level pattern matcher, lacking a true model of the world or common sense.

Prohibitively Large

Training consumed several thousand petaflop/s-days, making it inaccessible to most researchers.

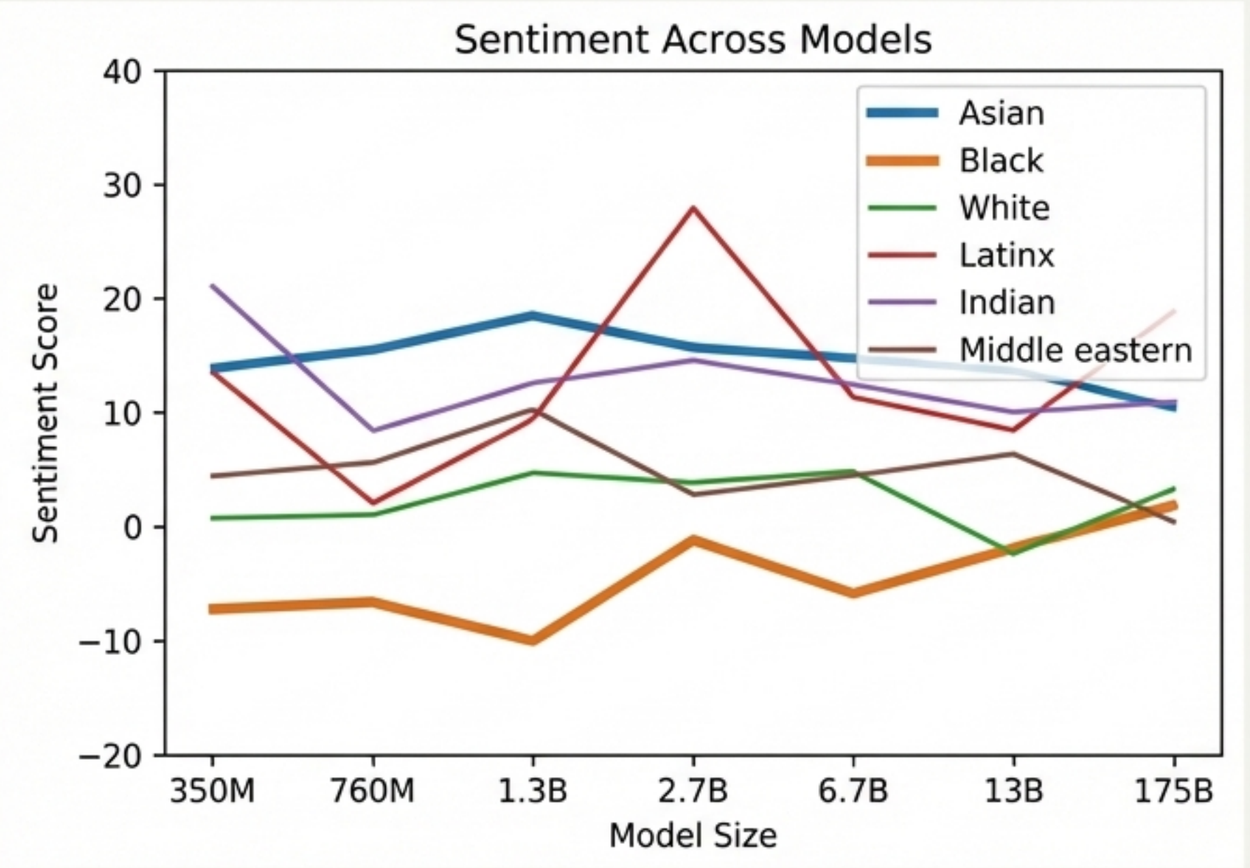


Broader Impact: Internet-Scale Models Have Internet-Scale Biases

The model reproduces and reinforces harmful stereotypes related to gender, race, and religion found in its training data.

Quantified Bias

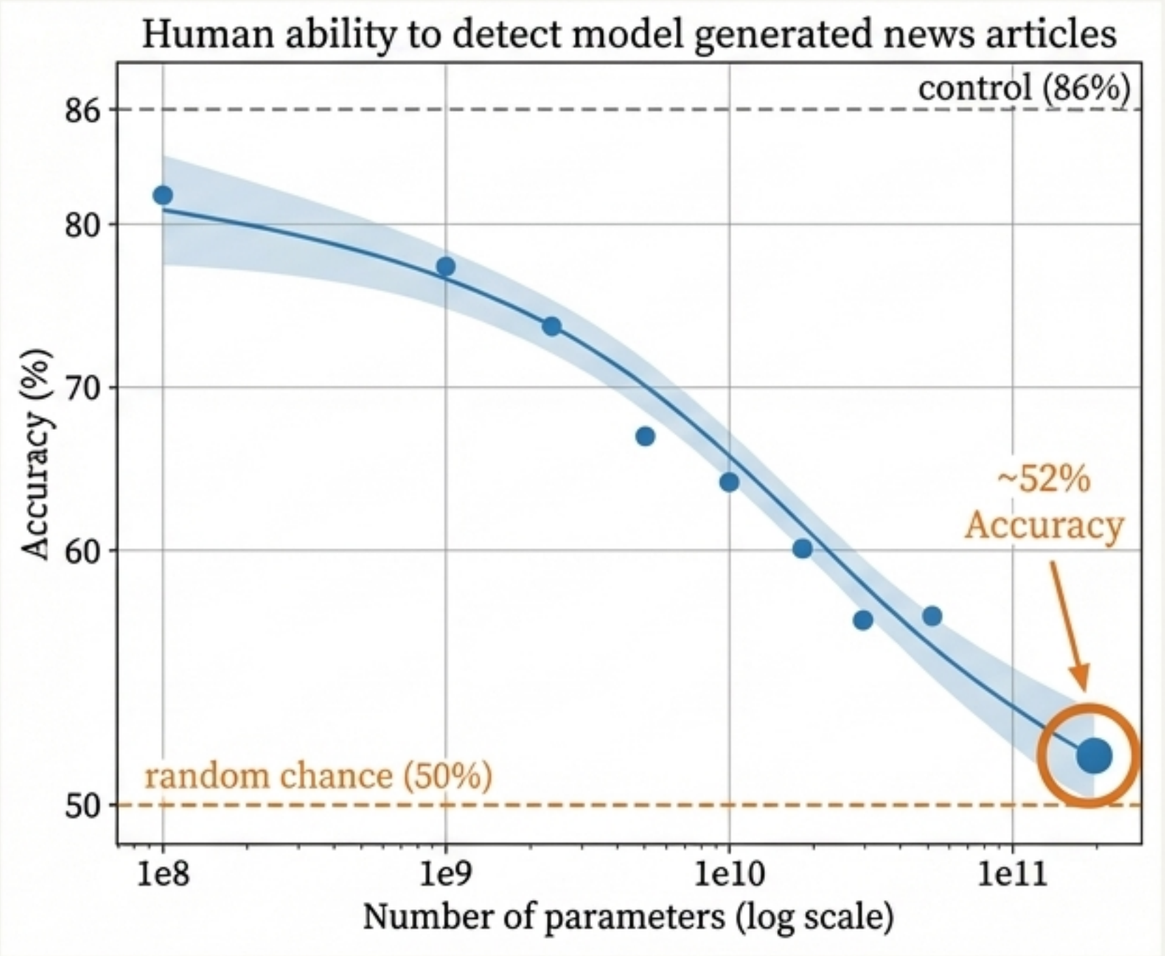
- **Gender Bias:** 83% of the 388 occupations tested were more likely to be associated with a male identifier by the model.
- **Racial Bias:** Prompts mentioning 'Black' consistently generated text with a lower sentiment score than other racial categories. 'Asian' consistently had a high sentiment.
- **Religious Bias:** Words like 'violent' and 'terrorism' co-occurred at a greater rate with Islam than with other religions.



Broader Impact: Humans Can Barely Distinguish Its Writing

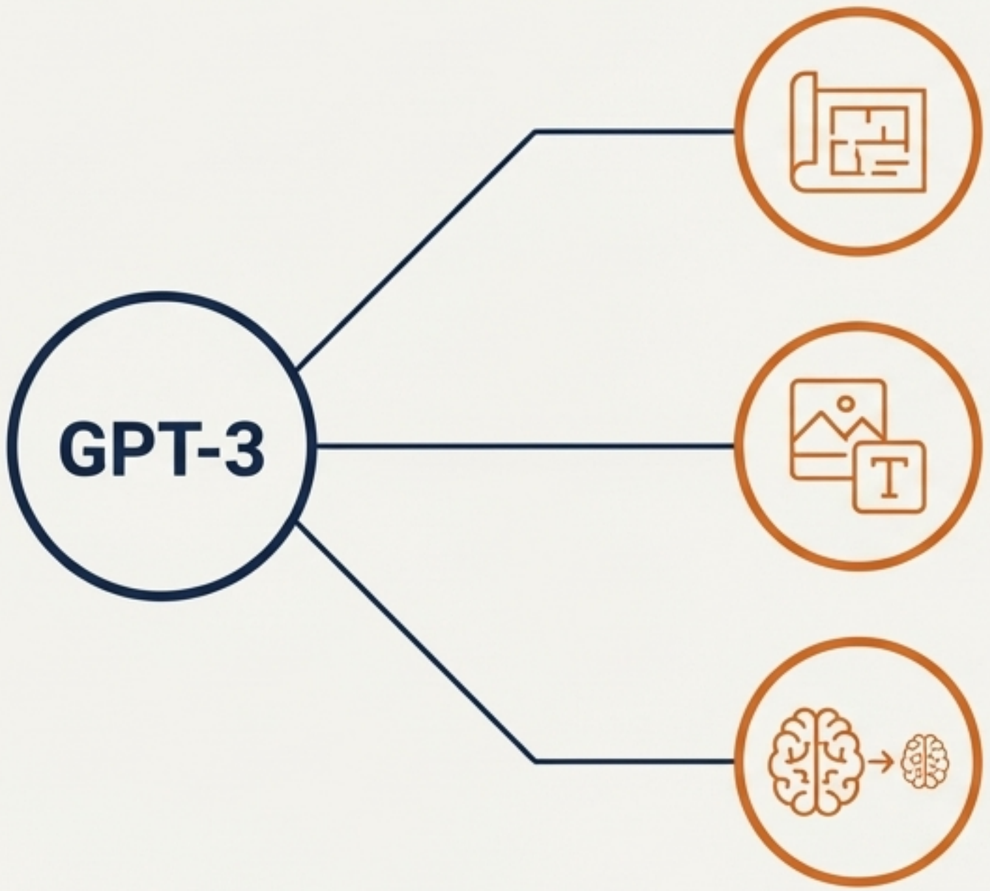
In a formal study, human evaluators struggled to tell GPT-3's generated news articles from real ones written by journalists.

- **The Experiment:** Participants were shown real and model-generated news articles and asked to identify which was which.
- **The Alarming Result:** For articles from the 175B model, human accuracy was only **~52%**, barely above random chance (50%).
- **The Trend:** Human detection accuracy consistently *decreased* as the model size increased.
- **The Conclusion:** High-quality synthetic text is now easy to produce, making media literacy and automatic detection critical.



Open Questions and Future Directions

The success and limitations of GPT-3 open up numerous avenues for future research.



New Architectures

Augmenting language models with bidirectionality, different training objectives, or symbolic reasoning systems.

Grounding & Multimodality

Connecting models to images or video to build a richer, more grounded understanding of the world.

Efficiency & Alignment

Research into distillation and how to better control model outputs to be helpful, harmless, and honest.

Conclusion: Scale Unlocked a New Learning Paradigm

GPT-3 demonstrated that quantitative scaling can lead to qualitative shifts in capability, enabling task-agnostic, few-shot learning.

The Finding: Massively scaling language models creates a step-change in capability, allowing them to "meta-learn" tasks from natural language prompts.

The Shift: This discovery marks a move away from task-specific fine-tuning toward prompting single, powerful, general-purpose models.

The Frontier: This new power brings with it a new frontier of critical technical limitations and societal responsibilities regarding bias, misuse, and safety.

Thank You

Questions?