Machine Learning for Natural Language Processing

Modern Language Models

UNIVERSITY OF
GOTHENBURG

CHALMERS

Richard Johansson

richajo@chalmers.se

2022: Language models saw a popular breakthrough

The technologies already existed. What has changed now?

2022: Language models saw a popular breakthrough

The technologies already existed. What has changed now?

This was a HCI breakthrough: The models have become user-friendly, easy to access, and developed to do many things

What is the key difference?

GPT-1/2/3, OPT, ...

Write a Python one-liner that finds the longest string in a list of strings.

It should be as short as possible and use list comprehensions only, not for loops, maps. Write a Python one-liner that finds the maximal integer stored as a value in a dict.

What is the key difference?

GPT-1/2/3, OPT, ...

Write a Python one-liner that finds the longest string in a list of strings.

It should be as short as possible and use list comprehensions only, not for loops, maps.
Write a Python one-liner that finds the maximal integer stored as a value in a dict.

InstructGPT, ChatGPT, PaLM, ...

Write a Python one-liner that finds the longest string in a list of strings.

```
max(the_list, key=len)
```

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang

March 2023

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4, was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang

March 2023

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4, was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

What do we use recent language models for?

Not just chatbots!

RI

Beskriv Chalmers värdegrund.

Chalmers värdegrund är en samling principer och värderingar som styr skolans verksamhet och dess relationer med omvärlden. Här är en kort beskrivning av Chalmers värdegrund:

1. Kunskap och vetenskap: Chalmers är dedikerat till att främja kunskap och vetenskap genom forskning och utbildning. Skolan arbetar för att producera vetenskapligt kunniga och samhällsmedvetna individer som kan bidra till samhällets utveckling.
2. Lärande och personlig utveckling: Chalmers strävar efter att skapa en miljö som främjar lärande och personlig utveckling. Skolan värnar om individens potential och strävar efter att ge studenter de verktyg och resurser de behöver för att växa och utvecklas.
3. Hållbarhet: Chalmers har ett starkt fokus på hållbarhet och arbetar för att skapa en hållbar framtid genom forskning och utbildning. Skolan strävar efter att utbilda studenter som kan bidra till en hållbar utveckling och som kan arbeta för att lösa de utmaningar som vår planet står inför.
4. Samhällsengagemang: Chalmers är engagerat i samhället och arbetar för att bidra till samhällets utveckling genom forskning och utbildning. Skolan strävar efter att vara en del av samhället och att samarbeta med andra organisationer och företag för att skapa en bättre framtid.
5. Internationell prägel: Chalmers är en internationell skola med en mångfald av studenter och anställda från olika delar av världen. Skolan värdesätter och främjar den internationella erfarenheten och arbetar för att ge studenter möjligheter att få en global utbildning och perspektiv.

Challenges and Applications of Large Language Models

Jean Kaddour[α, †, *], Joshua Harris[β, *], Maximilian Mozes[α],
Herbie Bradley[γ, δ, ϵ], Roberta Raileanu[ζ], and Robert McHardy[η, *]

[α]University College London [β]UK Health Security Agency [γ]EleutherAI
[δ]University of Cambridge [ϵ]Stability AI [ζ]Meta AI Research [η]InstaDeep

Abstract

Large Language Models (LLMs) went from non-existent to ubiquitous in the machine learning discourse within a few years. Due to the fast pace of the field, it is difficult to identify the remaining challenges and already fruitful application areas. In this paper, we aim to establish a systematic set of open problems and application successes so that ML researchers can comprehend the field's current state more quickly and become productive.

Contents

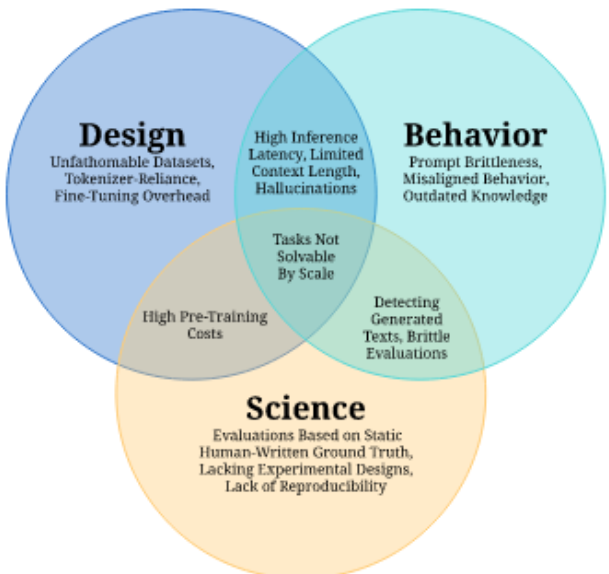


Figure 1: **Overview of LLM Challenges.** *Design*ing LLMs relates to decisions taken before deployment. *Behavior*ial challenges occur during deployment. *Science* challenges hinder academic progress.

Language models are becoming components

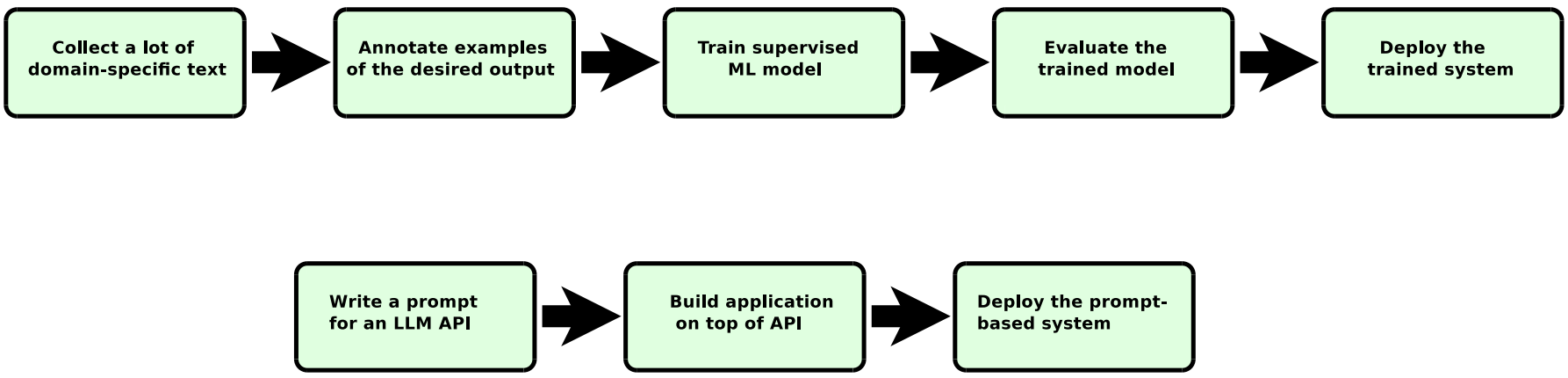
How that changes engineering practices is an open question

As of 2023, natural language is used to control the models and to give inputs and outputs

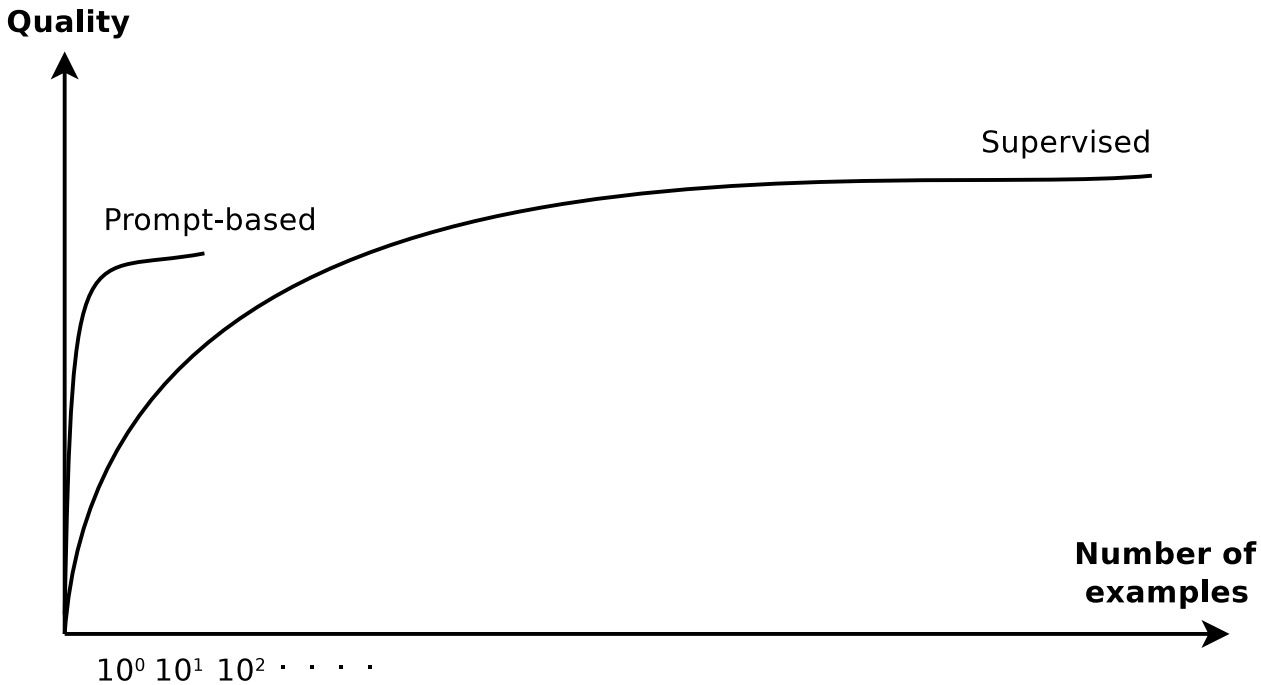
Workflows


Collect a lot of domain-specific text
Annotate examples of the desired output
Train supervised ML model
Evaluate the trained model
Deploy the trained system

Workflows


Collect a lot of domain-specific text
Annotate examples of the desired output
Train supervised ML model
Evaluate the trained model
Deploy the trained system
Write a prompt for an LLM API
Build application on top of API
Deploy the prompt-based system

How much can we get for our data points? (General idea)



Main significance is in ease of prototyping?

PromptMaker: Prompt-based Prototyping with Large Language Models

Ellen Jiang*
Google Research, PAIR team
USA
ellenj@google.com

Kristen Olson*
Google Research, PAIR team
USA
kolson@google.com

Edwin Toh*
Google Research, PAIR team
USA
edwintoh@google.com

Alejandra Molina
Google Research, PAIR team
USA
alemolinata@google.com

Aaron Donsbach
Google Research, PAIR team
USA
donsbach@google.com

Michael Terry
Google Research, PAIR team
USA
michaelterry@google.com

Carrie J. Cai
Google Research, PAIR team
USA
cjcai@google.com

ABSTRACT

Prototyping is notoriously difficult to do with machine learning (ML), but recent advances in large language models may lower the barriers to people prototyping with ML, through the use of natural language prompts. This case study reports on the real-world experiences of industry professionals (e.g. designers, program managers, front-end developers) prototyping new ML-powered feature ideas via **prompt-based prototyping**. Through interviews with eleven practitioners during a three-week sprint and a workshop, we find that prompt-based prototyping reduced barriers of access by substantially broadening who can prototype with ML, sped up the prototyping process, and grounded communication between collaborators. Yet, it also introduced new challenges, such as the need to reverse-engineer prompt designs, source example data, and debug and evaluate prompt effectiveness. Taken together, this case study provides important implications that lay the groundwork toward a new future of prototyping with ML.

learning [10, 16, 17]. Prototyping with AI can be uniquely challenging, due to the difficulty of: determining a priori what an AI can or cannot do, finding skilled AI technical collaborators, and envisioning AI uses that do not yet exist [17]. Recently, advances in Large Language Models (LLM) like GPT-3 [3] have lowered the barrier to authoring new machine learning functionality on-the-fly by allowing users to feed natural language prompts to an LLM, a practice known as "prompt programming" [1, 3]. For example, given a general purpose LLM, a user could customize it to act like a specialized English-to-French translation engine, by giving the LLM a natural language prompt containing pairs of English and French examples: "English: how are you? French: comment allez-vous? English: goodbye! French: au revoir! English: hello! French: ". Given this prompt, the LLM is likely to output the French translation: "bonjour!"

While the translation example above serves as a toy example, in practice there are numerous *types* of ML functionality a designer may want to prototype, ranging from innovative new application ideas (e.g., given the weather, generate clothing advice; given fo-

Towards Agile Text Classifiers for Everyone

Maximilian Mozes[1,2*†] **Jessica Hoffmann**[1*] **Katrin Tomanek**[1] **Muhamed Kouate**[1†]
Nithum Thain[1] **Ann Yuan**[1] **Tolga Bolukbasi**[1] **Lucas Dixon**[1]
[1]Google Research
[2]University College London
{jhoffmann,katrintomanek,kouate,nthain,annyuan,tolgab,ldixon}@google.com
maximilian.mozes@ucl.ac.uk

Abstract

Text-based safety classifiers are widely used for content moderation and increasingly to tune generative language model behavior—a topic of growing concern for the safety of digital assistants and chatbots. However, different policies require different classifiers, and safety policies themselves improve from iteration and adaptation. This paper introduces and evaluates methods for agile text classification, whereby classifiers are trained using small, targeted datasets that can be quickly developed for a particular policy. Experimenting with 7 datasets from three safety-related domains, comprising 15 annotation schemes, led to our key finding: prompt-tuning large language models, like PaLM 62B, with a labeled dataset of as few as 80 examples can achieve state-of-the-art performance. We argue that this

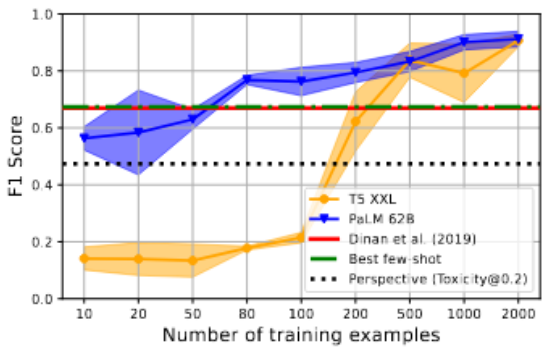


Figure 1: Prompt-tuning PaLM 62B and T5 XXL with as few as 80 and 500 examples, respectively, outperforms both in-context learning (12-shot) on PaLM 62B and a BERT model fine-tuned on 24,000 training examples as reported in Dinan et al. (2019) for the ParlAI Single Adversarial dataset.

Ways to control the behavior of LLMs

Prompting: all information about the task (instructions, examples) are in the prompt

Fine-tuning: we make some (small) adaptation of the model parameters, based on a dataset

Cost models of API-based LLMs

Capabilities of the underlying model

Number of tokens in prompts

Number of tokens in generated output

Whether or not the model is fine-tuned

“Hallucinations” in language model outputs

The output is plain false, or somewhat false

OR: The output is unrelated to the input

RI Please list three important papers by Devdatt Dubhashi.

Devdatt Dubhashi is a prominent computer scientist who has made significant contributions to the fields of algorithms, data structures, and machine learning. Here are three important papers authored or co-authored by him:

1. "A Randomized Algorithm for Minimum Cuts," co-authored with Karger, Klein, and Tarjan, which was published in the Journal of the ACM in 1995. This paper introduced a randomized algorithm for finding minimum cuts in undirected graphs that has an expected running time of O(n^2 log n), where n is the number of vertices in the graph. This algorithm is widely used in practice and has become a cornerstone of modern algorithms research.
2. "Concentration of Measure for the Analysis of Randomized Algorithms," co-authored with Panconesi, which was published in the Journal of Computer and System Sciences in 1997. This paper introduced a powerful tool for the analysis of randomized algorithms called the concentration of measure phenomenon. This tool has since been used to analyze a wide range of randomized algorithms and has become a key component of modern probabilistic analysis techniques.
3. "Machine Learning and Optimization: A Unifying View," co-authored with Bengtsson, which was published in the Journal of Machine Learning Research in 2005. This paper introduced a unified framework for understanding the relationship between machine learning and optimization, two important fields in computer science. This framework has since been used to develop new machine learning algorithms and to analyze existing ones, and has helped to bridge the gap between machine learning theory and practice.

Large Language Models Struggle to Learn Long-Tail Knowledge

Nikhil Kandpal
UNC Chapel Hill
nkandpa2@cs.unc.edu

Haikang Deng
UNC Chapel Hill
haikang@live.unc.edu

Adam Roberts
Google Research
adarob@google.com

Eric Wallace
UC Berkeley
ericwallace@berkeley.edu

Colin Raffel
UNC Chapel Hill
craffel@gmail.com

Abstract

The internet contains a wealth of knowledge—from the birthdays of historical figures to tutorials on how to code—all of which may be learned by language models. However, there is a huge variability in the number of times a given piece of information appears on the web. In this paper, we study the relationship between the knowledge memorized by large language models and the information in their pre-training datasets. In particular, we show that a language model's ability to answer a fact-based question relates to how many documents associated with that question were seen during pre-training. We identify these relevant documents by entity linking pre-training datasets and counting documents that contain the same entities as a given question-answer pair. Our results demonstrate strong correla-

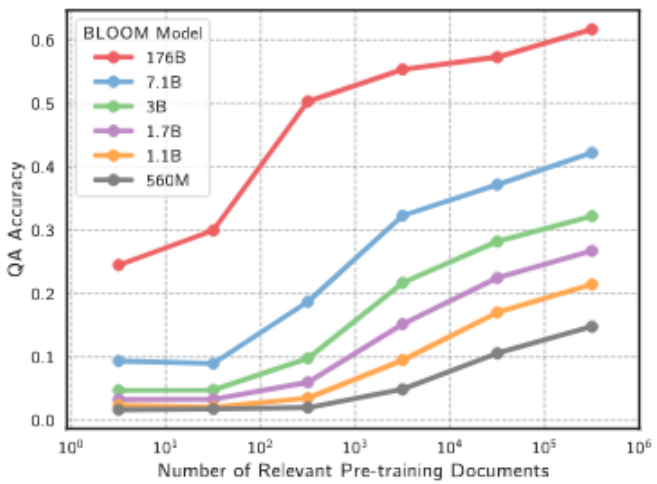


Figure 1: Language models struggle to capture the long-tail of information on the web. Above, we plot accuracy for the BLOOM model family on TriviaQA as a function of how many documents in the model's pre-training data are relevant to each question.

(As of 2023): Poor metalinguistic capabilities

"How many occurrences of the word 'the' are there in this sentence?"

"Please write the following sentence in the opposite order"

"Please write a text about Gothenburg in exactly 20 words."

"Write a sentence about Chalmers, where the last word is 'bananas.' "

Next: How to train LMs to follow instructions

References

S. Bubeck, V. Chandrasekaran, and R. Eldan et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712.

E. Jiang, K. Olson, E. Toh, A. Molina, A. Donsbach, M. Terry, and C. J. Cai. 2022. PromptMaker: Prompt-based prototyping with large language models. In *CHI Conference on Human Factors in Computing Systems*.

J. Kaddour, J. Harris, and M. Mozes et al. 2023. Challenges and applications of large language models. arXiv:2307.10169.

N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *ICML*.

M. Mozes, J. Hoffmann, and K. Tomanek et al. 2023. Towards agile text classifiers for everyone. arXiv:2302.06541.