Yes prompts work, no Don’t ask why!

-By nitkan

Why prompts?

- Prompts work because they give model context about what we expect it to do.
- Expect model to take into consideration the *semantics* of the given prompt to better solve the task with just few data examples

- Example: Using "Summarize the passage: [Passage] " is better than using just "[Passage]" as input when we can just afford few datapoints

GPT: Language models are few-shot learners

Discrete Prompts

They are essentially words from the model's vocabulary used for task instruction

Discrete Prompts

Started with GPT from OpenAI.

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

Jared 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 M. Ziegler** **Jeffrey Wu** **Clemens Winter**

Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**

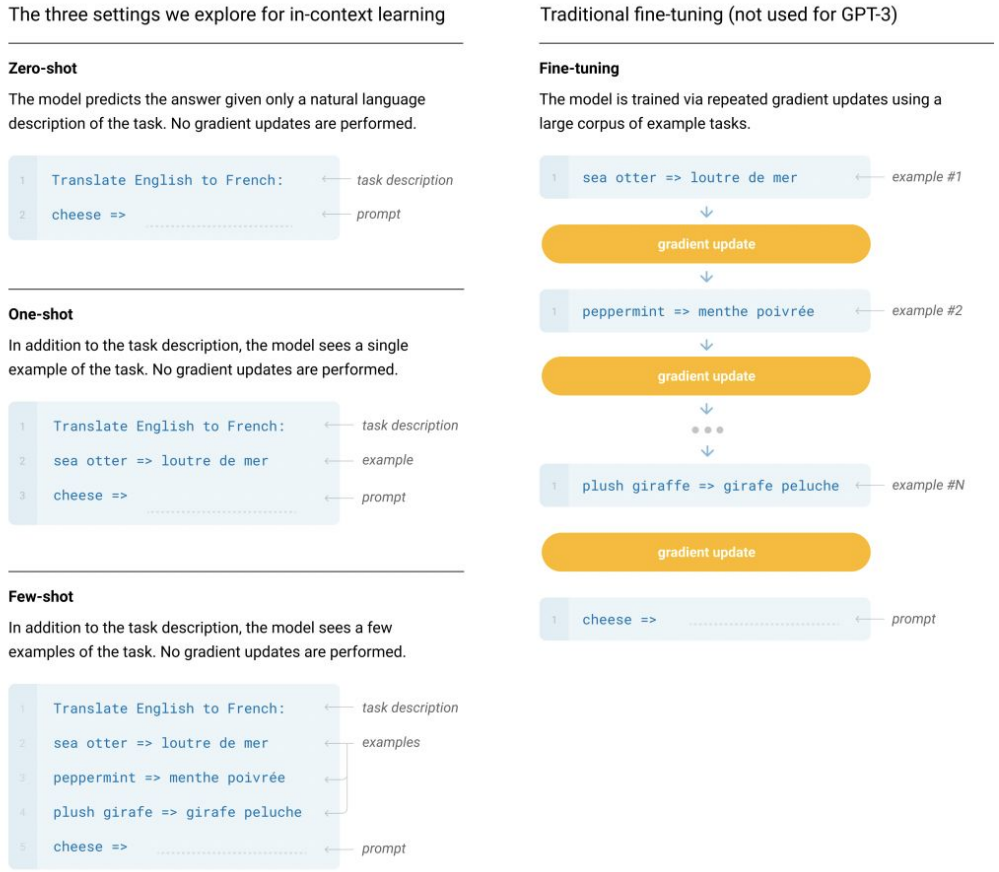
Benjamin Chess **Jack Clark** **Christopher Berner**

Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

OpenAI

Abstract

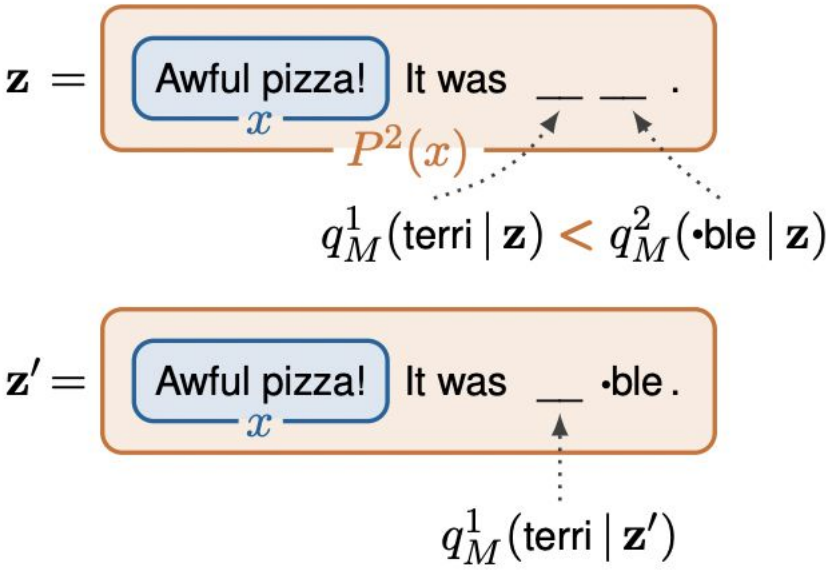
Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.



[GPT: Language models are few-shot learners]

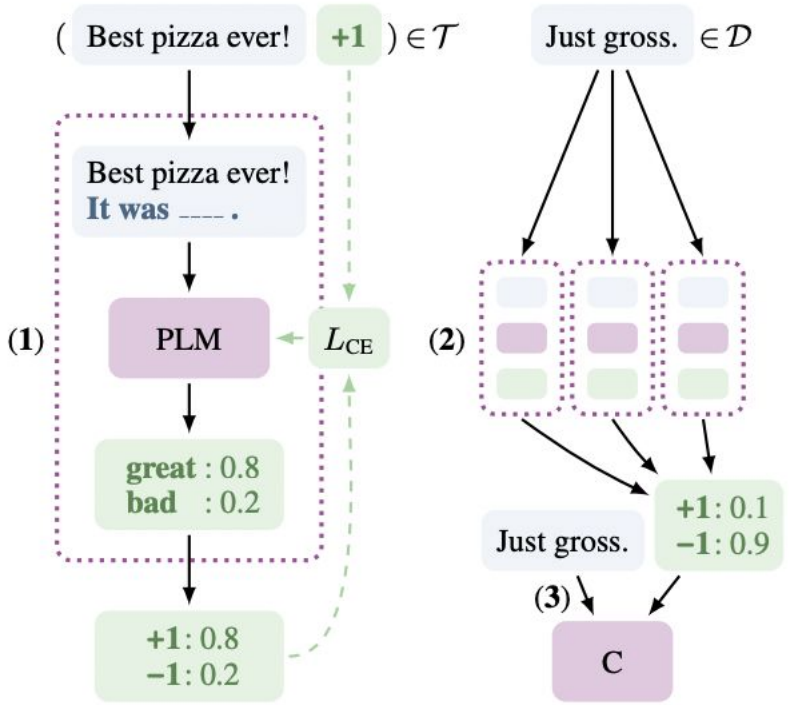
Discrete Prompts

It's not just the size that matters!



[Schick and Schütze, 2021b;]

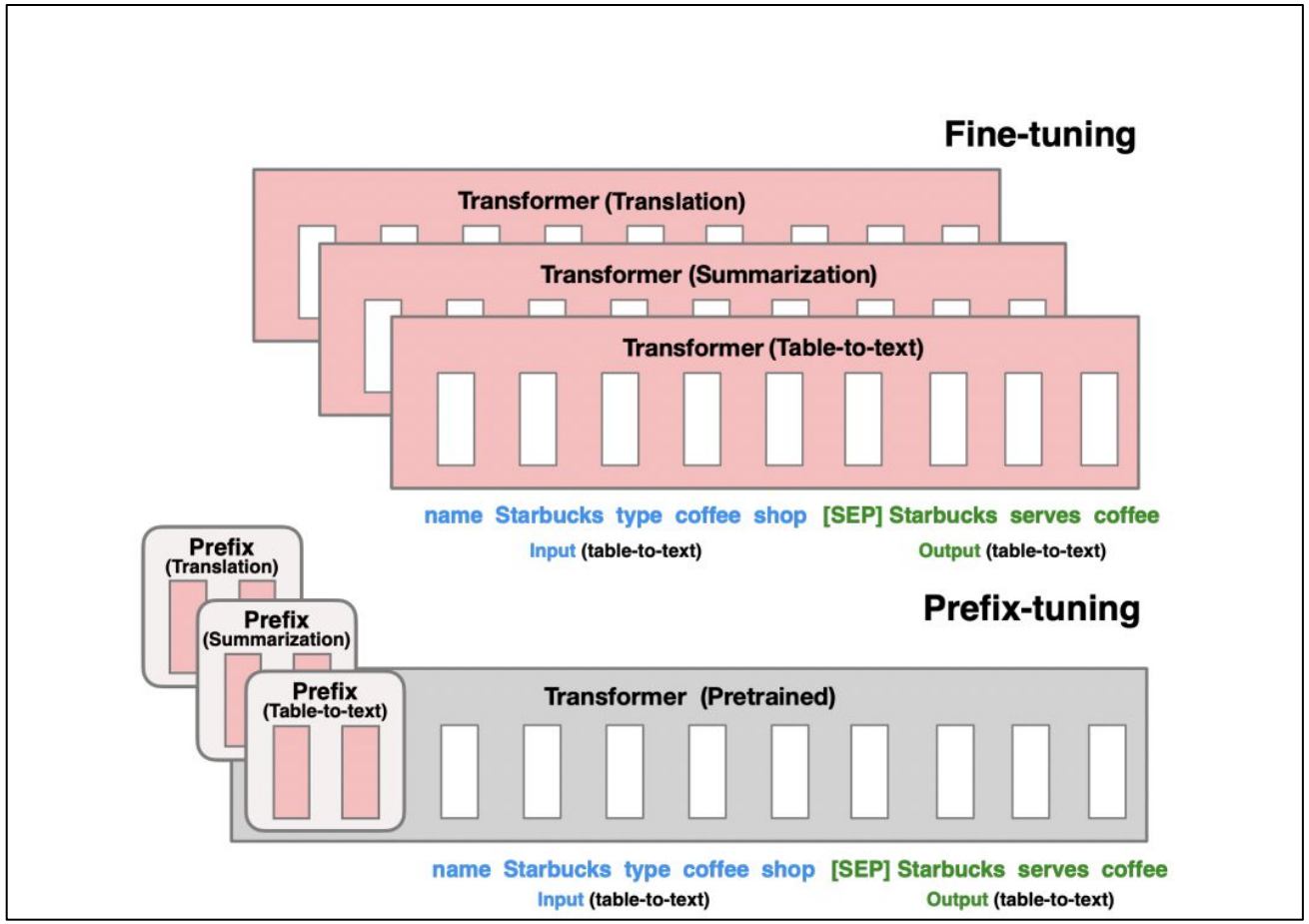
Patter Exploiting Training (PETs)



[Schick and Schütze, 2021a]

Continuous Prompts

They are randomly (almost always) initialised prepended tokens whose embeddings are trained with the model frozen. Anologous to searching for a suitable prompt for the task.



Li and Lang, 2021

Prefix Tuning

- Prefix-tuning, a lightweight alternative to fine-tuning for natural language generation (NLG) tasks, inspired by prompting.
- The task specific prefix tokens are randomly initialised
- Only the prefix tokens of each layer are finetuned instead of the whole model
- This way the transformer is frozen and just the task specific prefix contains knowledge of the task

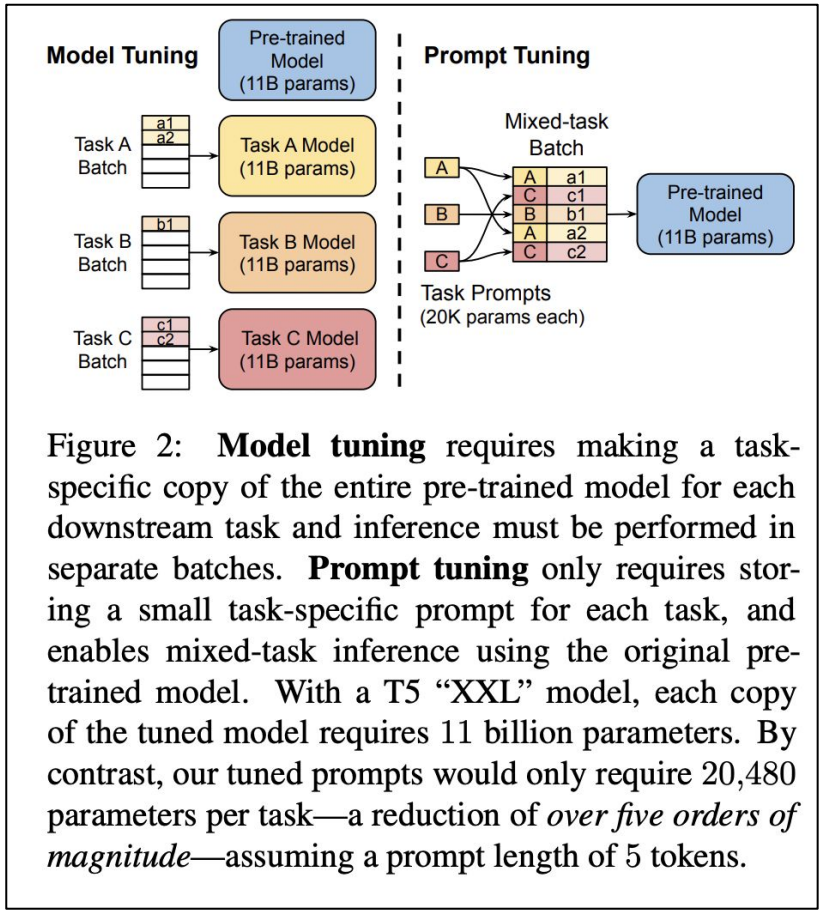
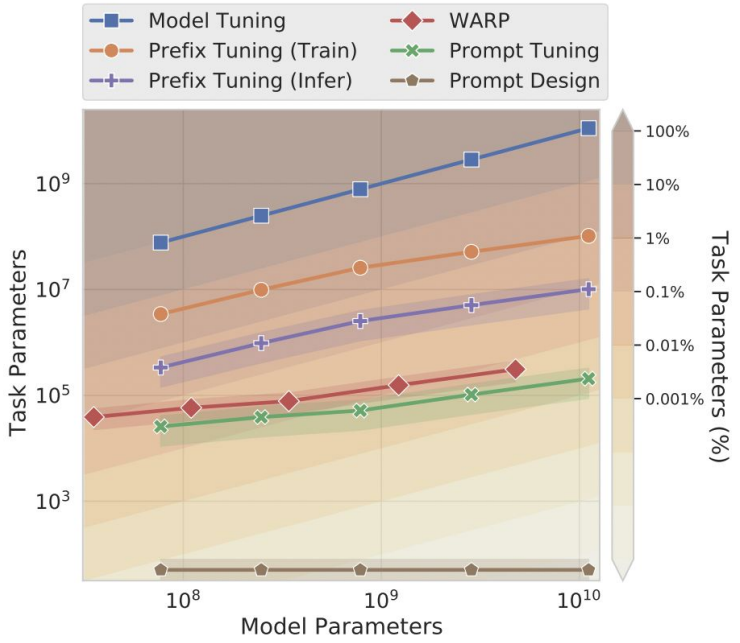


Figure 2: **Model tuning** requires making a task-specific copy of the entire pre-trained model for each downstream task and inference must be performed in separate batches. **Prompt tuning** only requires storing a small task-specific prompt for each task, and enables mixed-task inference using the original pre-trained model. With a T5 "XXL" model, each copy of the tuned model requires 11 billion parameters. By contrast, our tuned prompts would only require 20,480 parameters per task—a reduction of *over five orders of magnitude*—assuming a prompt length of 5 tokens.

Lester et al, 2021

Prompt Tuning

- Prompt tuning is essence a simplification of Prefix Tuning (Li and Liang 2021).
- They show that just fine tuning the embedding matrix of the prepended prompts suffices for transferring task specific knowledge.



Do they work as expected?

Conclusion from NAACL 2022: No!

Paper link
NAACL 2022

Do Prompt-Based Models Really Understand the Meaning of Their Prompts?

Albert Webson[1,2] and Ellie Pavlick[1]

{albert_webson, ellie_pavlick}@brown.edu
[1]Department of Computer Science, Brown University
[2]Department of Philosophy, Brown University

TL: DR.

Prompt-Based models work just as fast with many prompts that are intentionally irrelevant or even pathologically misleading as they do with instructively "good" prompts. Question is does the model pay attention to the semantics of the input prompt at all while learning the task?

Category	Examples
instructive	{prem} Are we justified in saying that "{hypo}"? Suppose {prem} Can we infer that "{hypo}"?
misleading-moderate	{prem} Can that be paraphrased as: "{hypo}"? {prem} Are there lots of similar words in "{hypo}"?
misleading-extreme	{prem} is the sentiment positive? {hypo} {prem} is this a sports news? {hypo}
irrelevant	{prem} If bonito flakes boil more than a few seconds the stock becomes too strong. "{hypo}"?
null	{premise} {hypothesis} {hypothesis} {premise}

Paper of Interest

NAACL 2022

PROMPT WAYWARDNESS: The Curious Case of Discretized Interpretation of Continuous Prompts

Daniel Khashabi[‡] **Xinxi Lyu**[†] **Sewon Min**[†]
Lianhui Qin[†] **Kyle Richardson**[‡] **Sean Welleck**[†‡]
Hannaneh Hajishirzi[†‡] **Tushar Khot**[‡] **Ashish Sabharwal**[‡] **Sameer Singh**[‡♡] **Yejin Choi**[†‡]

[†]University of Washington [‡]Allen Institute for AI [♡]University of California-Irvine

Paper link
NAACL 2022

Switch to paper