ICML
International Conference
On Machine Learning

Black-Box Tuning for Language-Model-as-a-Service

Tianxiang Sun

School of Computer Science, Fudan University

https://txsun1997.github.io/

Our Team

Tianxiang Sun[1]

Yunfan Shao[1]

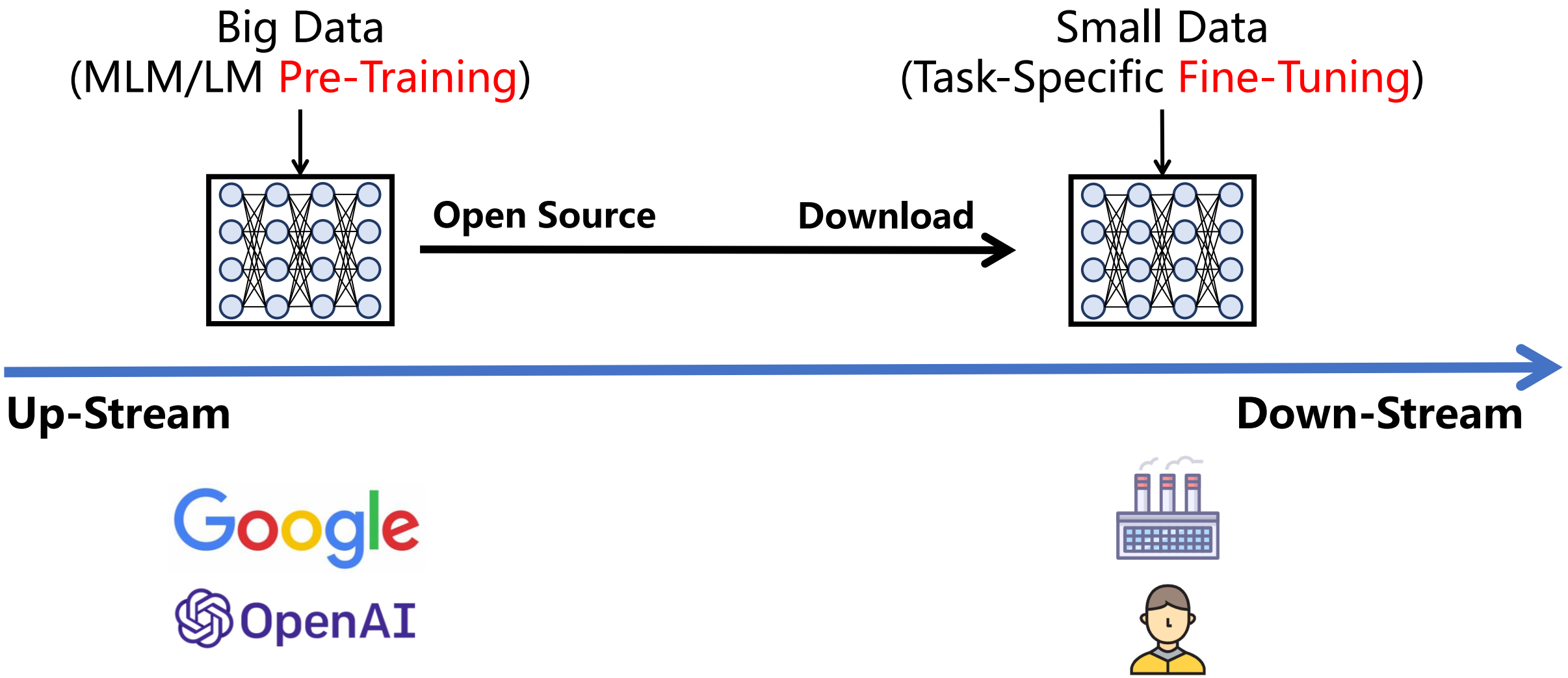
Hong Qian[2]

Xuanjing Huang[1]

Xipeng Qiu[1,3]

[1] Fudan University [2] East China Normal University [3] Peng Cheng Laboratory

Pre-train, then fine-tune



When language models become larger…

In the era of large language models (LLMs)…

- **Servers** often do not open-source the weights of LLMs due to commercial reasons
- **Users** usually do not have enough resources to run LLMs

The emergent ability of LLMs

- Manually craft text prompt to query LLMs
- In-context learning (GPT-3, Brown et al., 2020)

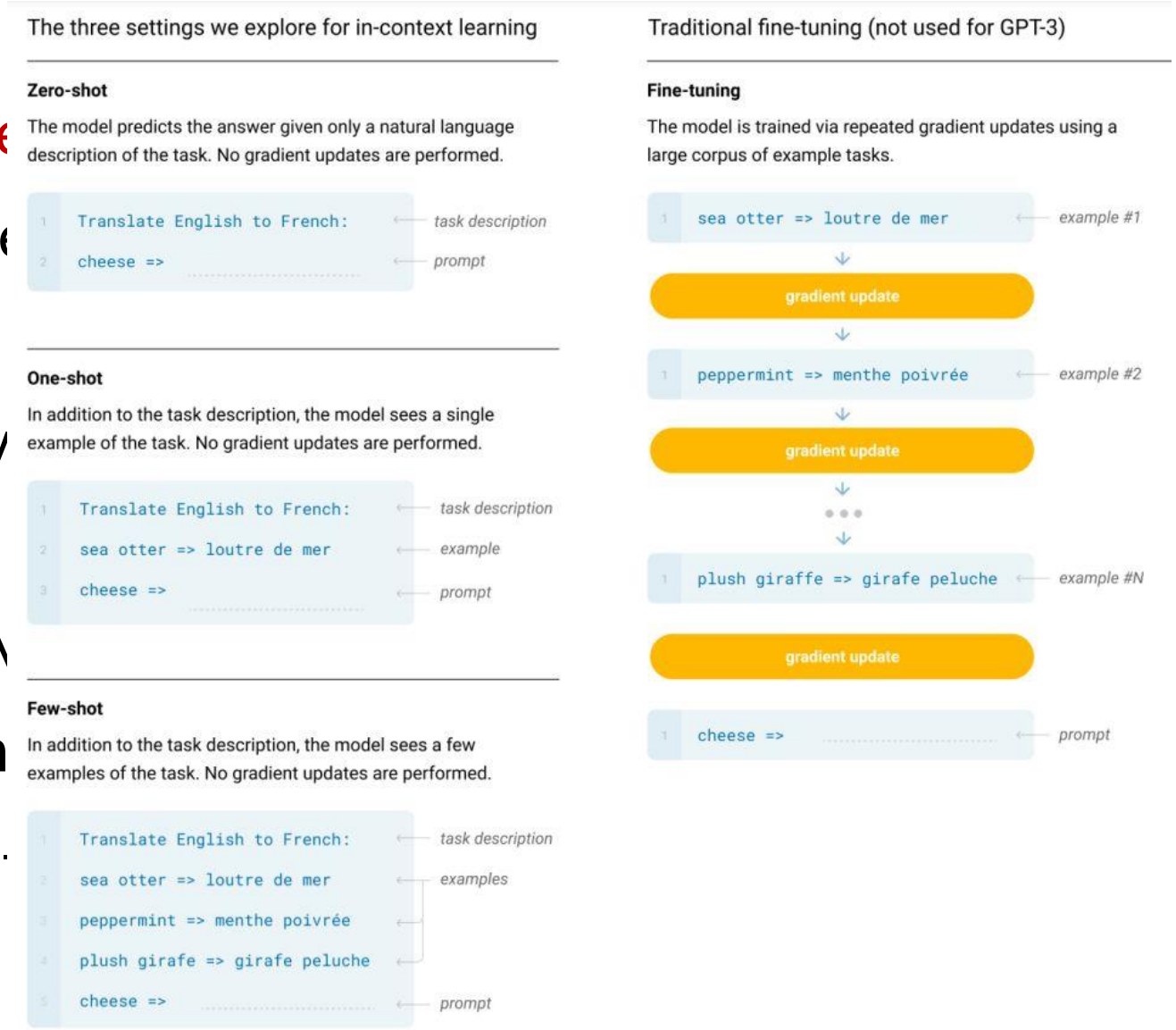
When language models become larger…

In the era of large language

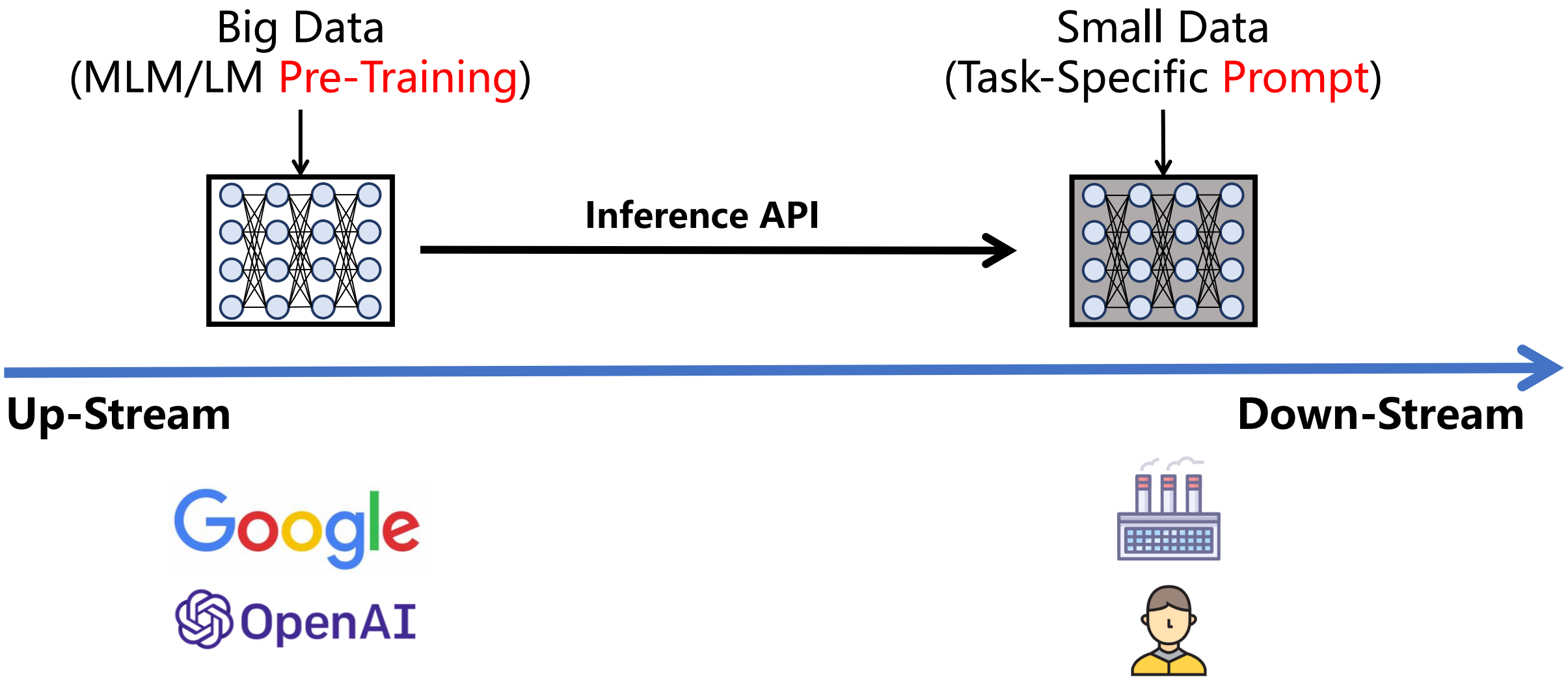
- **Servers** often do not ope
  commercial reasons
- **Users** usually do not hav

The emergent ability of LLM

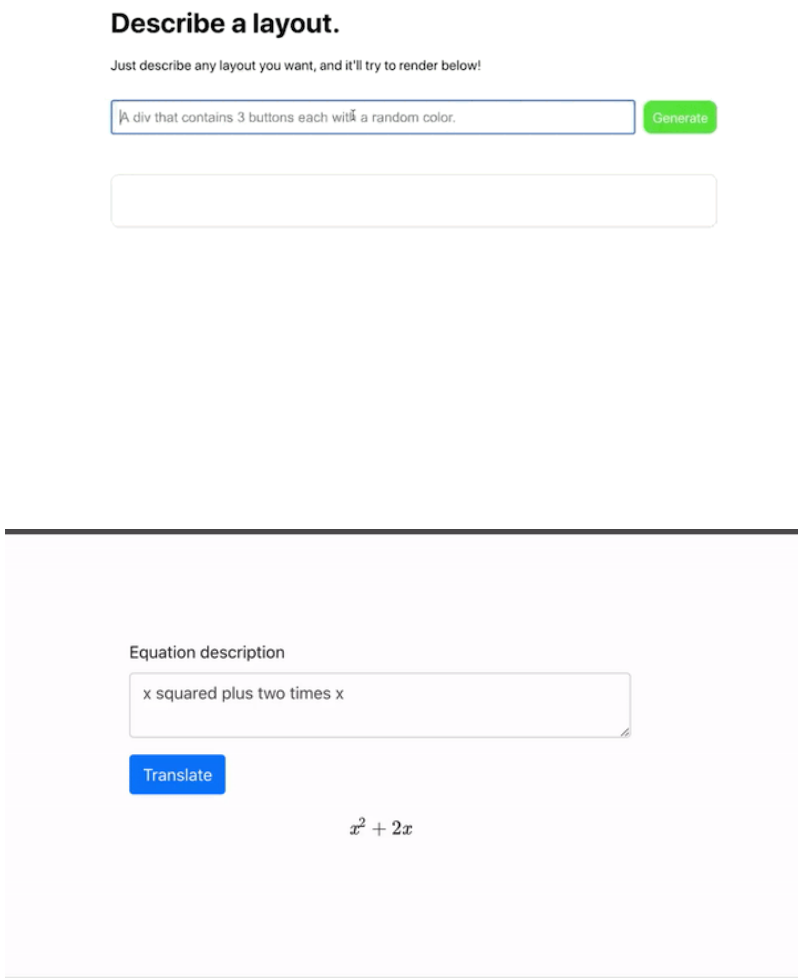
- Manually craft text prom
- In-context learning (GPT-



Language-Model-as-a-Service (LMaaS)



LMaaS has encouraged a lot of apps



Products

Select product

Collections

- New
- Popular
- Upcoming
- Requested

Categories

All	319
A/B Testing	2
Ad Generation	3
AI Copywriting	37
AI Writing Assistants	1
API Design	1
Avatars	1
Blog writing	2
Book Writing	1

New

See all →

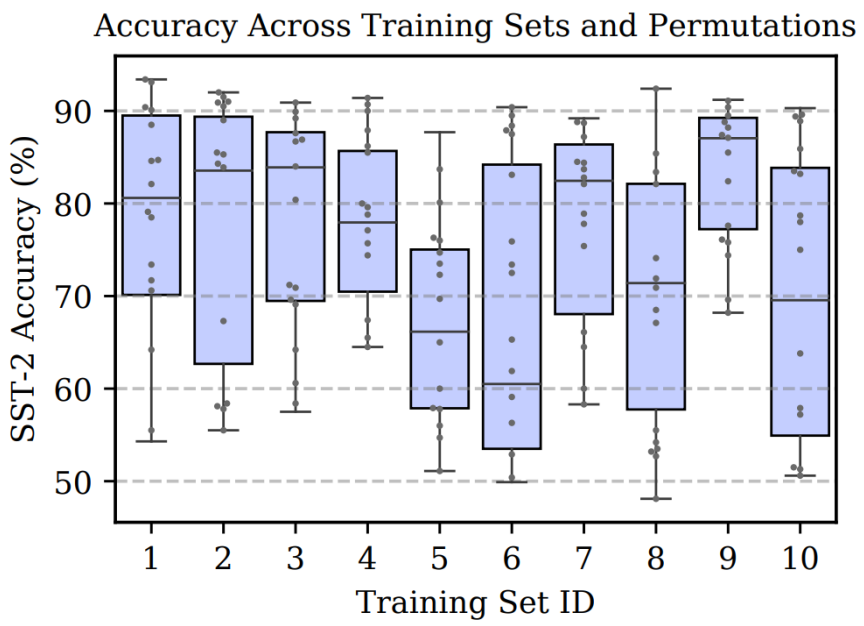
Recently added GPT-3 apps

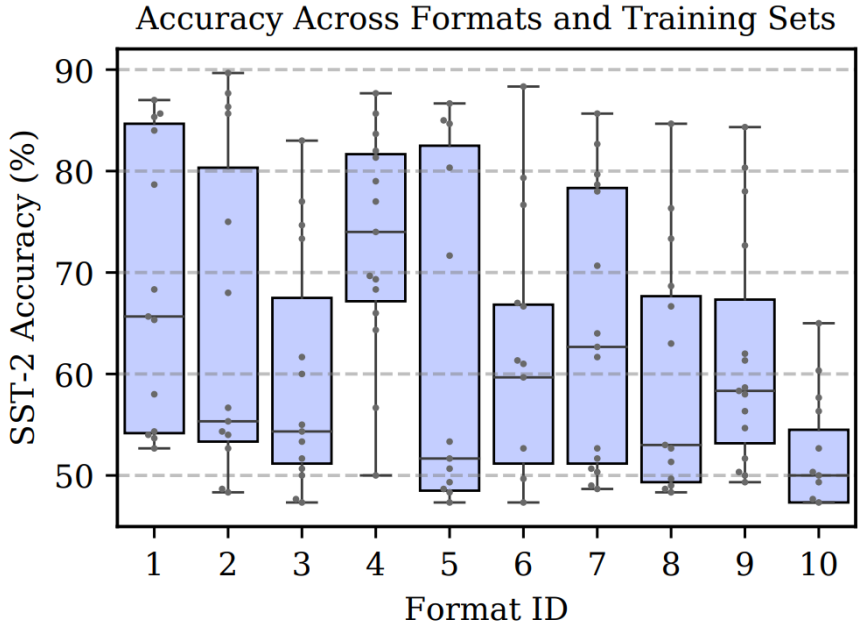
Customer Service ActiveChat.ai	Chatbots AI Buddy	Humor AI Guru
LegalTech aiLawDocs	Chatbots AskBrian	Developer Tools Azure OpenAI Service
Generative Art Botto	Image captioning ClipClap	Healthcare Curai
Code Generation DeepGenX	Summarization Delv AI	API Design Design an API with ...
Recruiting Drafted	Language Learning Duolingo	Research Assistants Elicit

https://gpt3demo.com/

However...

The performance of manual prompt and in-context learning highly depend on the choice of prompt and demonstrations, and lags far behind model tuning.





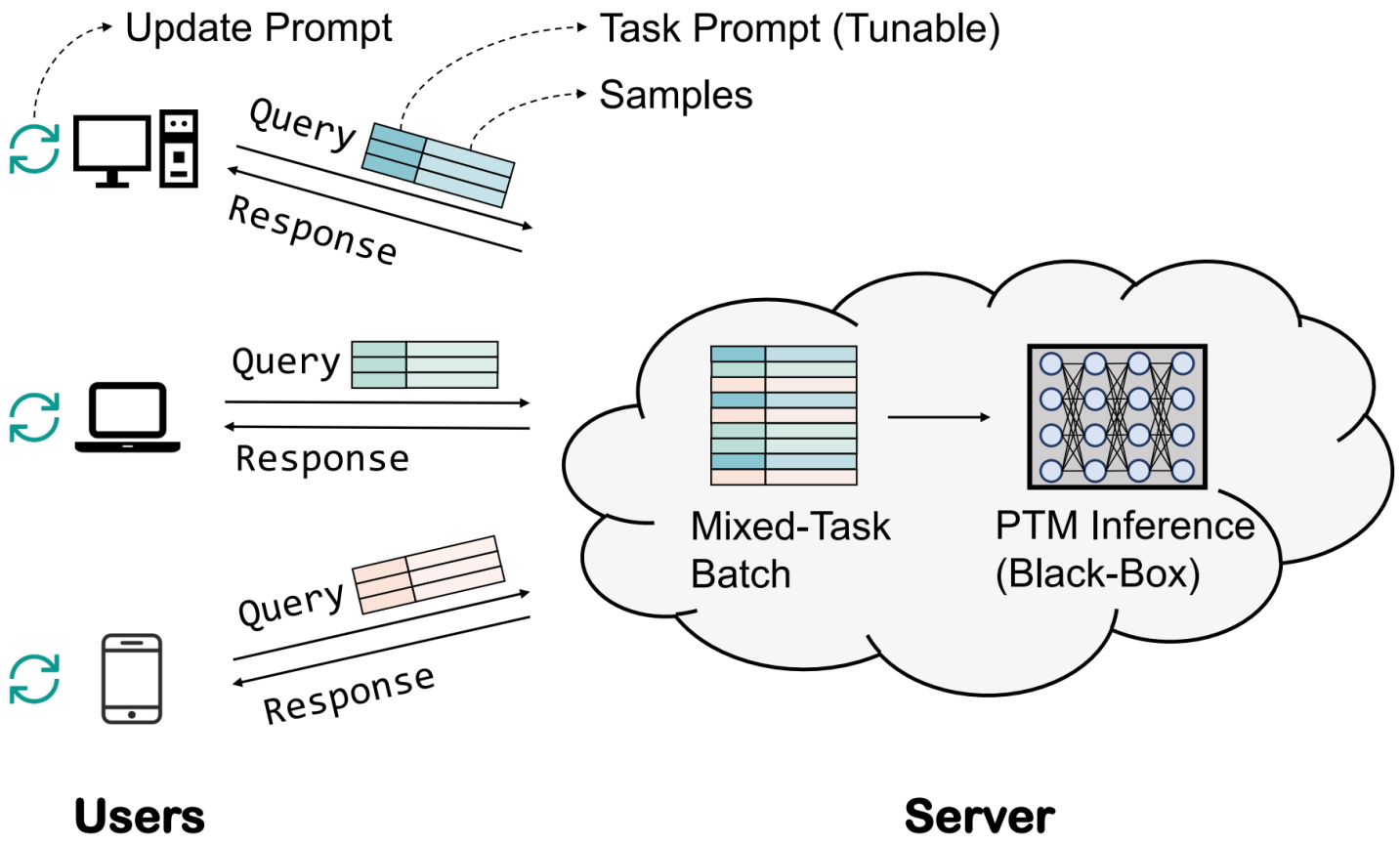
Zhao et al. Calibrate Before Use: Improving Few-Shot Performance of Language Models. ICML 2021

To make LLMs benefit more people...

Can we optimize the prompt with the API feedback? (without expensive backpropagation)

Objective:

$$\mathbf{p}^\star = \arg\min_{\mathbf{p}\in\mathcal{P}} \mathcal{L}(f(\mathbf{p}; \tilde{X}), \tilde{Y})$$



A challenge of high dimensionality

Considering optimization of the continuous prompt, the dimensionality can be tens of thousands (say we are going to optimize 50 prompt tokens, each with 1k dimensions, there are 50k parameters to be optimized.)

Derivative-free optimization (DFO) can struggle with high-dimensional problems, **except for** the case when the problem has a low intrinsic dimensionality.

Note: Intrinsic dimensionality is the minimal number of parameters needed to represent the problem

A challenge of high dimensionality

Considering optimization of the continuous prompt, the dimensionality

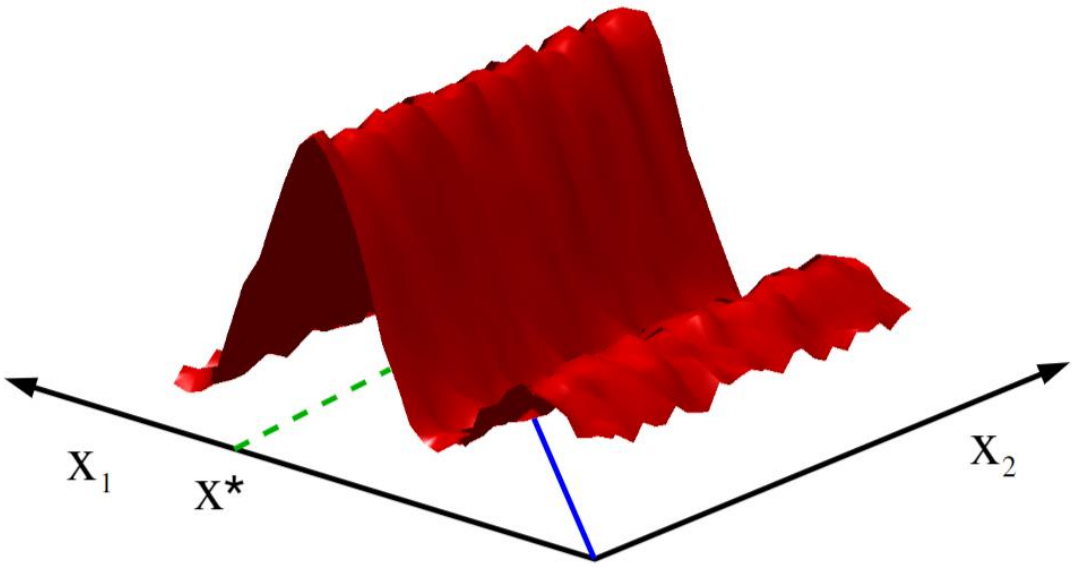
can b okens,

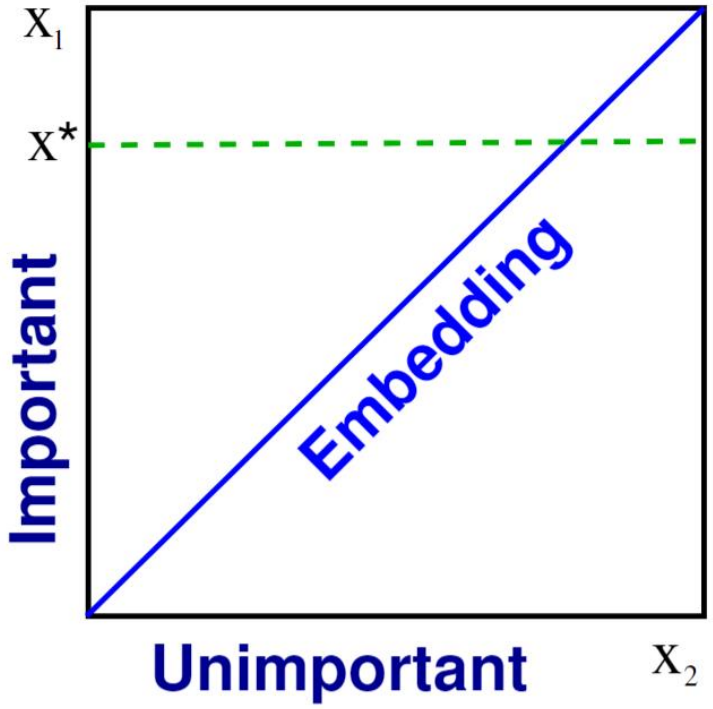
each)

Deriv nal

probl

dime

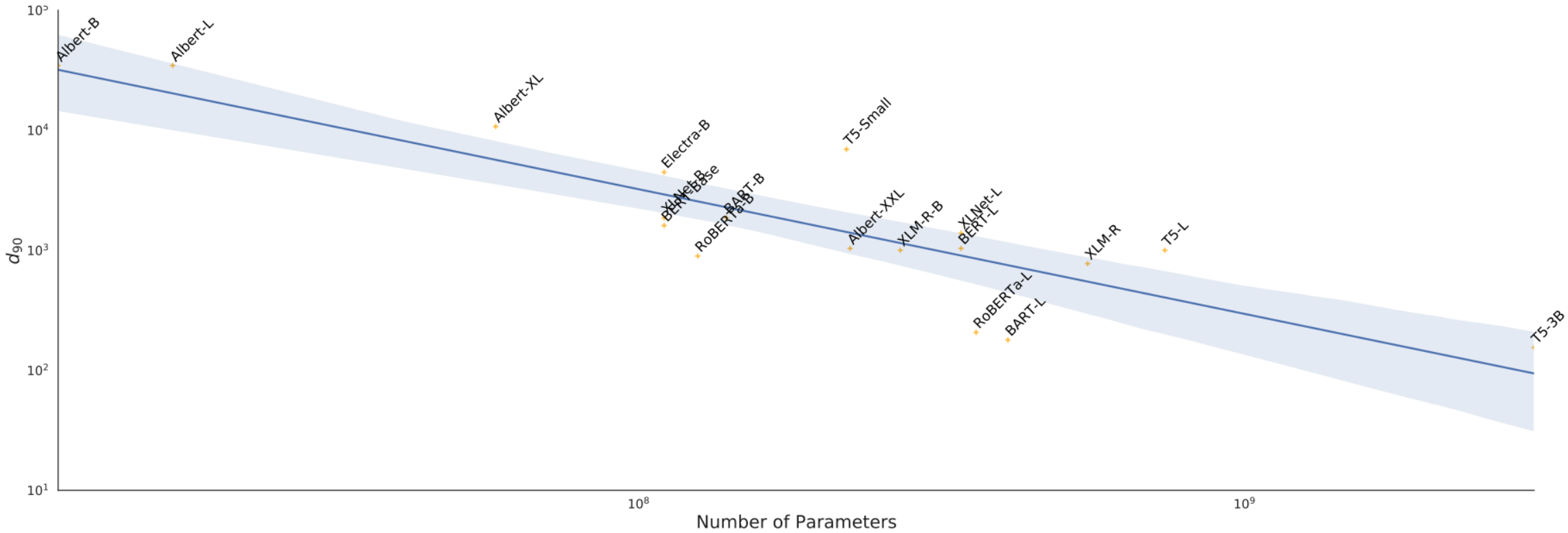




Wang et al. Bayesian optimization in a billion dimensions via random embeddings. J. Artif. Intell. Res. 2016

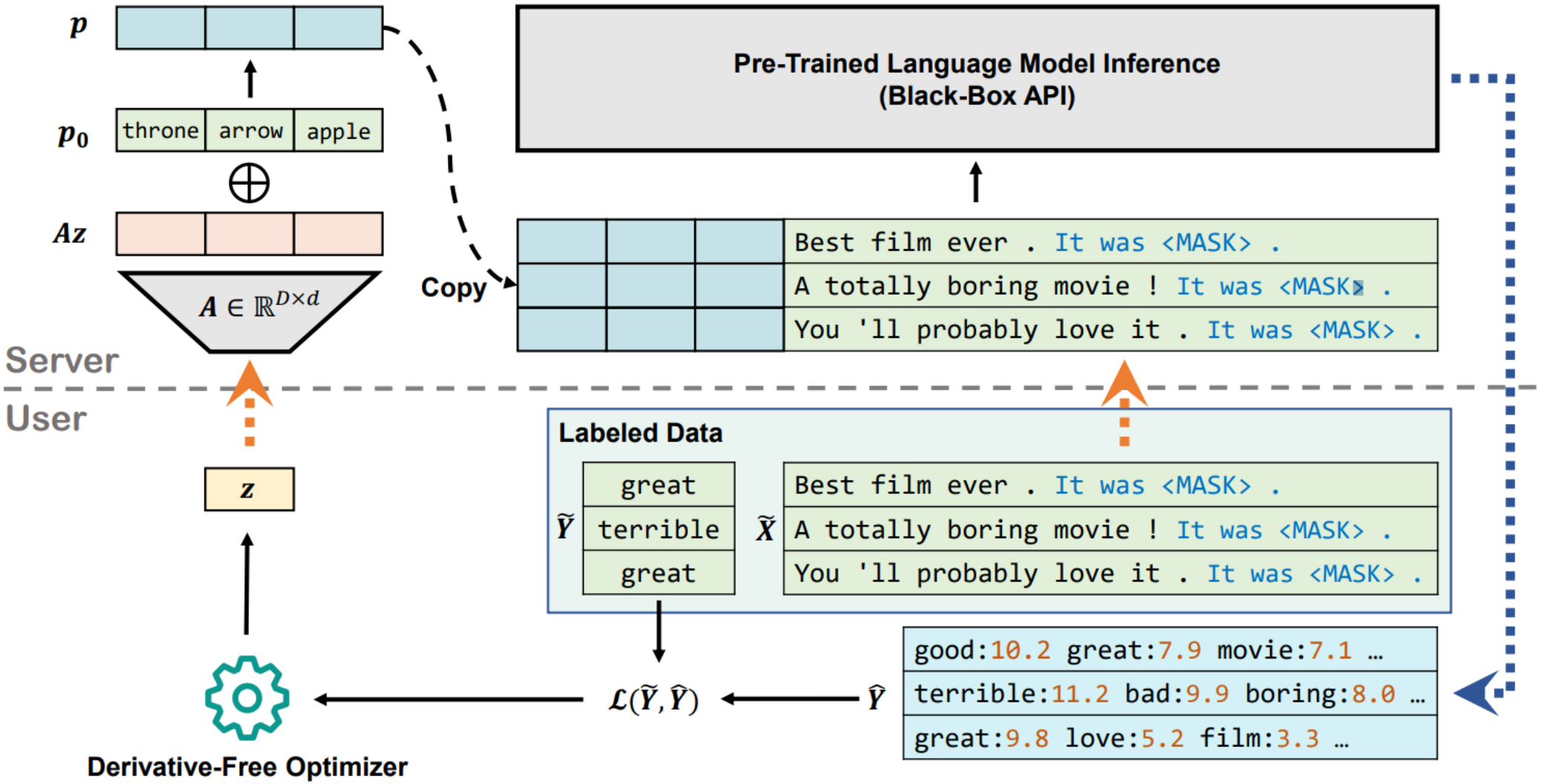
Fortunately…

LLMs have a very low intrinsic dimensionality!



Aghajanyan et al. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. ACL 2021

Black-Box Tuning



Experiments

16-shot (per class) learning with RoBERTa-large (350M)

Method	SST-2 acc	Yelp P. acc	AG's News acc	DBPedia acc	MRPC F1	SNLI acc	RTE acc	Avg.
Gradient-Based Methods								
Prompt Tuning	68.23 ±3.78	61.02 ±6.65	84.81 ±0.66	87.75 ±1.48	51.61 ±8.67	36.13 ±1.51	54.69 ±3.79	63.46
+ Pre-trained prompt	/	/	/	/	77.48 ±4.85	64.55 ±2.43	77.13 ±0.83	74.42
P-Tuning v2	64.33 ±3.05	92.63 ±1.39	83.46 ±1.01	97.05 ±0.41	68.14 ±3.89	36.89 ±0.79	50.78 ±2.28	70.47
Model Tuning	85.39 ±2.84	91.82 ±0.79	86.36 ±1.85	97.98 ±0.14	77.35 ±5.70	54.64 ±5.29	58.60 ±6.21	78.88
Gradient-Free Methods								
Manual Prompt	79.82	89.65	76.96	41.33	67.40	31.11	51.62	62.56
In-Context Learning	79.79 ±3.06	85.38 ±3.92	62.21 ±13.46	34.83 ±7.59	45.81 ±6.67	47.11 ±0.63	60.36 ±1.56	59.36
Feature-MLP	64.80 ±1.78	79.20 ±2.26	70.77 ±0.67	87.78 ±0.61	68.40 ±0.86	42.01 ±0.33	53.43 ±1.57	66.63
Feature-BiLSTM	65.95 ±0.99	74.68 ±0.10	77.28 ±2.83	90.37 ±3.10	71.55 ±7.10	46.02 ±0.38	52.17 ±0.25	68.29
Black-Box Tuning	89.56 ±0.25	91.50 ±0.16	81.51 ±0.79	87.80 ±1.53	61.56 ±4.34	46.58 ±1.33	52.59 ±2.21	73.01
+ Pre-trained prompt	/	/	/	/	75.51 ±5.54	83.83 ±0.21	77.62 ±1.30	**83.90**

Experiments

Detailed comparison on SST-2 and AG News

	Deployment-Efficient	**As-A-Service**	**Test Accuracy**	**Training Time**	**Memory Footprint** User	Server	**Upload** per query	**Download** per query
SST-2 (max sequence length: 47)								
Prompt Tuning	√	×	72.6	15.9 mins	-	5.3 GB	-	-
Model Tuning	×	×	87.8	9.8 mins	-	7.3 GB	-	-
Feature-MLP	√	√	63.8	7.0 mins	20 MB	2.8 GB	4 KB	128 KB
Feature-BiLSTM	√	√	66.2	9.3 mins	410 MB	2.8 GB	4 KB	6016 KB
Black-Box Tuning	√	√	89.4	10.1 (6.1*) mins	30 MB	3.0 GB	6 KB	0.25 KB
AG's News (max sequence length: 107)								
Prompt Tuning	√	×	84.0	30.2 mins	-	7.7 GB	-	-
Model Tuning	×	×	88.4	13.1 mins	-	7.3 GB	-	-
Feature-MLP	√	√	71.0	13.5 mins	20 MB	3.6 GB	20 KB	256 KB
Feature-BiLSTM	√	√	73.1	19.7 mins	500 MB	3.6 GB	20 KB	27392 KB
Black-Box Tuning	√	√	82.6	21.0 (17.7*) mins	30 MB	4.6 GB	22 KB	1 KB

A Gradient-Free Future for LLMs?

Limitations of black-box tuning:

- Slow convergence on many-label classification (e.g., DBPedia)
- Requirement of prompt pre-training (gradient) on difficult tasks (e.g., SNLI)

Current version of black-box tuning is just a lower bound:

- Prompt/verbalizer engineering, prompt ensemble, prompt pre-training…

Code is publicly available!

- https://github.com/txsun1997/Black-Box-Tuning

ICML
International Conference
On Machine Learning

Thanks!

Tianxiang Sun

School of Computer Science, Fudan University

https://txsun1997.github.io/