GenAI – the new kind of software

Pratyush Kumar

22nd November 2023 / Accel Event

Broad topics

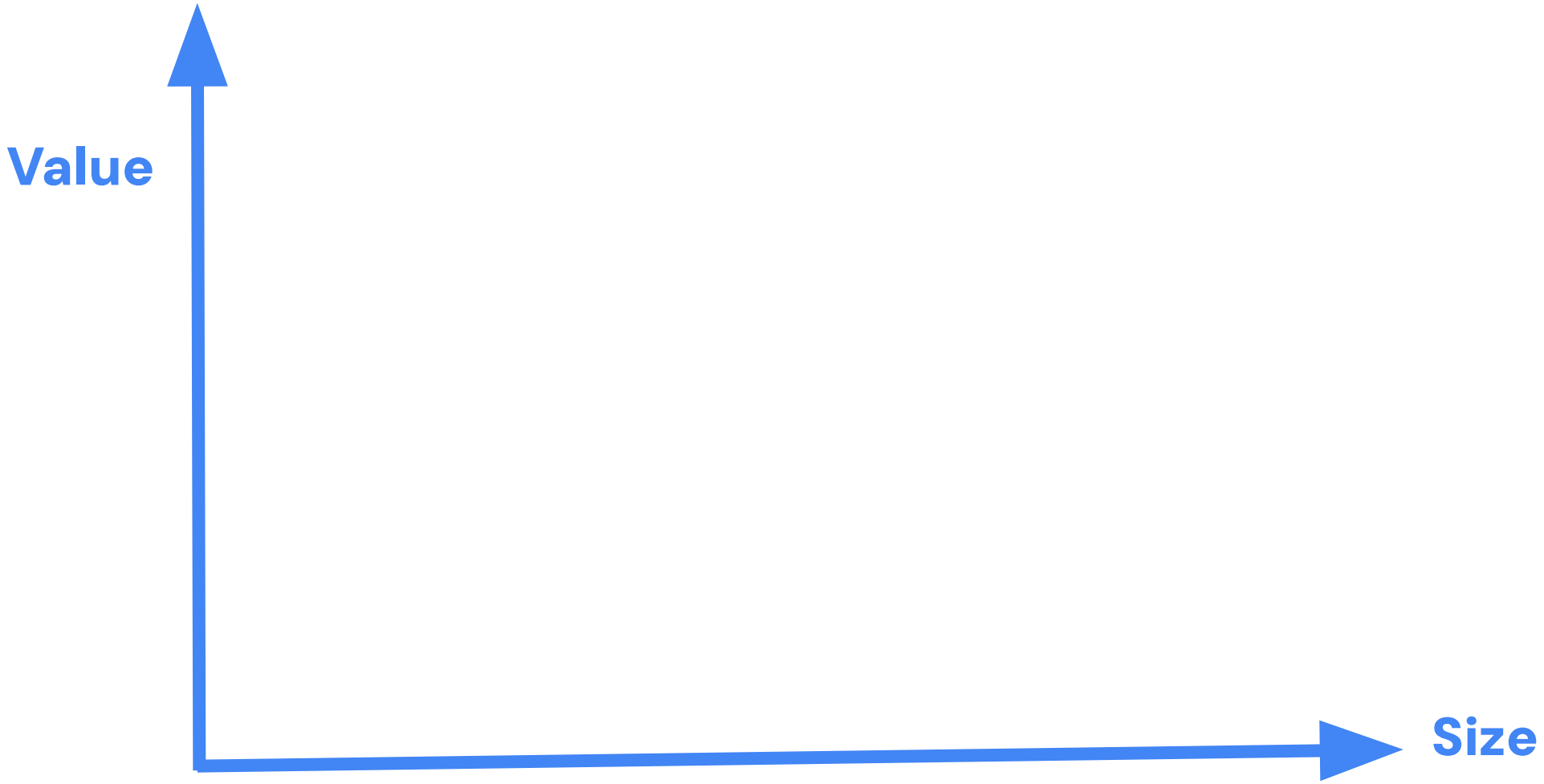
Useful to think of GenAI as a new type of software

Why is this type of software extremely generalisable

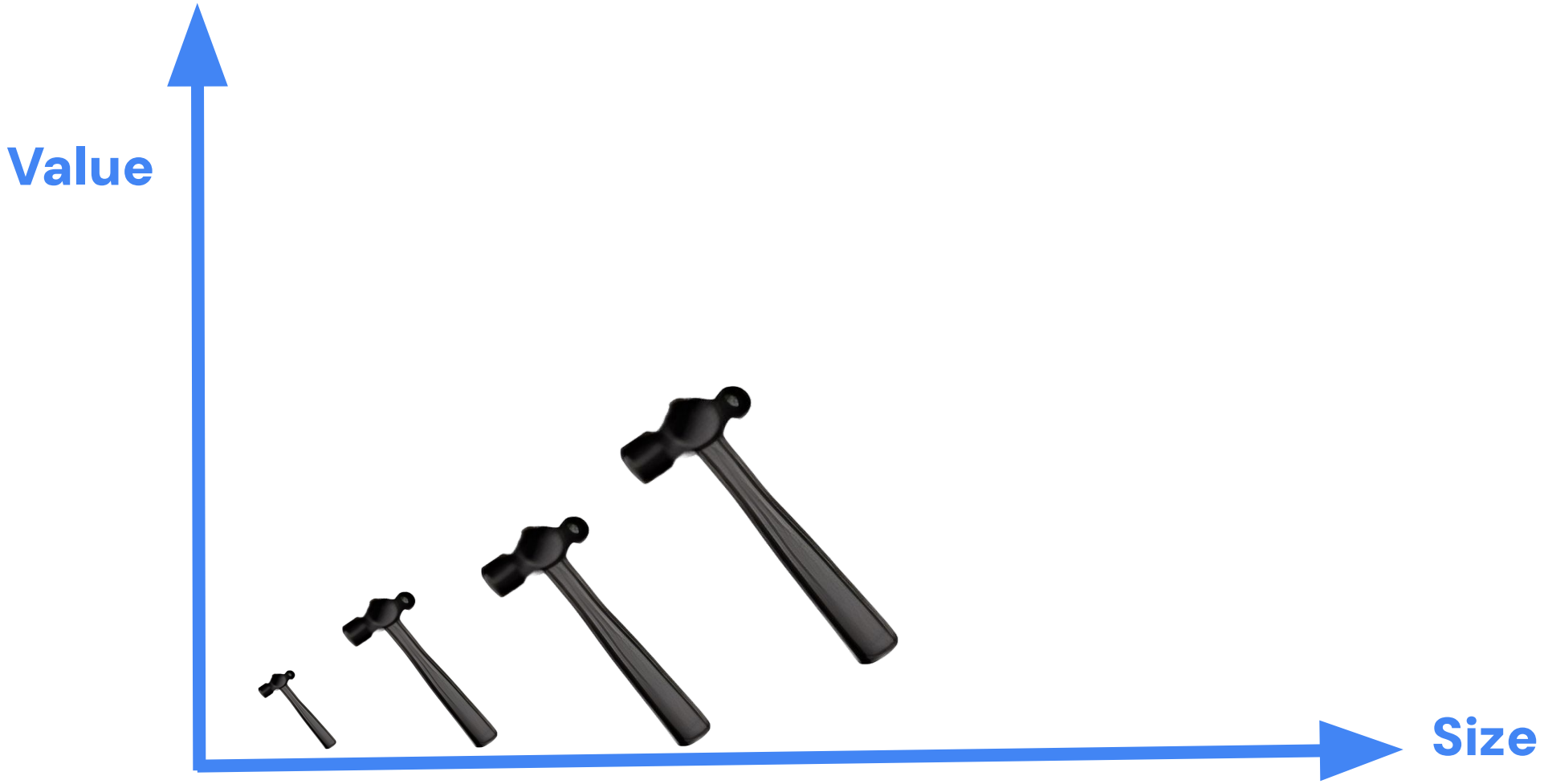
What are the various options in building this software

What can we use this software for in Bharat

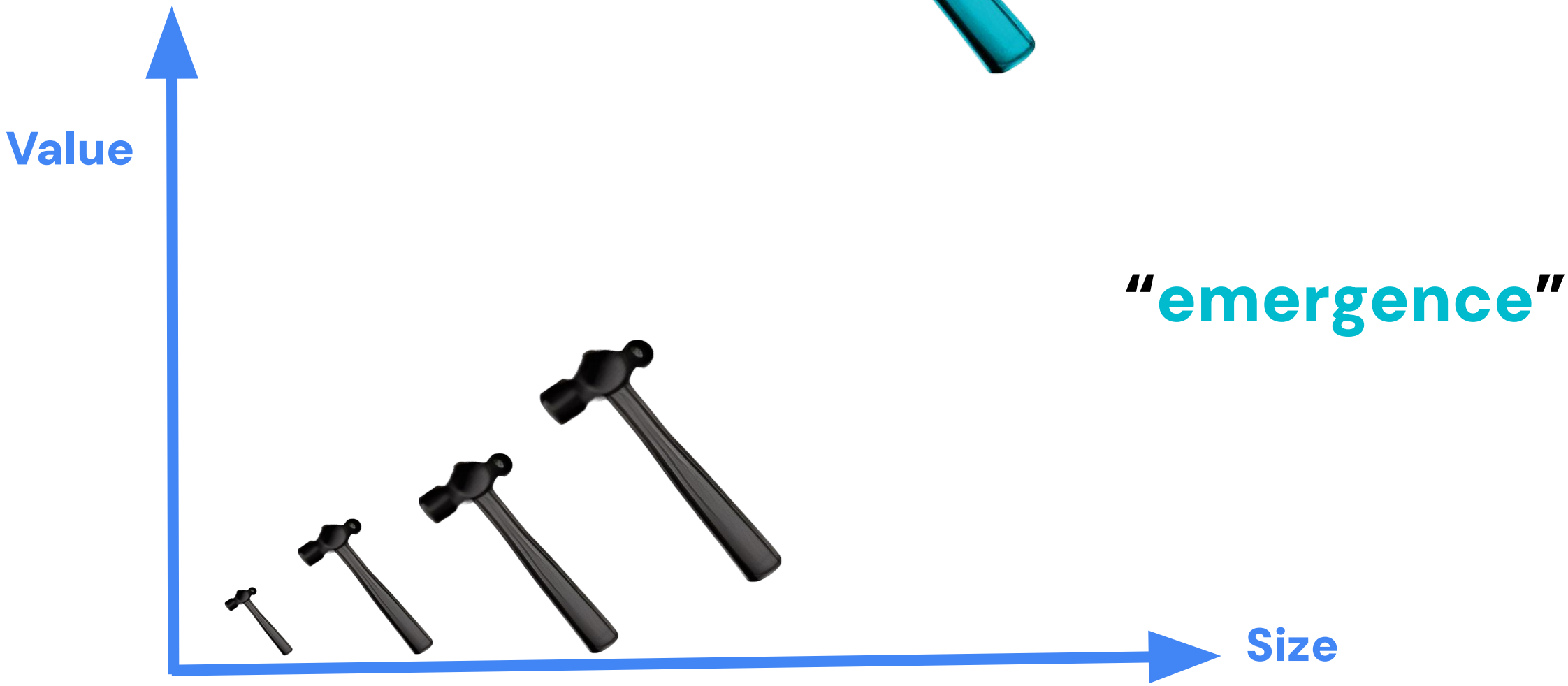
We have never had a situation like this …



We have never had a situation like this …

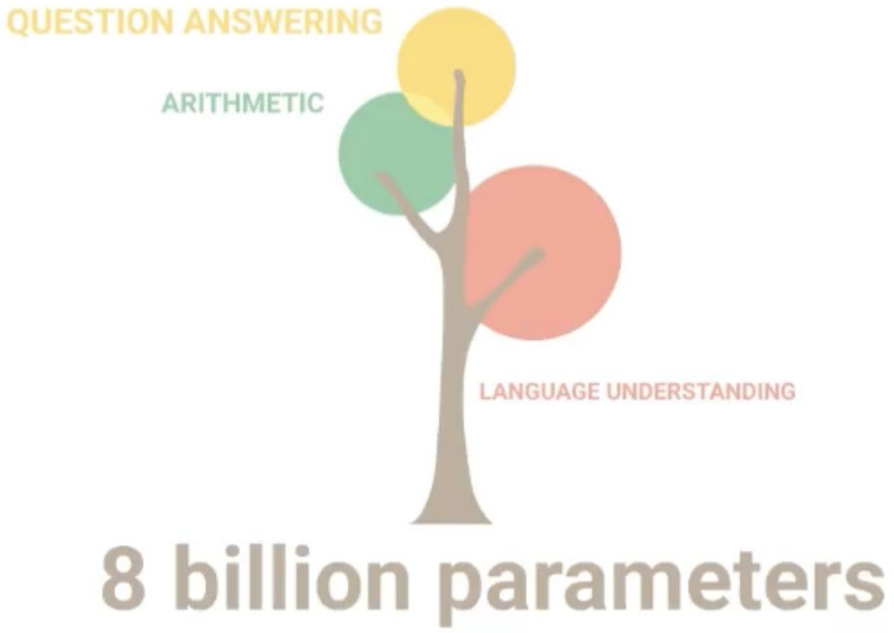


We have never had a situation like this …



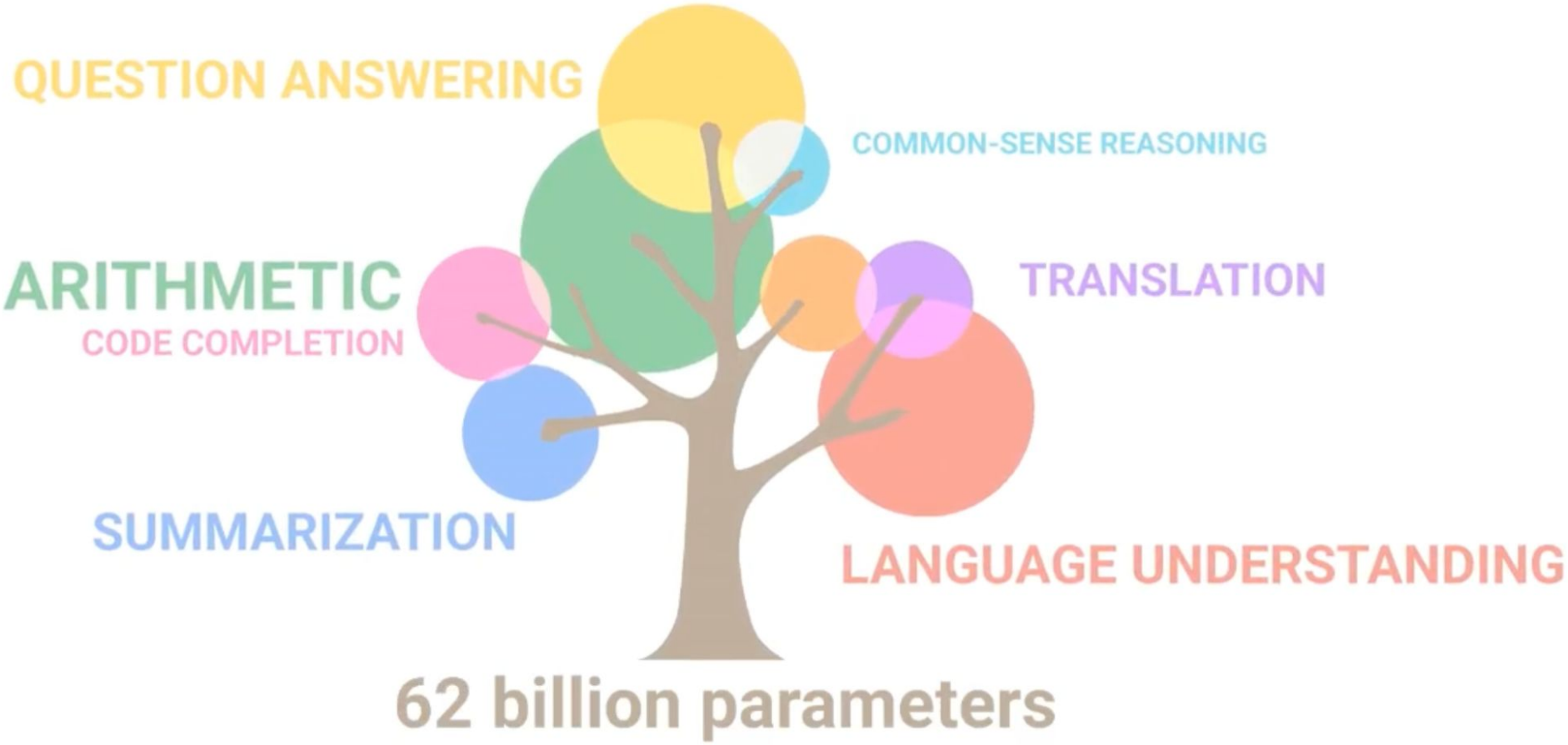
"emergence"

What emerges with large language models?



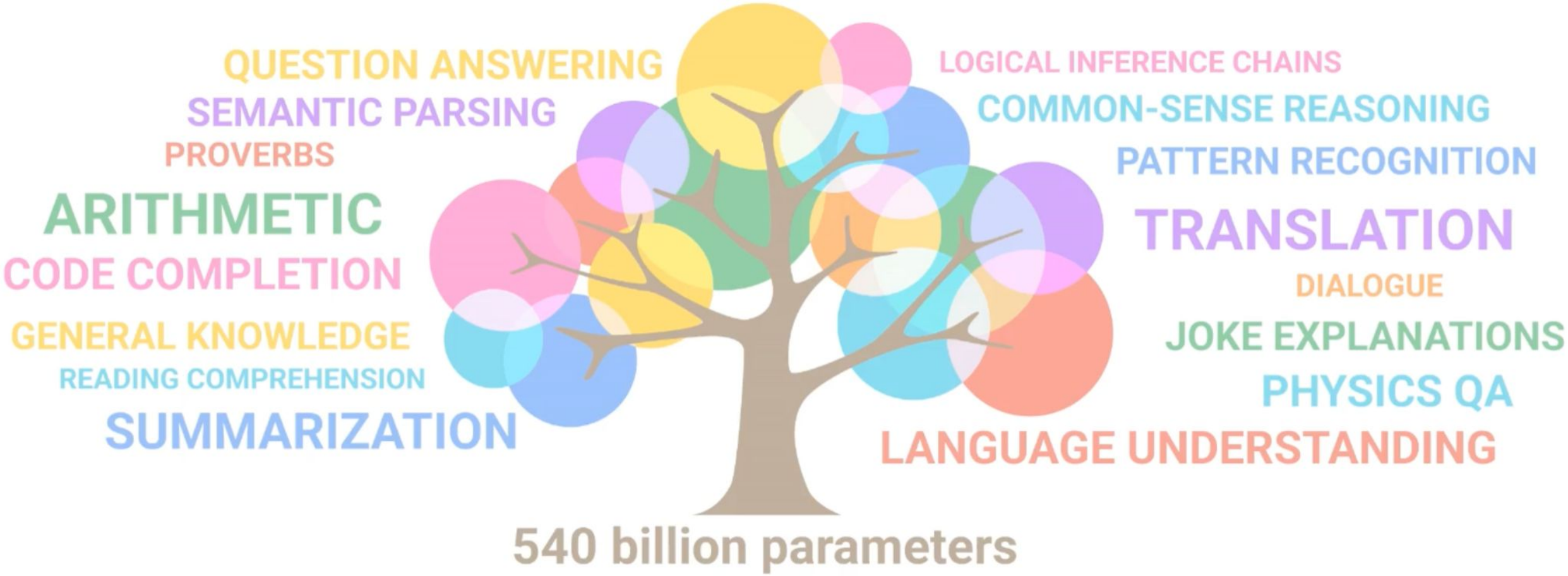
Source: https://www.assemblyai.com/blog/emergent-abilities-of-large-language-models/

What emerges with large language models?



Source: https://www.assemblyai.com/blog/emergent-abilities-of-large-language-models/

What emerges with large language models?



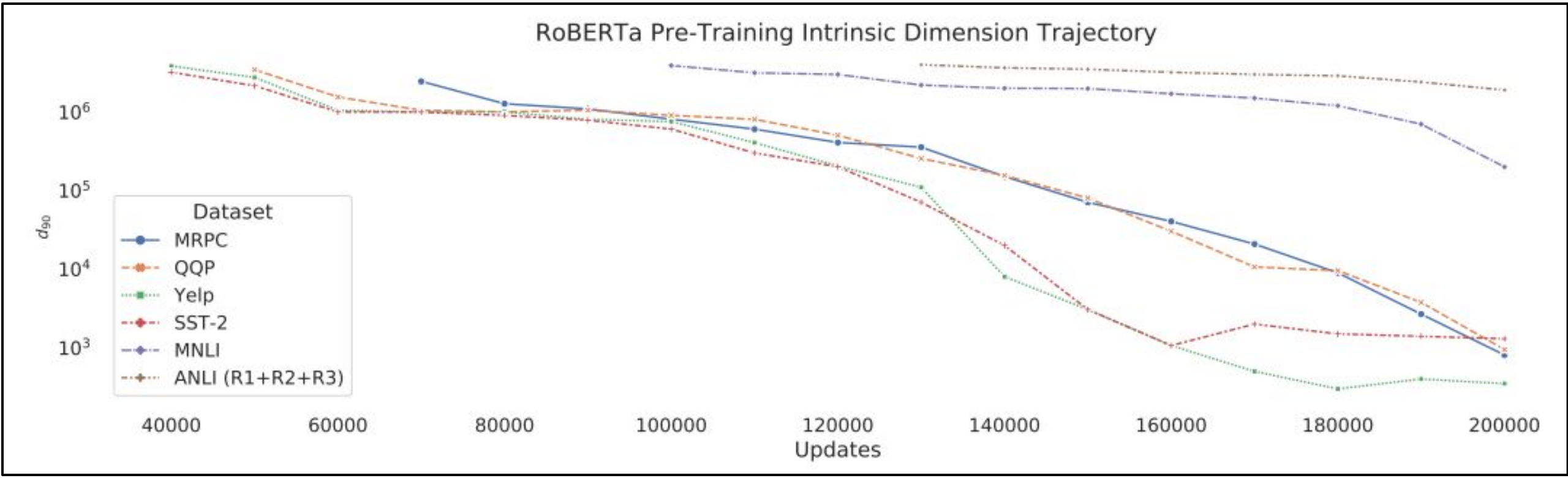
Source: https://www.assemblyai.com/blog/emergent-abilities-of-large-language-models/

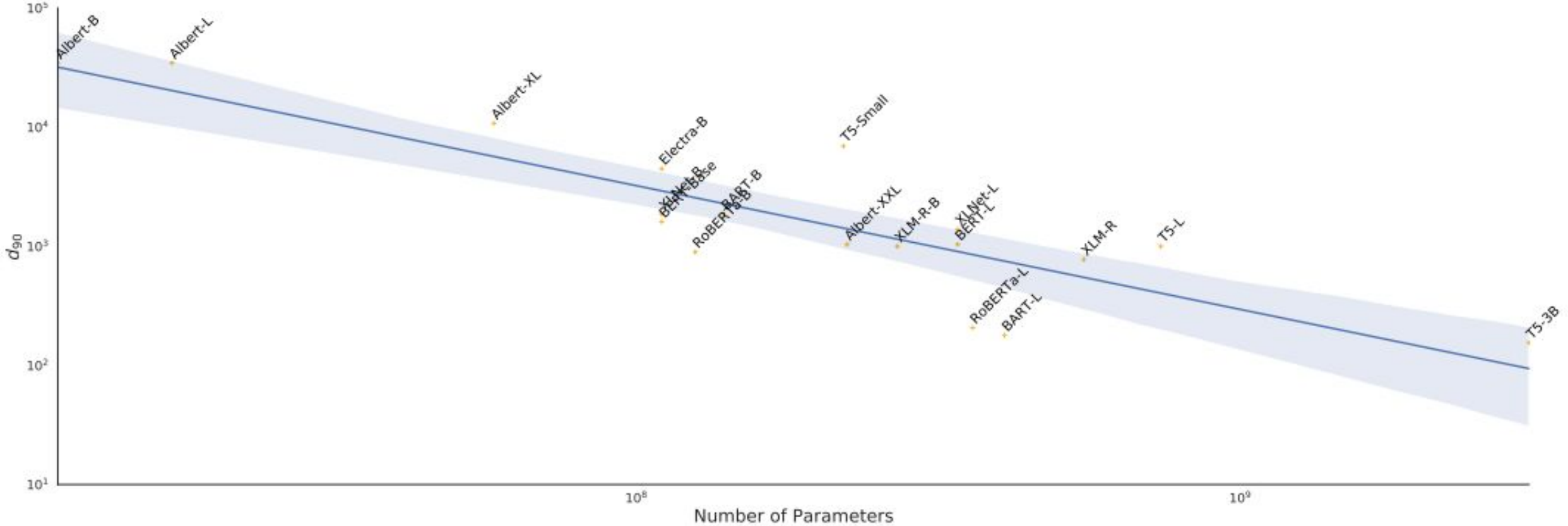
INTRINSIC DIMENSIONALITY EXPLAINS THE EFFECTIVENESS OF LANGUAGE MODEL FINE-TUNING

Armen Aghajanyan, Luke Zettlemoyer, Sonal Gupta
Facebook
{armenag,lsz,sonalgupta}@fb.com

ABSTRACT

Although pretrained language models can be fine-tuned to produce state-of-the-art results for a very wide range of language understanding tasks, the dynamics of this process are not well understood, especially in the low data regime. Why can we use relatively vanilla gradient descent algorithms (e.g., without strong regularization) to tune a model with hundreds of millions of parameters on datasets with only hundreds or thousands of labeled examples? In this paper, we argue that analyzing fine-tuning through the lens of intrinsic dimension provides us with empirical and theoretical intuitions to explain this remarkable phenomenon. We empirically show that common pre-trained models have a very low intrinsic dimension; in other words, there exists a low dimension reparameterization that is as effective for fine-tuning as the full parameter space. For example, by optimizing only 200 trainable parameters randomly projected back into the full space, we can tune a RoBERTa model to achieve 90% of the full parameter performance levels on MRPC. Furthermore, we empirically show that pre-training implicitly minimizes intrinsic dimension and, perhaps surprisingly, larger models tend to have lower intrinsic dimension after a fixed number of pre-training updates, at least in part explaining their extreme effectiveness. Lastly, we connect intrinsic dimensionality with low dimensional task representations and compression based generalization bounds to provide intrinsic-dimension-based generalization bounds that are independent of the full parameter count.


RoBERTa Pre-Training Intrinsic Dimension Trajectory
d_{90}
10^6
10^5
10^4
10^3
Dataset
MRPC
QQP
Yelp
SST-2
MNLI
ANLI (R1+R2+R3)
40000
60000
80000
100000
120000
140000
160000
180000
200000
Updates


10^5
10^4
10^3
10^2
10^1
d_{90}
Albert-B
Albert-L
Albert-XL
Electra-B
T5-Small
BART-B
RoBERTa-B
Albert-XXL
XLM-R-B
XLNet-L
BERT-L
XLM-R
T5-L
RoBERTa-L
BART-L
T5-3B
10^8
10^9
Number of Parameters

The trikuti (the confluence of the three)

- Pre-training
- Fine-tuning
- Prompting

The Full Stack of AI

Currently: how do I know if i should refine prompt, finetune model, or pretrain model. cost? business model?

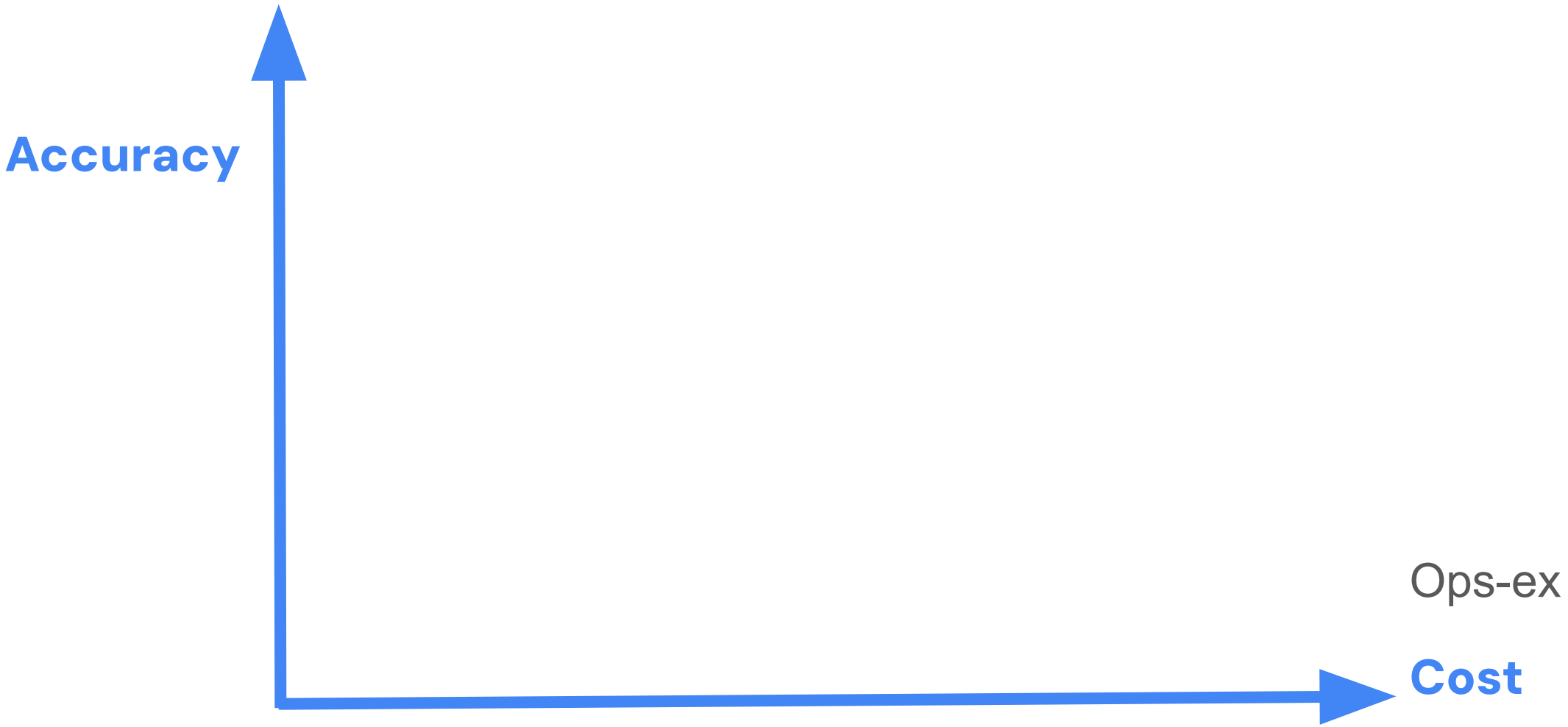
The trikuti (the confluence of the three)

- Pre-training
- Fine-tuning
- Prompting

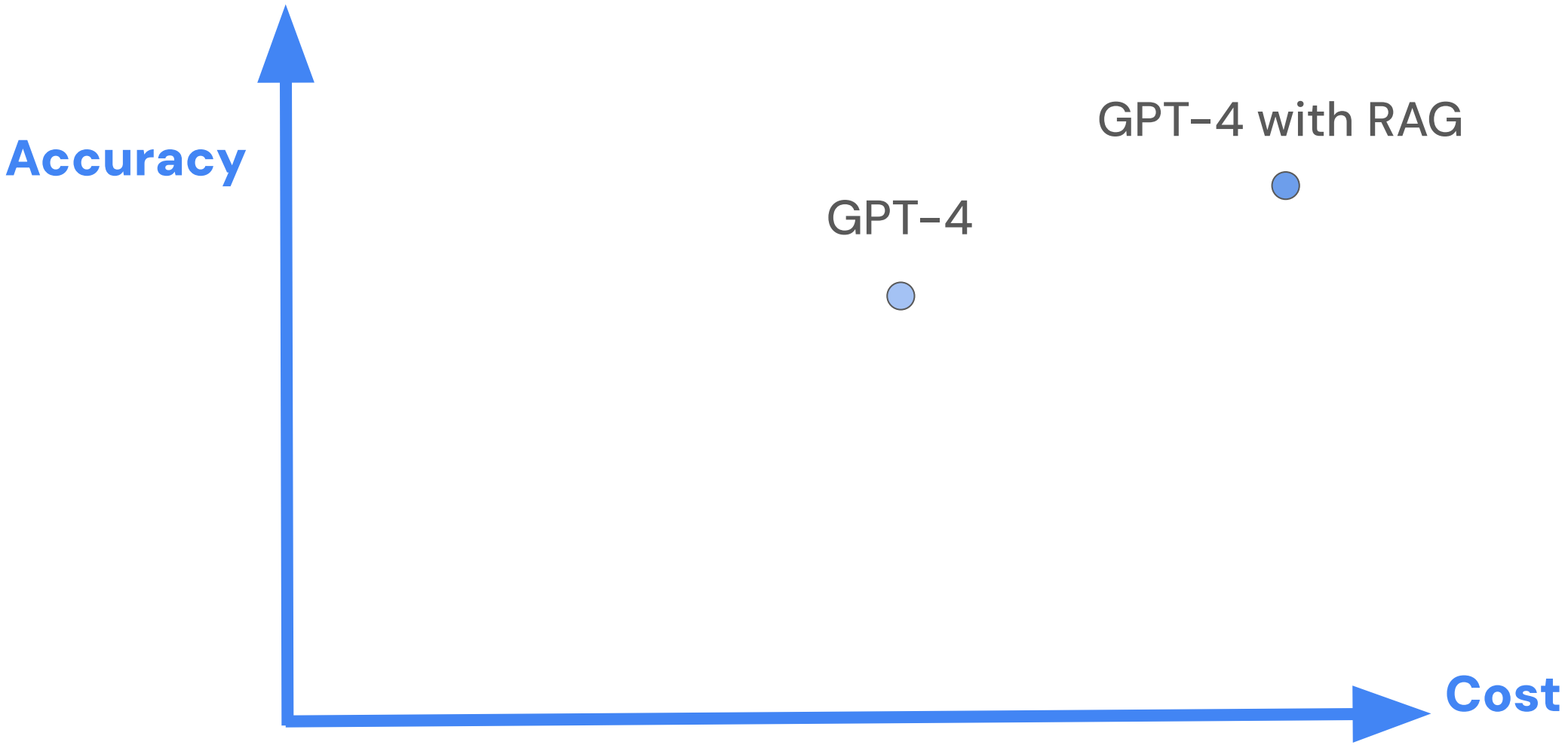
The Full Stack of AI

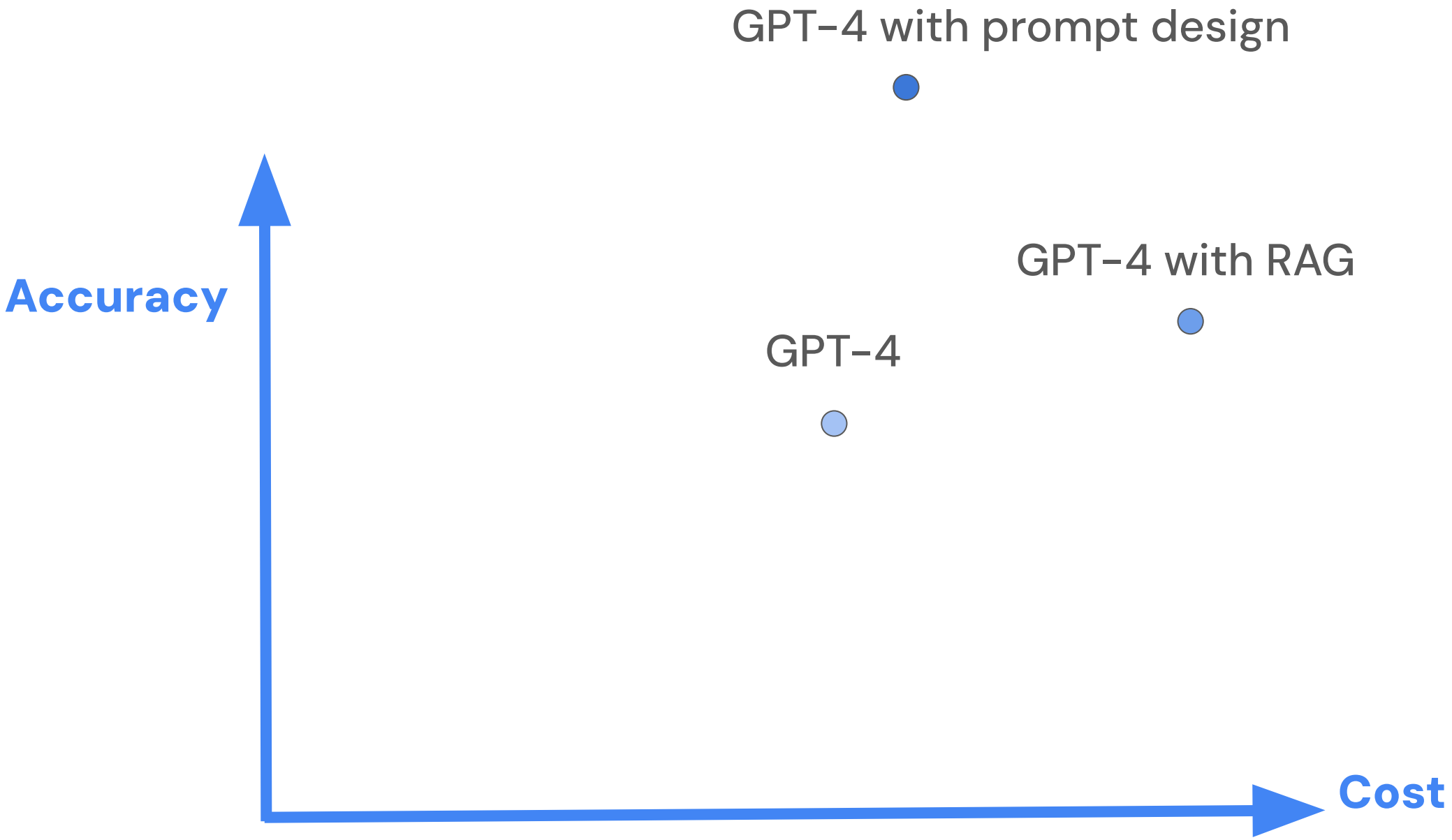
take: cost/expertise requirement for all three layers will rapidly fall, need to prepare for full stack

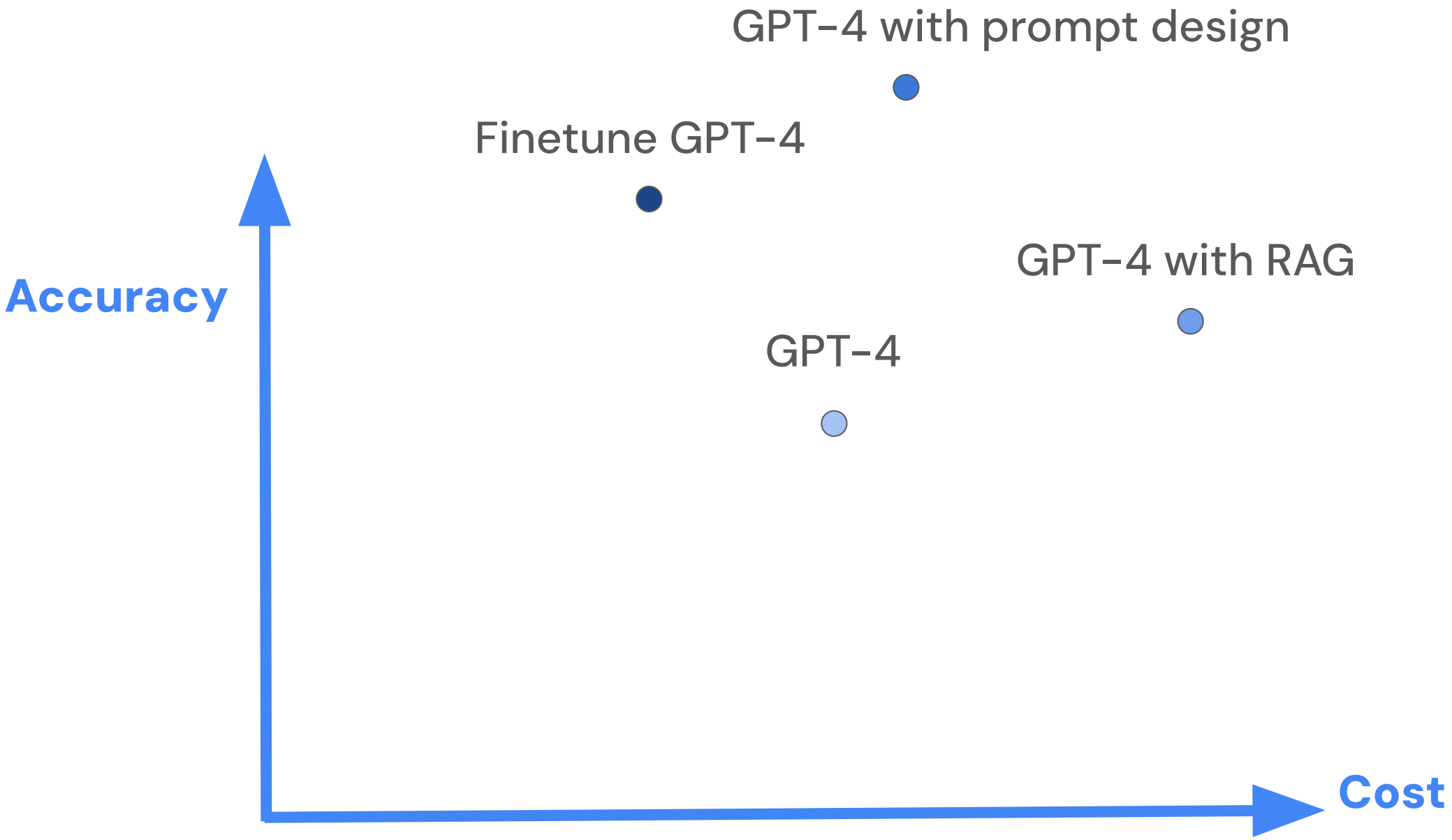
What are the options in this type of software (assuming we have hit ‘scale’)

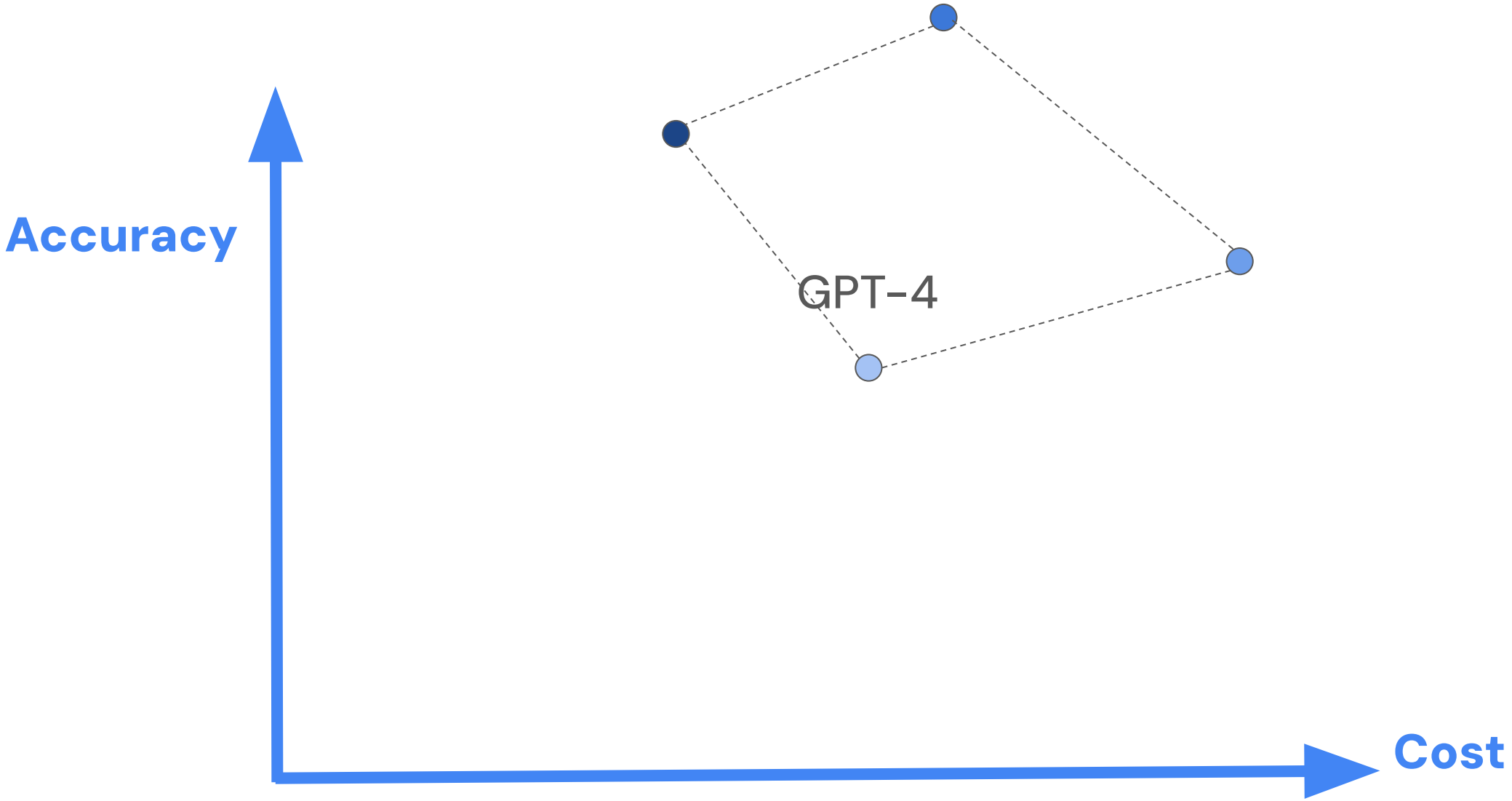


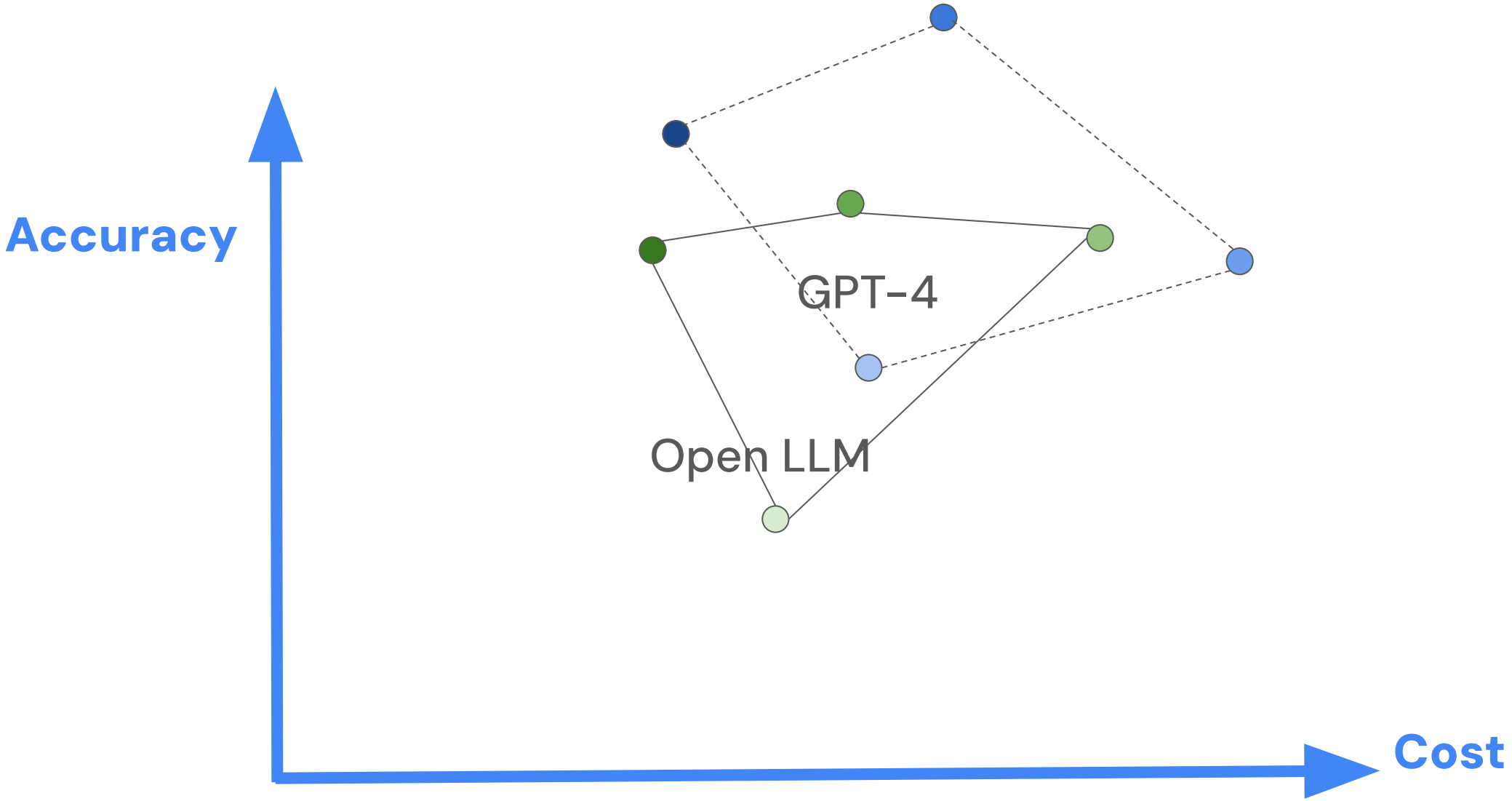

Accuracy
GPT-4
Cost

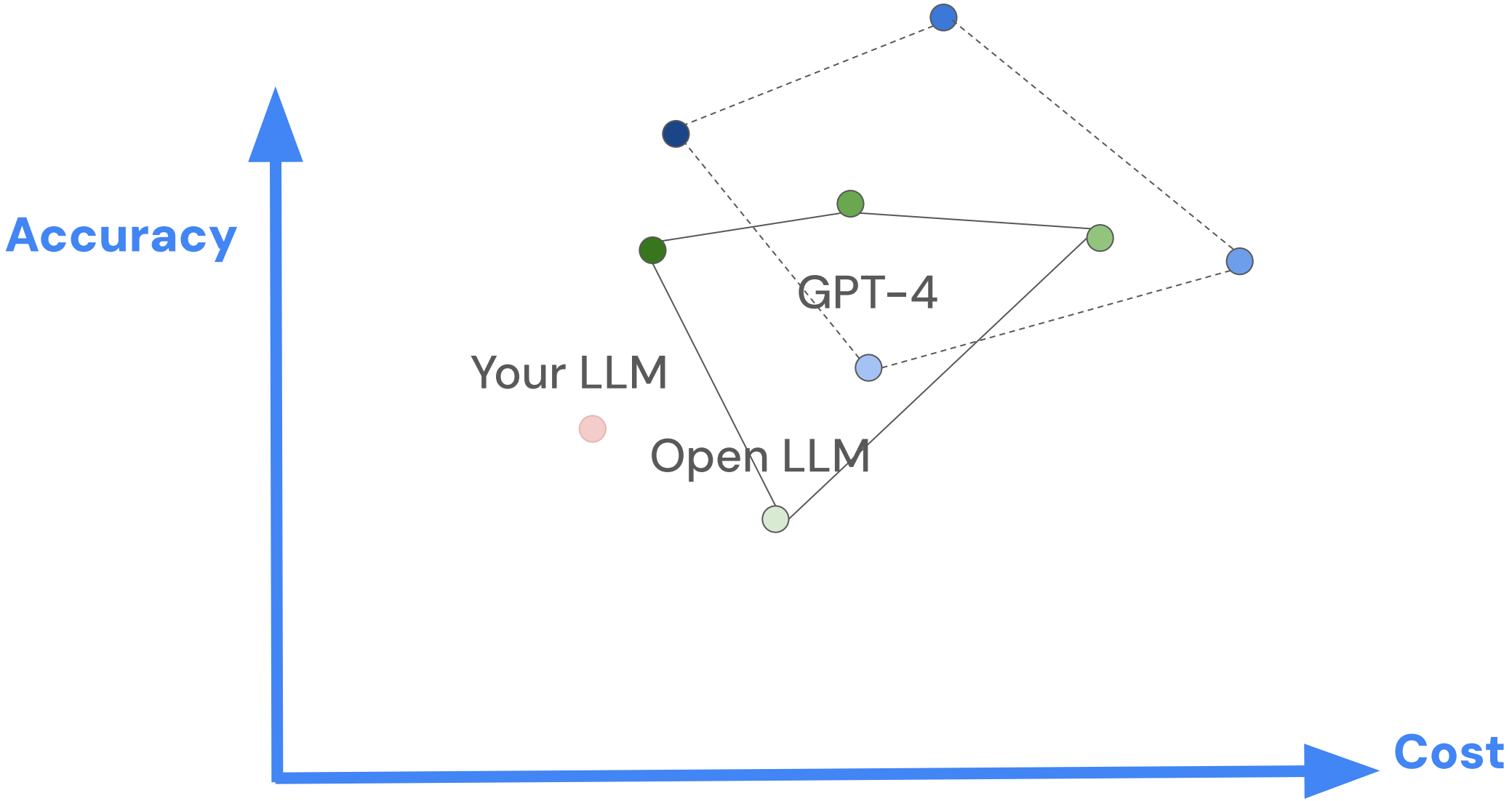

Accuracy
GPT-4 with RAG
GPT-4
Cost


GPT-4 with prompt design
GPT-4 with RAG
GPT-4
Accuracy
Cost


GPT-4 with prompt design
Finetune GPT-4
GPT-4 with RAG
GPT-4
Accuracy
Cost


Accuracy
GPT-4
Cost


Accuracy
GPT-4
Open LLM
Cost


Accuracy
GPT-4
Your LLM
Open LLM
Cost

Jais and *Jais-chat*: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models

Neha Sengupta[1] Sunil Kumar Sahu[1] Bokang Jia[1] Satheesh Katipomu[1]
Haonan Li[2] Fajri Koto[2] William Marshall[3] Gurpreet Gosal[3]
Cynthia Liu[3] Zhiming Chen[3] Osama Mohammed Afzal[2] Samta Kamboj[1]
Onkar Pandit[1] Rahul Pal[1] Lalit Pradhan[1] Zain Muhammad Mujahid[2]
Massa Baali[2] Xudong Han[2] Sondos Mahmoud Bsharat[2] Alham Fikri Aji[2]
Zhiqiang Shen[2] Zhengzhong Liu[2] Natalia Vassilieva[3] Joel Hestness[3] Andy Hock[3]
Andrew Feldman[3] Jonathan Lee[1] Andrew Jackson[1] Hector Xuguang Ren[2]
Preslav Nakov[2] Timothy Baldwin[2] Eric Xing[2]

[1]Inception, UAE
[2]Mohamed bin Zayed University of Artificial Intelligence, UAE
[3]Cerebras Systems

Abstract

We introduce *Jais* and *Jais-chat*, new state-of-the-art Arabic-centric foundation and instruction-tuned open generative large language models (LLMs). The models are based on the GPT-3 decoder-only architecture and are pretrained on a mixture of Arabic and English texts, including source code in various programming languages. With 13 billion parameters, they demonstrate better knowledge and reasoning capabilities in Arabic than any existing open Arabic and multilingual models by a sizable margin, based on extensive evaluation. Moreover, the models are competitive in English compared to English-centric open models of similar size, despite being trained on much less English data. We provide a detailed description of the training, the tuning, the safety alignment, and the evaluation of the models. We release two open versions of the model —the foundation *Jais* model, and an instruction-tuned *Jais-chat* variant— with the aim of promoting research on Arabic LLMs.

Shows how to use a GPT-3 architecture and build a model from scratch

(reaches GPT3.5 level for targeted use-cases)

BloombergGPT: A Large Language Model for Finance

Shijie Wu[1,*], **Ozan İrsoy**[1,*], **Steven Lu**[1,*], **Vadim Dabravolski**[1], **Mark Dredze**[1,3], **Sebastian Gehrmann**[1], **Prabhanjan Kambadur**[1], **David Rosenberg**[2], **Gideon Mann**[1]

[1] **Bloomberg, New York, NY USA**

[2] **Bloomberg, Toronto, ON Canada**

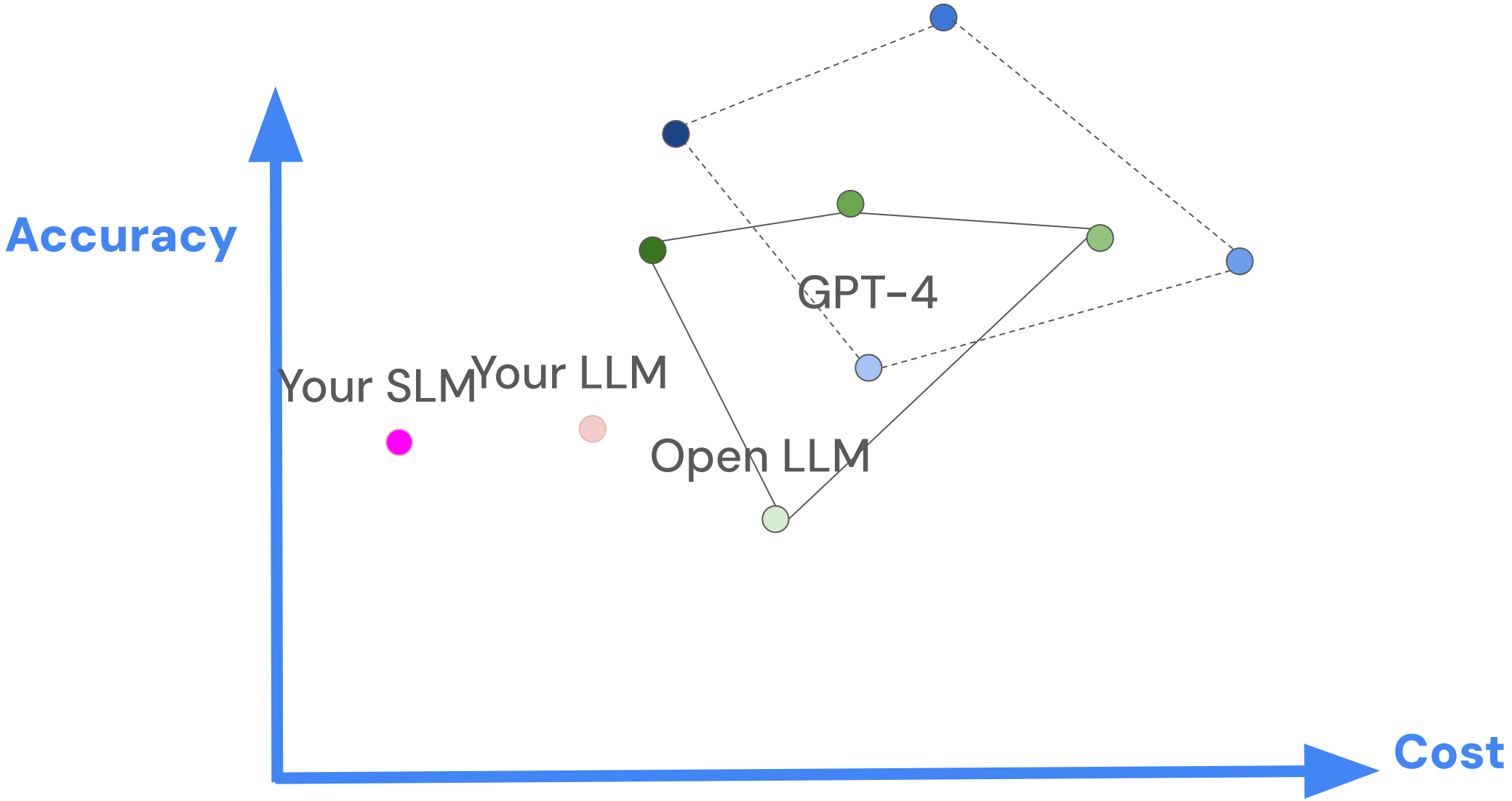
[3] **Computer Science, Johns Hopkins University, Baltimore, MD USA**

Abstract

The use of NLP in the realm of financial technology is broad and complex, with applications ranging from sentiment analysis and named entity recognition to question answering. Large Language Models (LLMs) have been shown to be effective on a variety of tasks; however, no LLM specialized for the financial domain has been reported in literature. In this work, we present BloombergGPT, a 50 billion parameter language model that is trained on a wide range of financial data. We construct a 363 billion token dataset based on Bloomberg's extensive data sources, perhaps the largest domain-specific dataset yet, augmented with 345 billion tokens from general purpose datasets. We validate BloombergGPT on standard LLM benchmarks, open financial benchmarks, and a suite of internal benchmarks that most accurately reflect our intended usage. Our mixed dataset training leads to a model that outperforms existing models on financial tasks by significant margins without sacrificing performance on general LLM benchmarks. Additionally, we explain our modeling choices, training process, and evaluation methodology. We release Training Chronicles (Appendix C) detailing our experience in training BloombergGPT.

Contents

Builds a domain-specific model with a custom tokenizer


Accuracy
Your SLM
Your LLM
Open LLM
GPT-4
Cost

TinyStories: How Small Can Language Models Be and Still Speak Coherent English?

Ronen Eldan* and Yuanzhi Li†

Microsoft Research

April 2023

Abstract

Language models[4, 5, 21] (LMs) are powerful tools for natural language processing, but they often struggle to produce coherent and fluent text when they are **small**. Models with around 125M parameters such as GPT-Neo (small) [3] or GPT-2 (small) [23] can rarely generate coherent and consistent English text beyond a few words even after extensive training. This raises the question of whether the emergence of the ability to produce coherent English text only occurs at larger scales (with hundreds of millions of parameters or more) and complex architectures (with many layers of global attention).

In this work, we introduce **TinyStories**, a synthetic dataset of short stories that only contain words that a typical 3 to 4-year-olds usually understand, generated by GPT-3.5 and GPT-4. We show that TinyStories can be used to train and evaluate LMs that are much smaller than the state-of-the-art models (**below 10 million total parameters**), or have much simpler architectures (**with only one transformer block**), yet still produce fluent and consistent stories with several paragraphs that are diverse and have almost perfect grammar, and demonstrate reasoning capabilities.

We also introduce a new paradigm for the evaluation of language models: We suggest a framework which uses GPT-4 to grade the content generated by these models as if those were stories written by students and graded by a (human) teacher. This new paradigm overcomes the flaws of standard benchmarks which often require the model's output to be very structured, and moreover it provides a multidimensional score for the model, providing scores for different capabilities such as grammar, creativity and instruction-following.

We hope that TinyStories can facilitate the development, analysis and research of LMs, especially for low-resource or specialized domains, and shed light on the emergence of language capabilities in LMs.

1 Introduction

Natural language is rich and diverse. It is not only a system of rules and symbols, but also a way of conveying and interpreting meaning [32]. To understand and produce language, one needs not only to master the technical rules of grammar and knowledge of vocabulary, but also to have sufficient factual information and to be able to reason

Shows that models with tens of millions of params (not billions) can be pretrained to generate tiny stories

Textbooks Are All You Need

Suriya Gunasekar Yi Zhang Jyoti Aneja Caio César Teodoro Mendes
Allie Del Giorno Sivakanth Gopi Mojan Javaheripi Piero Kauffmann
Gustavo de Rosa Olli Saarikivi Adil Salim Shital Shah Harkirat Singh Behl
Xin Wang Sébastien Bubeck Ronen Eldan Adam Tauman Kalai Yin Tat Lee
Yuanzhi Li

Microsoft Research

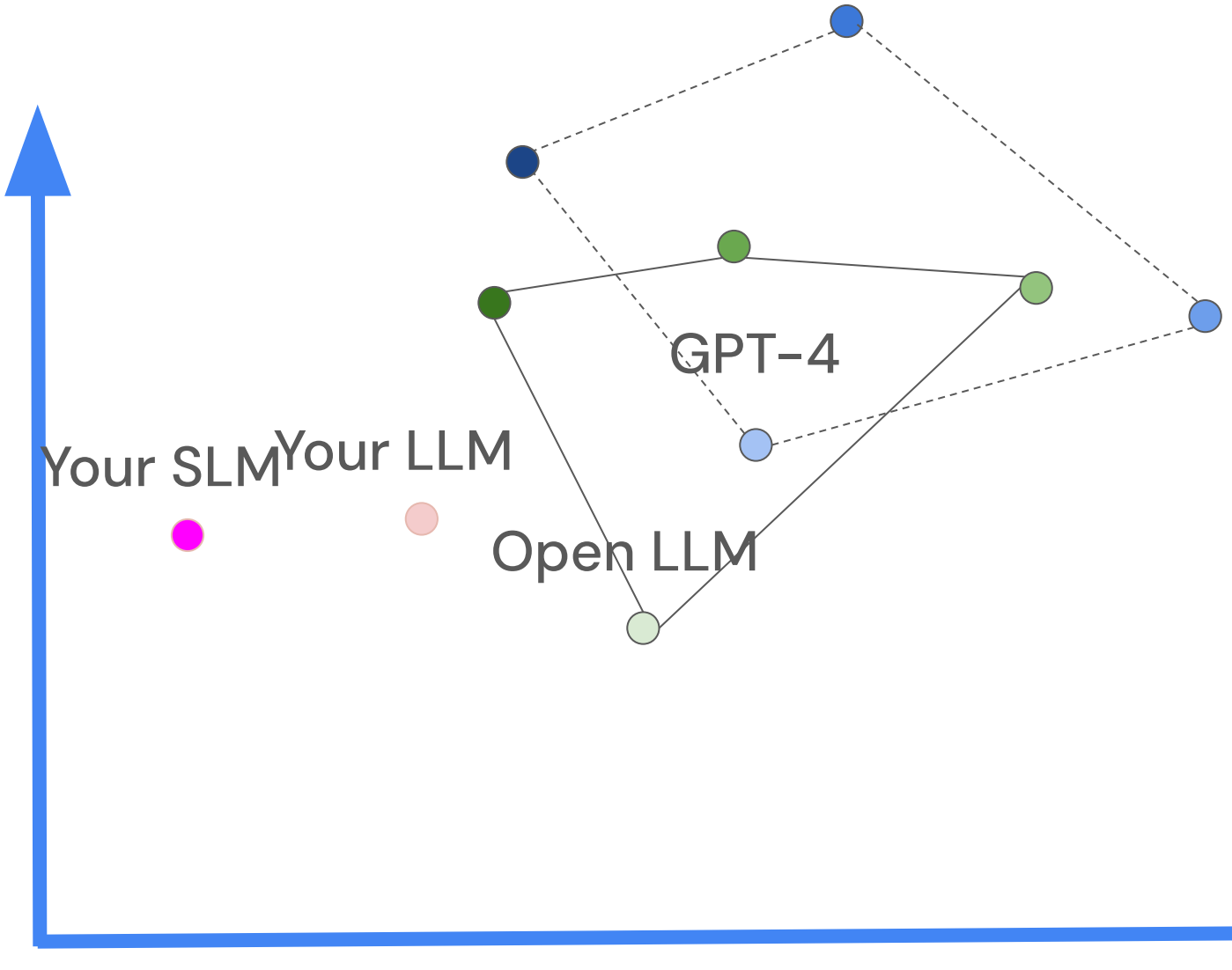
Abstract

We introduce **phi-1**, a new large language model for code, with significantly smaller size than competing models: **phi-1** is a Transformer-based model with 1.3B parameters, trained for 4 days on 8 A100s, using a selection of "textbook quality" data from the web (6B tokens) and synthetically generated textbooks and exercises with GPT-3.5 (1B tokens). Despite this small scale, **phi-1** attains **pass@1** accuracy 50.6% on HumanEval and 55.5% on MBPP. It also displays surprising emergent properties compared to **phi-1-base**, our model *before* our finetuning stage on a dataset of coding exercises, and **phi-1-small**, a smaller model with 350M parameters trained with the same pipeline as **phi-1** that still achieves 45% on HumanEval.

1 Introduction

The art of training large artificial neural networks has made extraordinary progress in the last decade, especially after the discovery of the Transformer architecture [VSP+17], yet the science behind this success remains limited. Amidst a vast and confusing array of results, a semblance of order emerged around the same time as Transformers were introduced, namely that performance improves somewhat predictably as one scales up either the amount of compute or the size of the network [HNA+17], a phenomenon which is now referred to as *scaling laws* [KMH+20]. The subsequent exploration of scale in deep learning was guided by these scaling laws [BMR+20], and discoveries of variants of these laws led to rapid jump in performances [HBM+22]. In this work, following the footsteps of Eldan and Li [EL23], we explore the improvement that can be obtained along a different axis: the *quality* of the data. It has long been known that higher quality data leads to better results, e.g., data cleaning is an important part of modern dataset creation [RSR+20], and it can yield other side benefits such as somewhat smaller datasets [LYR+23, YGK+23] or allowing for more passes on the data [MRB+23]. The recent work of Eldan and Li on TinyStories (a high quality dataset synthetically generated to teach English to neural networks) showed that in fact the effect

Shows that models trained on billions of tokens (not trillions) can effectively answer programming questions



Full-stack skill-set

- Prompt engineering
- RAG
- Evaluation
- Serving
- Curating finetuning data
- Finetuning
- Curating pretraining data
- Continual pretraining
- Synthetic pretrating data
- Pretraining from scratch

What can you do with Rs 10 of AI?

Currently:

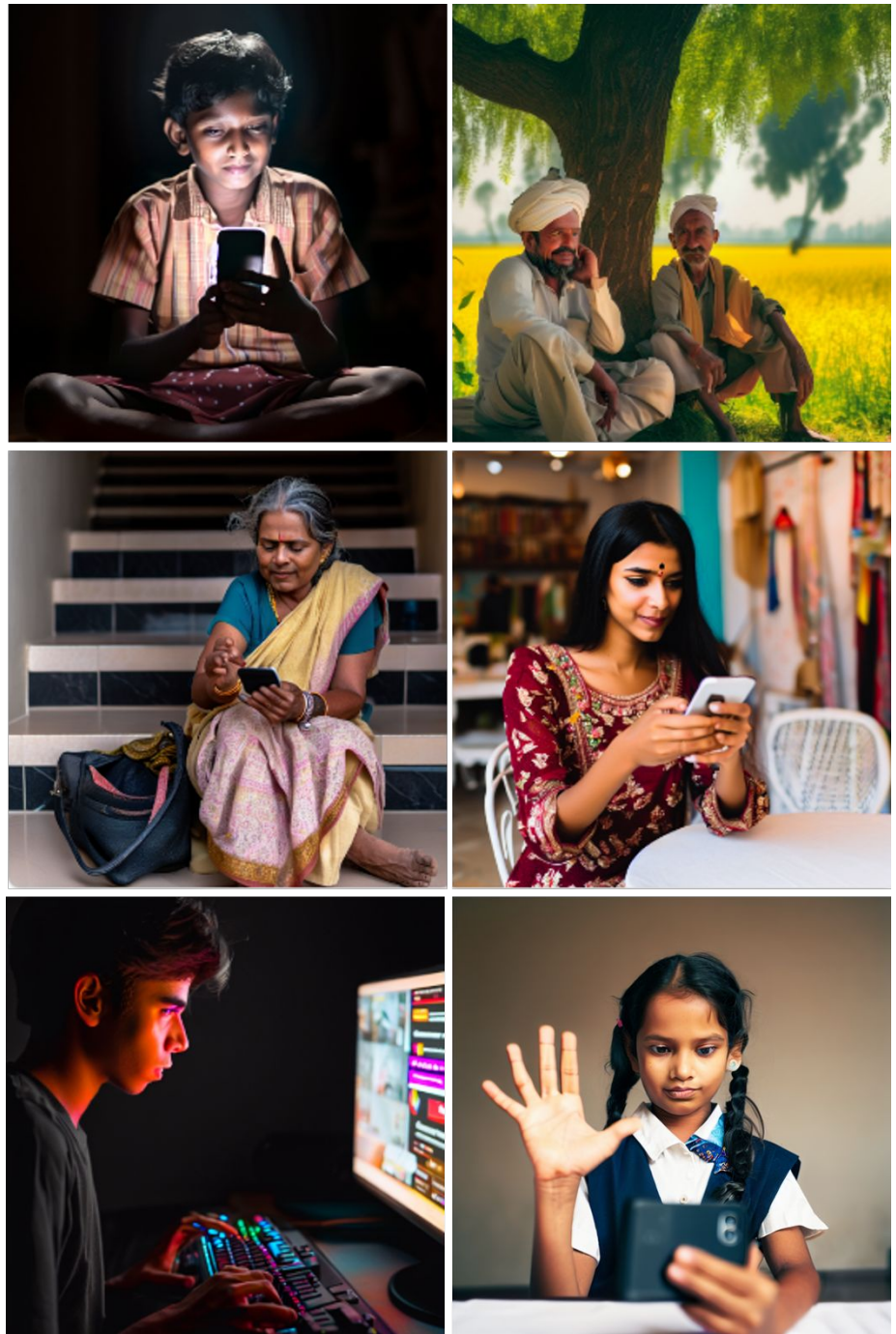
- Serve 1000 ads
- Recommend 100 youtube videos
- Translate a 500 word article

What can you do with Rs 10 of AI?

Currently:

- Serve 1000 ads
- Recommend 100 youtube videos
- Translate a 500 word article

With GenAI ->



Student: If Hanuman and Thanos were to meet, what would they talk about?

Figure 2: A high school student asks an LLM to generate a discussion between Hanuman and Thanos. Image generated by DALL-E 2."

Student: If Hanuman and Thanos were to meet, what would they talk about?

Scaffolded curiosity for children – overcome the bottleneck of content creation

Figure 2: A high school student asks an LLM to generate a discussion between Hanuman and Thanos. Image generated by DALL-E 2."

Farmer: ਮੈਂ 4 ਏਕੜ 'ਤੇ ਖੇਤੀ ਕਰਦਾ ਹਾਂ, ਕੀ ਮੈਂ ਪੀਐਮ ਕਿਸਾਨ ਲਈ ਯੋਗ ਹਾਂ?

(Translation: I farm on 4 acres, am I eligible for PM Kisan?)

Grounded access to government data

Figure 4: A farmer speaks in Punjabi to an LLM powered voice call system to answer custom queries reasoned over an English document. Image generated by DALL-E 2.

Lady: ನಾನು ಮೇ ತಿಂಗಳಿನಿಂದ ತಿಂಗಳಿಗೆ 20000 ಗಳಿಸುತ್ತೇನೆ ಮತ್ತು ಜುಲೈನಿಂದ 23000 ಕ್ಕೆ ಹೆಚ್ಚಿಸುತ್ತೇನೆ. ನಾನು ಪ್ರಸ್ತುತ 52800 ಉಳಿತಾಯವನ್ನು ಹೊಂದಿದ್ದೇನೆ. ನನ್ನ ಮಾಸಿಕ ವೆಚ್ಚಗಳು 15000. ಅಕ್ಟೋಬರ್ ಅಂತ್ಯದ ವೇಳೆಗೆ ನಾನು ದೀಪಾವಳಿಯ ಮೊದಲು ಎಷ್ಟು ಉಳಿಸಬಹುದು

(Translation: I earn 20000 per month from May and increase to 23000 from July. I currently have savings of 52800 My monthly expenses are 15000. By the end of October how much can I save before Diwali.)

NL interface to tools like spreadsheets

Figure 6: A domestic help converses with an LLM bot and do her finances without knowing how to operate a spreadsheet app. Image generated by DALL-E 2."

Movie buff: I recently watched the Mandalorian series. I liked it because of the both the sci-fi setting and the sage-like nature of Mandalorians. The father-son dynamics at the center stage is also delicately done. However, I did not quite appreciate the fight sequences which amounted to not much either drama wise or in weaponry.

What should I watch next?

Figure 10: A user of a video website builds a personal recommendation engine based on his nuanced preferences. Image generated by DALL-E 2.

Girl: <can teach an AI to learn her sign language by providing example images and then giving feedback in natural language>

Figure 12: A Deaf girl teaches her mobile to recognize her signs. Image generated by DALL-E 2.

Source Image: ISRO

...to play a meaningful role nationally, and in the community of nations, we must be second to none in the application of advanced technologies to the real problems of man and society.

”

Dr. Vikram A. Sarabhai
1919 - 1971