Review of *Language Models are Few-Shot Learners.* (2020)

OpenAI GPT-3

Matthew Evans

April 30, 2025

Overview

Conventional language models typically require extensive task-specific training to achieve satisfactory performance, thereby limiting their adaptability. Building on their prior GPT-2[1] work, the authors[2] investigated whether simply increasing model scale could confer broad task proficiency with minimal or no additional training, leveraging only a handful of examples or natural language instructions at inference.

Approach

Prior Work

The most relevant predecessors to GPT-3 include GPT-2, which introduced large-scale autoregressive language modeling with zero-shot capabilities, and BERT[3], which popularized fine-tuning of large pretrained transformers on downstream tasks. The paper also builds on work exploring scaling laws for neural networks[4], and earlier efforts in meta-learning and in-context learning like GPT and GPT-2, as well as unsupervised multitask learning exemplified by models like T5[5] and XLNet[6]. These works laid the foundation for GPT-3's focus on scaling and task-agnostic inference via natural language prompts.

Novelty

GPT-3 was introduced as a 175 billion-parameter autoregressive language model—over ten times larger than its predecessors—demonstrating substantial performance gains across a wide range of NLP tasks. By leveraging this scale, the authors showed that a single model can perform zero-shot, one-shot, and few-shot learning using only natural language prompts, eliminating the need for task-specific fine-tuning.

Empirical results confirmed predictable improvements in model performance as size increased, validating practical scaling laws. Without any gradient updates, GPT-3 achieved state-of-the-art or near-state-of-the-art results on benchmarks such as LAMBADA and TriviaQA using in-context learning alone (i.e., no fine-tuning).

A thorough analysis of model behavior highlighted both its broad generalization capabilities and its limitations. While GPT-3 excels at many tasks, it struggles with symbolic reasoning and is sensitive to prompt phrasing. These observations informed a balanced evaluation of its practical strengths and weaknesses.

Finally, the authors addressed broader impacts, including ethical considerations around bias, misinformation, and the environmental cost of large-scale training. They framed GPT-3 as a powerful yet double-edged advancement in AI, underscoring the need for responsible deployment and further research.

Considerations

Strengths

- Achieves state-of-the-art results across NLP benchmarks without task-specific fine-tuning.
- Leverages 175 billion parameters for in-context meta-learning in zero-, one-, and few-shot settings.
- Demonstrates consistent performance gains with increasing model scale.

Weaknesses

- The 175 B parameter scale and extensive pretraining demand prohibitively high compute resources, limiting accessibility and rendering retraining (e.g., to correct invalid data) infeasible.
- Such compute intensity incurs significant energy consumption, raising environmental and infrastructure concerns.
- Performance is highly sensitive to prompt design, necessitating manual in-distribution prompt tuning to achieve optimal results.
- The model exhibits weaknesses in symbolic and multi-step reasoning, as demonstrated by arithmetic and word-manipulation benchmarks.

Measures of Success

The authors systematically present GPT-3's performance across diverse NLP benchmarks. Despite no task-specific fine-tuning, GPT-3 achieves or approaches state-of-the-art results on numerous tasks. A summary of these evaluations follows.

Language Modeling

Language modeling tasks evaluate a model's ability to predict the next word in a sequence, reflecting its grasp of syntax, semantics, and context.

Strengths GPT-3 (175B) achieves SOTA in the zero-shot setting on PTB[1] (20.5 perplexity) and surpasses prior SOTA on LAMBADA[2] in few-shot (86.4%), demonstrating strong contextual prediction and effective in-context learning.

Weaknesses Few-shot performance on HellaSwag[3] (79.3%) and StoryCloze[4] (87.7%) remains below fine-tuned SOTA, indicating limitations in fully matching specialized models on structured commonsense tasks.

Closed Book Question Answering

Closed book QA tasks assess a model's ability to answer factual questions from internal knowledge without external context or retrieval.

Strengths GPT-3 (175B) surpasses SOTA on TriviaQA[5] in the few-shot setting (71.2%), showing strong internalization of factual content and effective in-context adaptation.

Weaknesses Fails to match SOTA on Natural Questions[6] and WebQuestions[7]; few-shot performance (29.9%, 41.5%) suggests limits in fine-grained recall and domain adaptation.

Translation

Translation tasks evaluate a model's ability to convert text between languages while preserving meaning, grammar, and fluency.

[1] The Penn Treebank (PTB) is a dataset of syntactically annotated text used to benchmark language modeling performance.

[2] The LAMBADA (LAnguage Modeling Broadened to Account for Discourse Aspects) benchmark tests a model's ability to predict the final word of a passage requiring broad context understanding.

[3] HellaSwag is an adversarially constructed dataset requiring models to choose the most plausible continuation of a short story or instruction.

[4] The StoryCloze Test challenges models to select the correct ending to a four-sentence story from two options, evaluating narrative understanding.

[5] TriviaQA is a question answering dataset consisting of trivia-style factoid questions with answers grounded in evidence documents.

[6] Natural Questions (NQ) is a large-scale dataset from Google with real user questions and corresponding answers found in Wikipedia articles.

[7] WebQuestions is a dataset of natural language questions sourced from web search queries, paired with answers derived from Freebase.

Strengths GPT-3 (175B) in the few-shot setting outperforms prior unsupervised (Neural Machine Translation) methods and approaches supervised SOTA on several X→En tasks (e.g., Ro→En: 39.5 BLEU[8]).

Weaknesses Performance on En→X directions, especially En→Ro (21.0 BLEU), remains far below supervised SOTA, reflecting training data imbalance and tokenizer bias toward English.

Winograd-Style Tasks

Winograd-style tasks test a model's ability to resolve pronoun references using contextual and commonsense reasoning.

Strengths GPT-3 (175B) achieves near-SOTA on WSC273[9] even in the zero-shot setting (88.3%), with a slight peak in one-shot (89.7%), demonstrating strong pretrained coreference capabilities.

Weaknesses Few-shot performance on the adversarial Winogrande task (77.7%) falls short of SOTA (84.6%), indicating limited robustness to more challenging or out-of-distribution examples.

Common Sense Reasoning

Common sense reasoning tasks evaluate a model's ability to apply everyday physical and causal knowledge to answer questions or select plausible outcomes.

Strengths GPT-3 (175B) achieves SOTA on PIQA[10] in all settings (e.g., 82.8% few-shot), outperforming fine-tuned baselines, and shows consistent scaling and few-shot gains on OpenBookQA[11].

Weaknesses Fails to match SOTA on ARC-Challenge (51.5%) and OpenBookQA (65.4%), with shallow improvements in few-shot settings and persistent gaps in multi-hop or science-based reasoning.

Reading Comprehension

Reading comprehension tasks assess a model's ability to extract or infer answers from passages of text, often requiring span selection or generative responses.

Strengths GPT-3 (175B) performs competitively on CoQA[12] in the few-shot setting (85.0 F1), approaching human-level and fine-tuned SOTA, and shows strong gains from zero- to few-shot on SQuADv2.

Weaknesses Few-shot performance on DROP[13] (36.5 F1), QuAC[14] (44.3 F1), and RACE[15] ($<$ 60% accuracy) remains far below SOTA, indicating limited capacity for discrete reasoning, dialog structure, and multi-step inference.

SuperGLUE

SuperGLUE is a benchmark suite of challenging NLP tasks designed to test a model's reasoning, inference, and understanding across multiple formats.

[8] The Bilingual Evaluation Understudy (BLEU) score measures the quality of machine-translated text by comparing it to one or more reference translations.

[9] WSC273 is the 273-example version of the Winograd Schema Challenge, designed to test pronoun resolution requiring commonsense reasoning.

[10] The Physical Interaction Question Answering (PIQA) benchmark evaluates a model's ability to reason about everyday physical situations.

[11] OpenBookQA is a multiple-choice question answering dataset requiring models to combine science facts with broad commonsense knowledge.

[12] The Conversational Question Answering (CoQA) dataset involves multi-turn QA where each question depends on previous context.

[13] The Discrete Reasoning Over Paragraphs (DROP) benchmark tests a model's ability to perform arithmetic and reasoning over paragraphs.

[14] The Question Answering in Context (QuAC) dataset contains information-seeking dialogues where the model must answer questions based on context.

[15] The Reading Comprehension from Examinations (RACE) dataset consists of English exam questions for middle and high school students.

Strengths GPT-3 (175B) in the few-shot setting matches or exceeds fine-tuned BERT-Large on four out of eight tasks, and approaches SOTA on COPA[16] and ReCoRD[17], demonstrating effective broad generalization.

Weaknesses Few-shot performance is weak on WiC[18] (49.4%) and MultiRC[19] (30.5% accuracy), with consistent underperformance on sentence-pair tasks requiring fine-grained semantic comparison.

NLI

Natural Language Inference (NLI) tasks require a model to determine whether a hypothesis logically follows from, contradicts, or is neutral with respect to a premise.

Strengths GPT-3 (175B) shows modest gains on RTE[20] in the few-shot setting, approaching the performance of a fine-tuned BERT-Large baseline.

Weaknesses Few-shot performance on the adversarial ANLI[21] benchmark remains well below SOTA, with smaller models performing near random, indicating limited robustness to difficult inference cases.

Synthetic and Qualitative Tasks

These tasks probe GPT-3's ability to perform on-the-fly reasoning, pattern recognition, and generative tasks not directly seen during training.

Strengths GPT-3 (175B) demonstrates strong few-shot performance on arithmetic, word manipulation, analogy solving, and naturalistic text generation, often producing human-like outputs without fine-tuning.

Weaknesses Fails on more complex or compositional tasks (e.g., 5-digit multiplication), with performance sharply degrading as problem complexity increases, revealing limits in systematic generalization and symbolic reasoning.

Impact

The GPT-3 paper had a profound impact on future work by demonstrating that large-scale language models can perform a wide variety of tasks using only natural language prompts, eliminating the need for task-specific fine-tuning in many cases. It established few-shot, one-shot, and zero-shot learning as viable paradigms for evaluating language models and shifted the research focus toward scaling laws, in-context learning, and prompt engineering. GPT-3's success inspired the development of even larger and more capable models, accelerated the deployment of foundation models across domains, and raised important questions about data contamination, evaluation practices, and ethical considerations in large-scale AI.

References

[1] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019.

[16] The Choice of Plausible Alternatives (COPA) task requires selecting the more plausible cause or effect of a given premise.

[17] The Reading Comprehension with Commonsense Reasoning Dataset (ReCoRD) evaluates a model's ability to fill in masked entities in passages based on commonsense inference.

[18] The Word-in-Context (WiC) task asks whether a target word has the same meaning in two different sentence contexts.

[19] The Multi-Sentence Reading Comprehension (MultiRC) dataset is a QA benchmark where each question can have multiple correct answers, requiring justification from multiple sentences.

[20] The Recognizing Textual Entailment (RTE) dataset tests whether a hypothesis can be inferred from a given premise.

[21] The Adversarial Natural Language Inference (ANLI) benchmark is a series of increasingly difficult NLI tasks constructed to challenge language models' inference capabilities.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[4] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020.

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[6] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019.