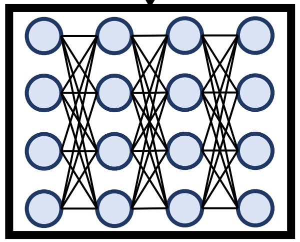
Black-Box Tuning for Language-Model-as-a-Service

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, Xipeng Qiu
Fudan University
txsun19@fudan.edu.cn

19 Jan 2022

The Old Paradigm: Pre-Training + Fine-Tuning

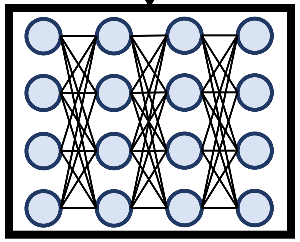
Big Data
(MLM/LM Pre-Training)



Open Source

Download

Small Data
(Task-Specific Fine-Tuning)



Up-Stream

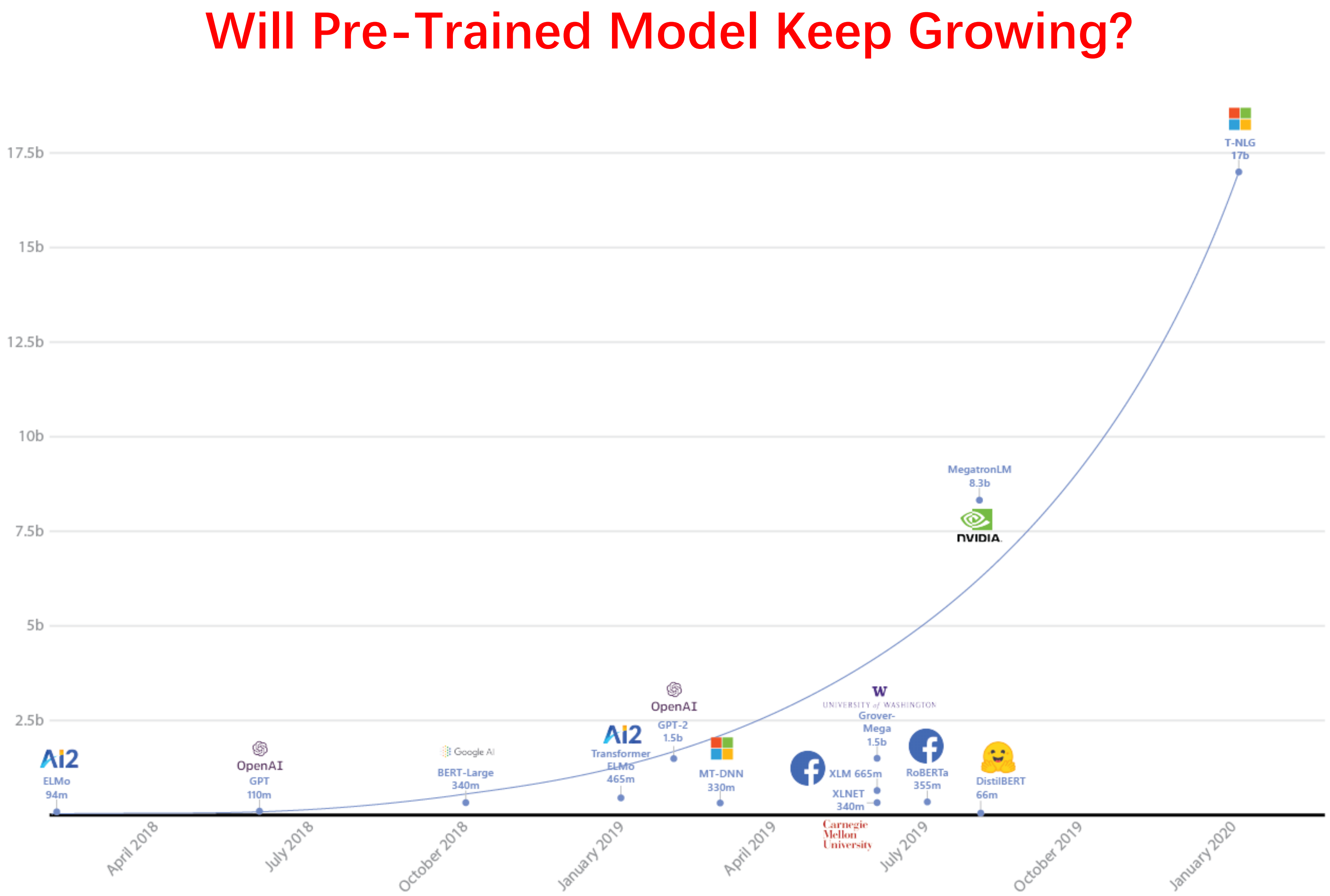
Down-Stream

Google

OpenAI

HUAWEI

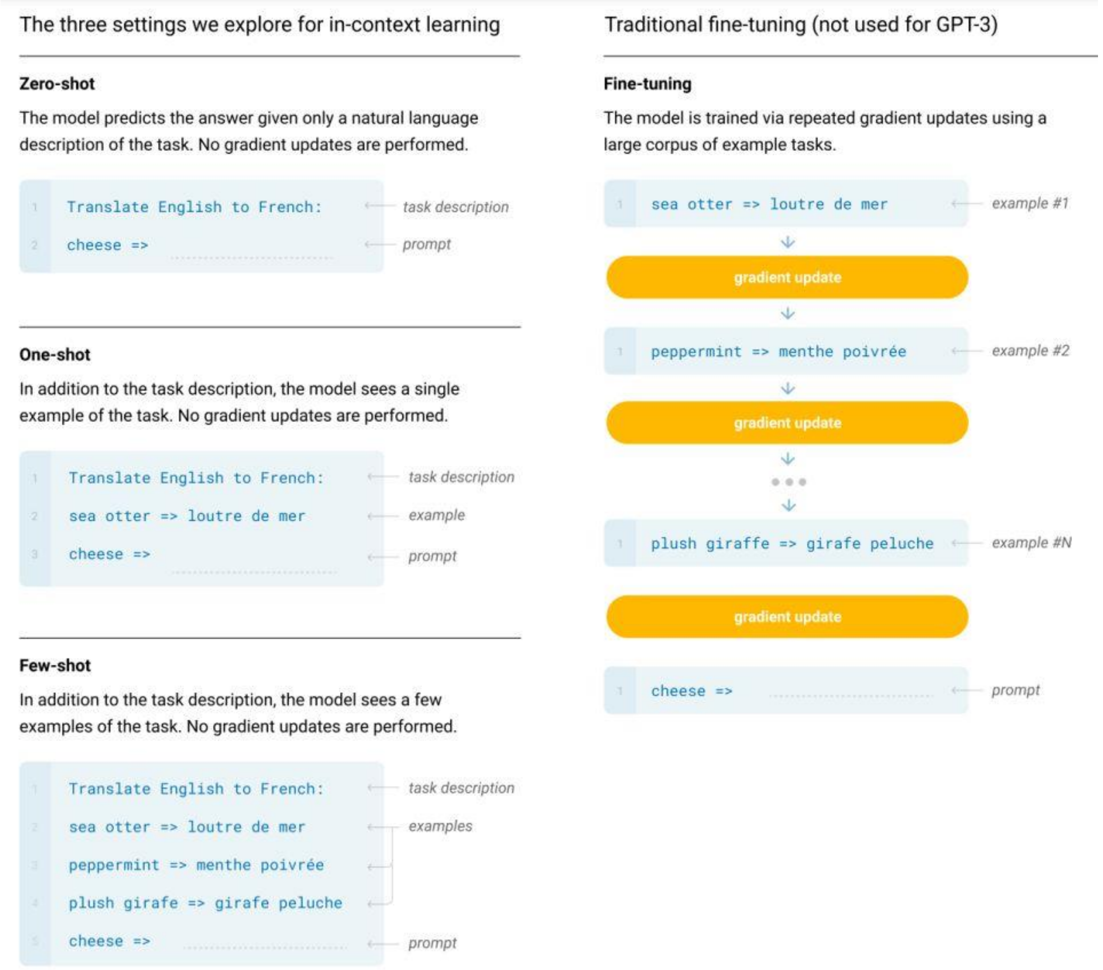
The Era of Big Models



The New Paradigm: LMaaS

- **LMaaS: Language-Model-as-a-Service**
- **A Milestone: GPT-3**
  - **Why in-context learning?**

1. Generalization of big models (one for all)
2. Backpropagation is expensive
3. Commercial use



The New Paradigm: LMaaS

- **LMaaS: Language-Model-as-a-Service**

Why did OpenAI choose to release an API instead of open-sourcing the models?

There are three main reasons we did this. First, commercializing the technology helps us pay for our ongoing AI research, safety, and policy efforts.

Second, many of the models underlying the API are very large, taking a lot of expertise to develop and deploy and making them very expensive to run. This makes it hard for anyone except larger companies to benefit from the underlying technology. We're hopeful that the API will make powerful AI systems more accessible to smaller businesses and organizations.

Third, the API model allows us to more easily respond to misuse of the technology. Since it is hard to predict the downstream use cases of our models, it feels inherently safer to release them via an API and broaden access over time, rather than release an open source model where access cannot be adjusted if it turns out to have harmful applications.

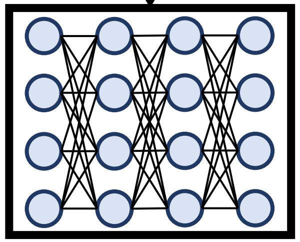
The New Paradigm: LMaaS

- **LMaaS: Language-Model-as-a-Service**

值得一提的是，张宏江强调，人工智能大模型时代的到来，为进入AI赛道提供了机会点，“超大数据+超大算力+超大模型”的大模型可应对多种任务。未来，大模型会形成类似电网的智能基础平台，像发电厂一样为全社会源源不断地供应“智力源”。从“大炼模型”到“炼大模型”，智能研究院成为“人工智能大模型”发展转折的推动者，悟道系列大模型成为这一进程中的标志性成果。

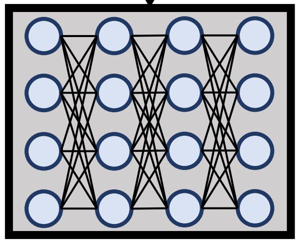
The New Paradigm: LMaaS

Big Data
(MLM/LM Pre-Training)



API

Small Data
(Task-Specific Prompt)



Up-Stream

Down-Stream

Google

OpenAI

HUAWEI

The New Paradigm: LMaaS

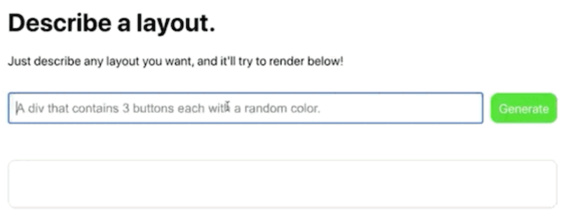
- **GPT-3 Pricing**

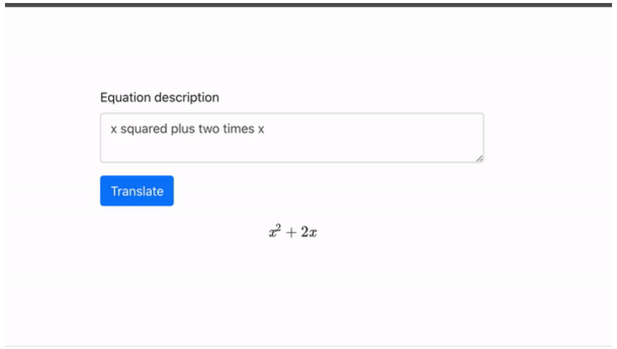
Per-model prices

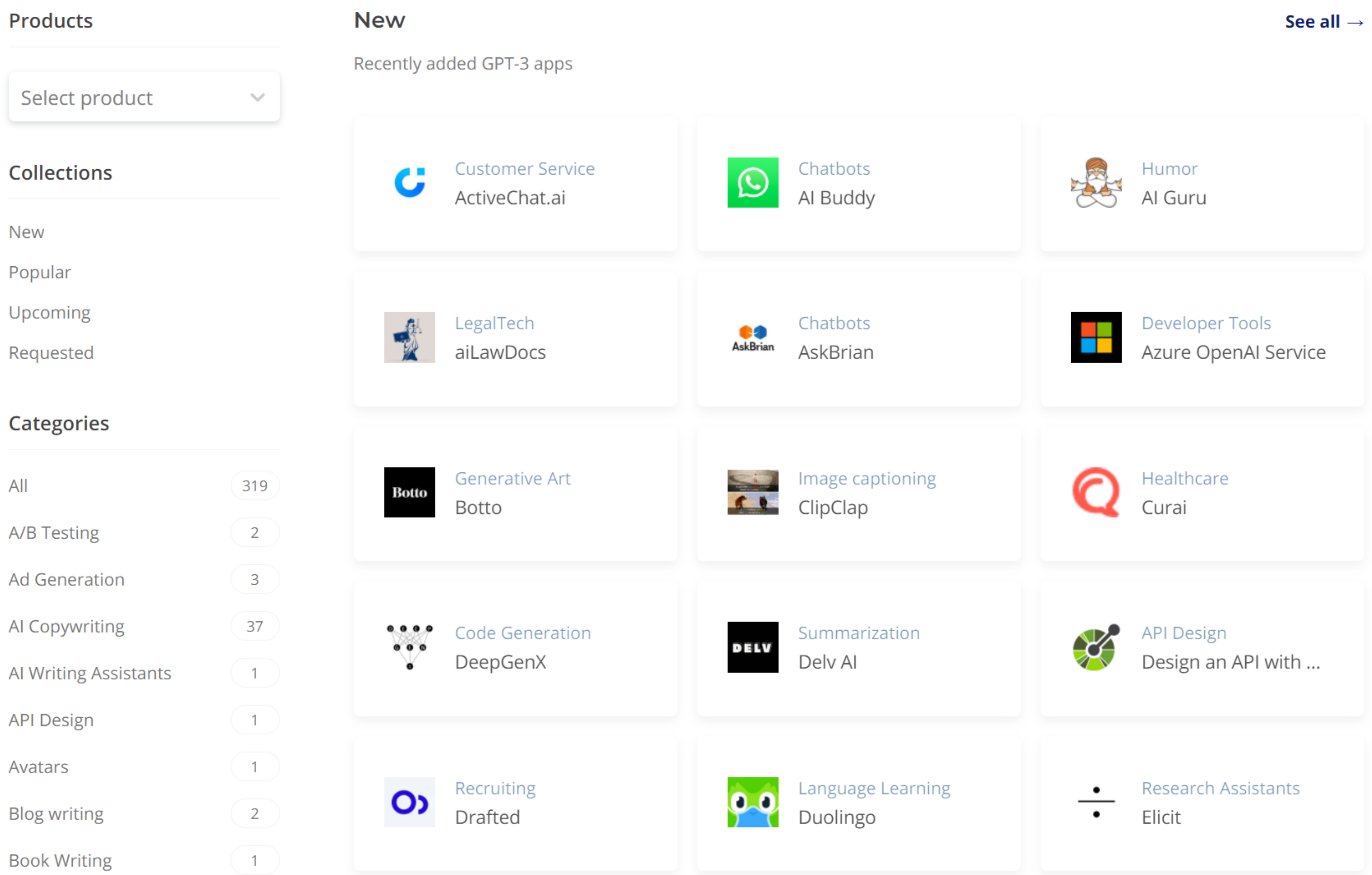
Ada Fastest	Babbage	Curie	Davinci Most powerful
$0.0008 /1K tokens	$0.0012 /1K tokens	$0.0060 /1K tokens	$0.0600 /1K tokens

The New Paradigm: LMaaS

- **GPT-3 Demos**







https://gpt3demo.com/

The New Paradigm: LMaaS

- **In China**

赛题说明

INTRODUCE

本次大赛的主题为基于悟道2.0大模型的创新应用开发，面向在校大学生、企业工程师、科研工作者等全球开发者全面开放。参赛选手需要依据对悟道能力的理解，结合社会关注的热点选择健康医疗、教育学习、社交生活、效率工具、环境自然或其他具有社会价值、产业价值的相关主题，提交一个潜在的智能应用方案并上线应用。

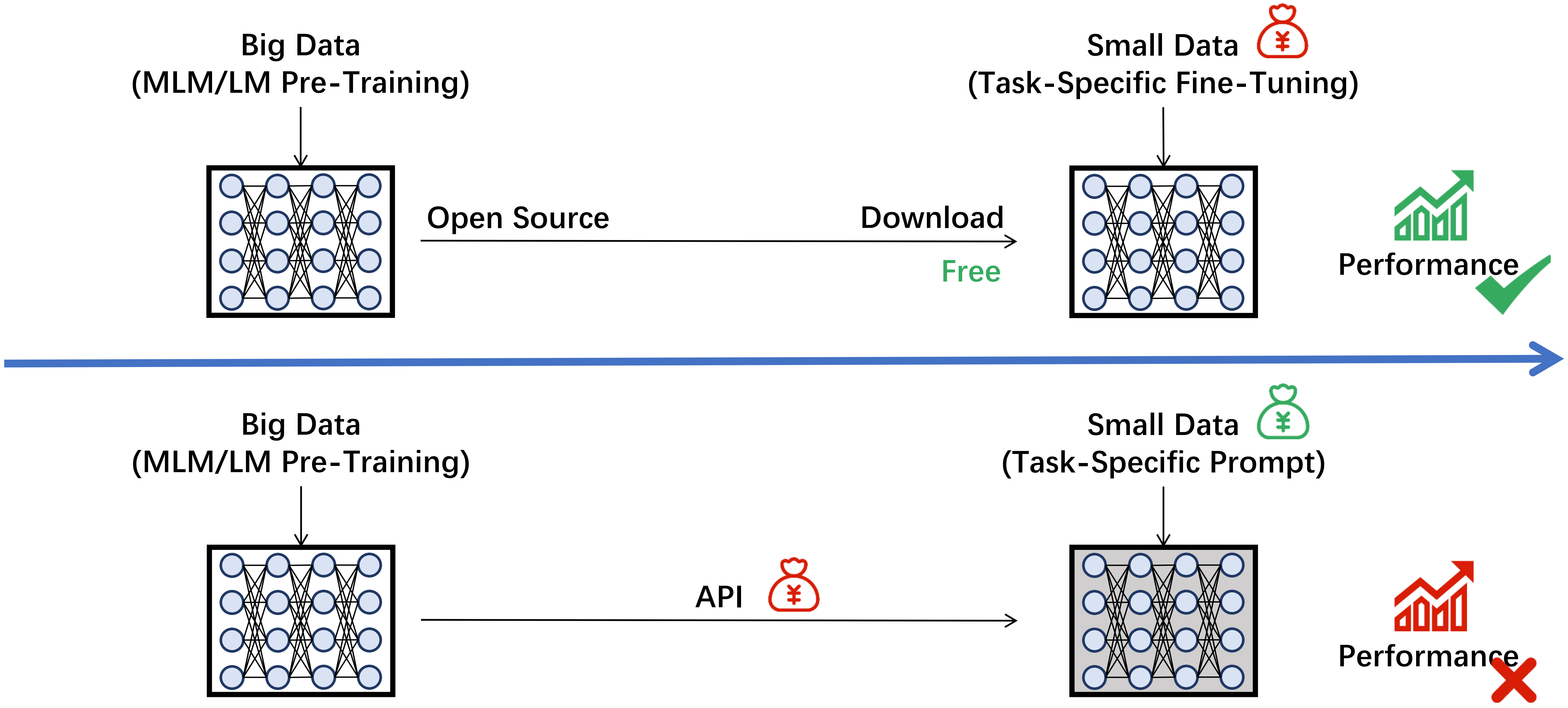
API 文档说明

本次比赛共提供 9 个应用的 API 接口，每支队伍每天访问次数有限。

- ✓ CogView API 文档，100 次 / 天。
- ✓ 藏头诗 API 文档，100 次 / 天。
- ✓ 获取图像的特征向量，1000 次 / 天。
- ✓ 获取文本的特征向量，1000 次 / 天。
- ✓ 快速写诗 API 文档，100 次 / 天。
- ✓ 宋词 API 文档，100 次 / 天。
- ✓ 问答 API 文档，100 次 / 天。
- ✓ 写诗 API 文档，100 次 / 天。
- ✓ 新闻 API 文档，100 次 / 天。

https://www.biendata.xyz/wudao/

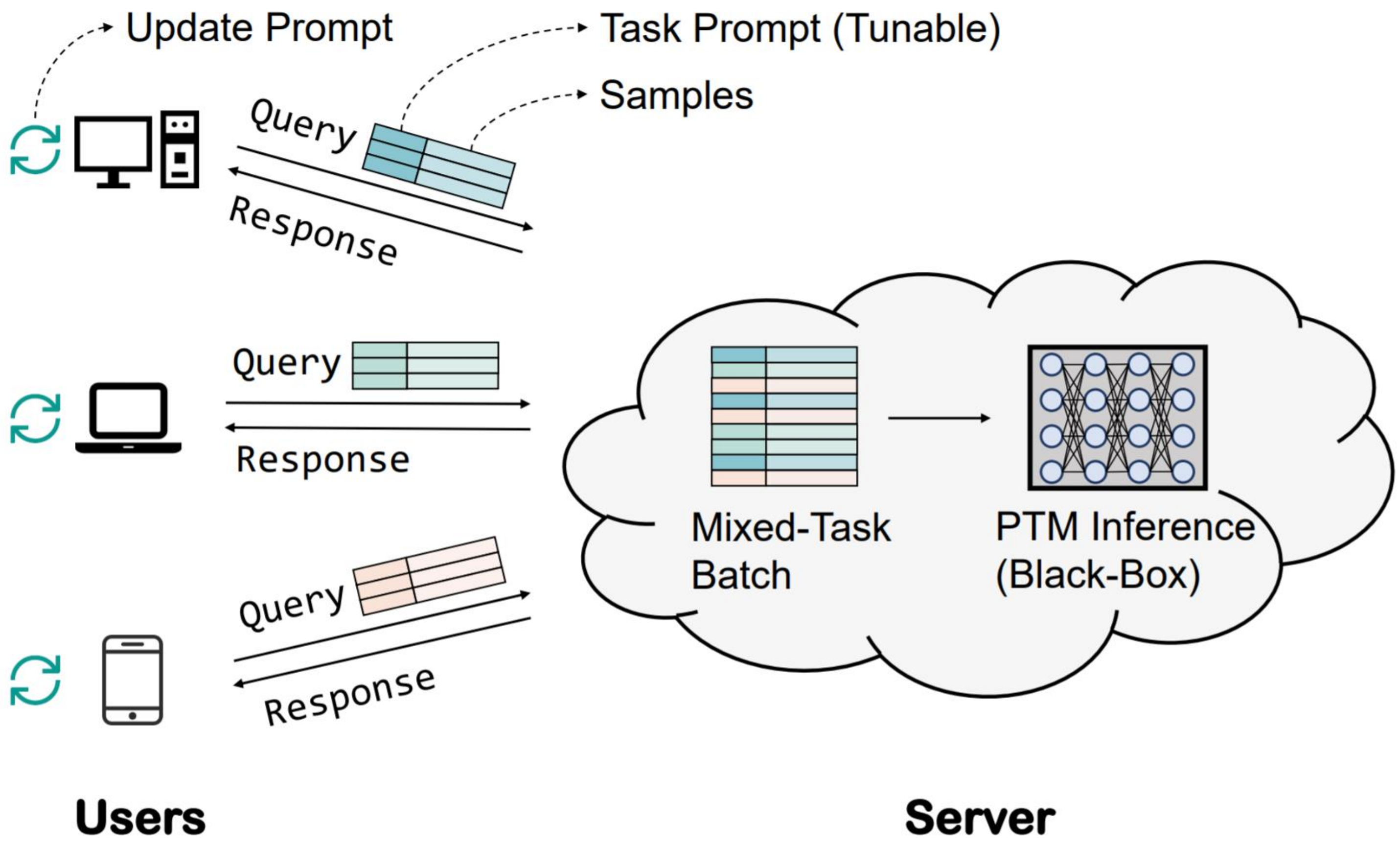
Grounding From the Cloud



Performance is the Key for grounding! (Who are the users?)

Grounding From the Cloud

- **Can we optimize the prompt with API responses?**

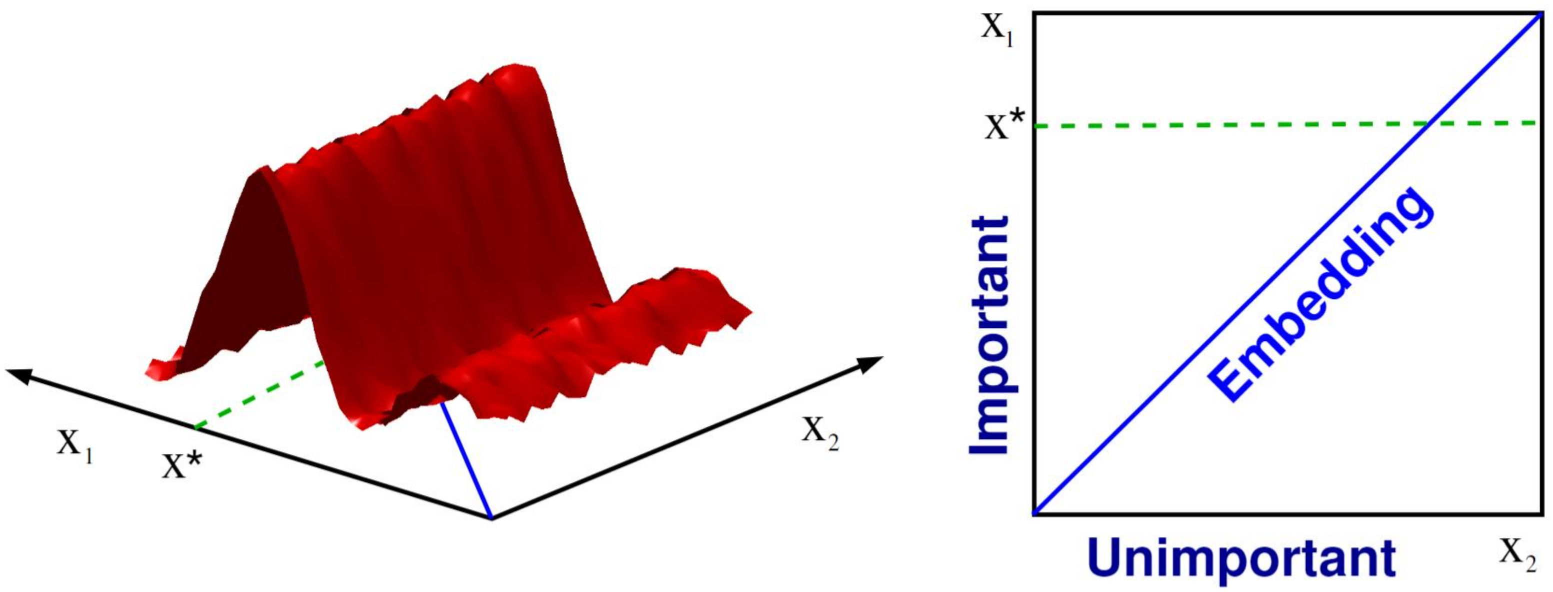


Grounding From the Cloud

- **Can we optimize the prompt with API responses?**
  - Yes, we can use derivative-free optimization (DFO)!
- **Then I tried a lot of methods to achieve this…**
  - Surrogate models
  - Synthetic gradient
  - Gradient distillation
  - Learning to learn
  - …
- **Finally I found these methods failed due to the dimension**
  - Even for prompt tuning, the dimension we need to optimize is >20x1024=20480
  - However, current DFO cannot well-handle optimization with >1k parameters ☹

Grounding From the Cloud

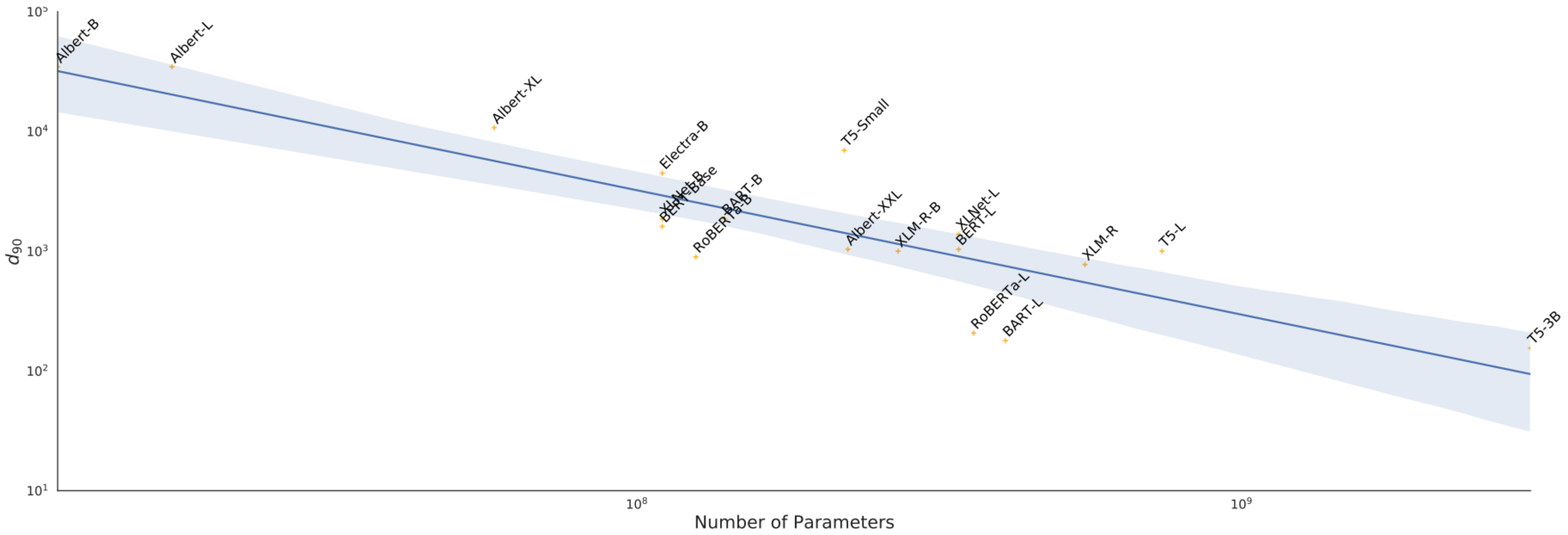
- **Fortunately…**
- **On the one hand, we can use DFO to solve a high-dimensional optimization as long as its intrinsic dimensionality is low**



Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research. 2016*

Grounding From the Cloud

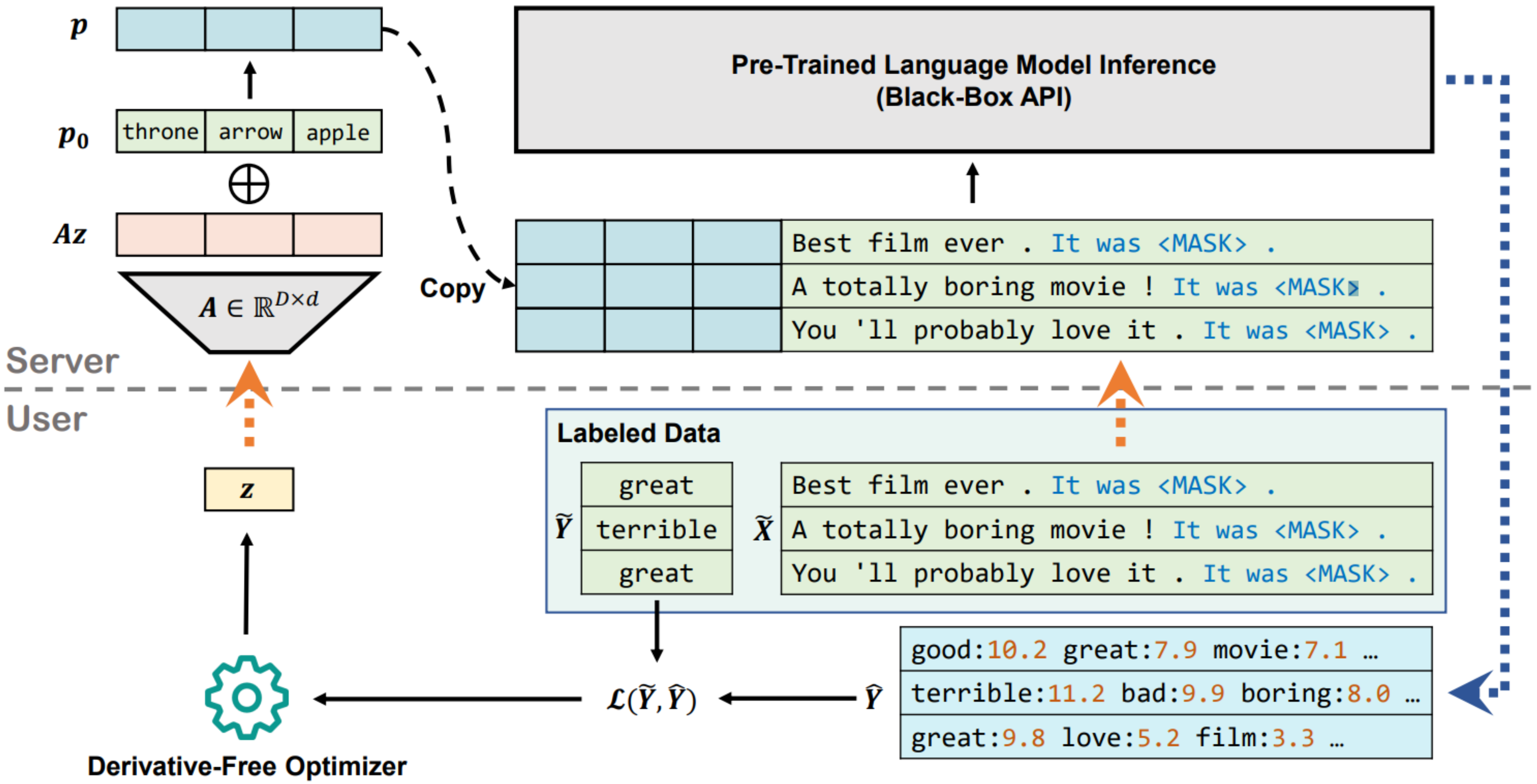
- **Fortunately…**
- **On the one hand, we can use DFO to solve a high-dimensional optimization as long as its intrinsic dimensionality is low**
- **On the other hand, large-scale pre-trained models have a very low intrinsic dimensionality.**



Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *ACL 2021*

Black-Box Tuning

- Combine the two hands, we have an implementation…



Black-Box Tuning

- **The CMA Evolution Strategy**

The CMA-ES (**E**volution **S**trategy with **C**ovariance **M**atrix **A**daptation)

Consider $P^{(t)} = \mathcal{N}\left(\boldsymbol{\mu}^{(t)}, {\sigma^{(t)}}^2 \boldsymbol{C}^{(t)}\right)$ where $\boldsymbol{\mu}^{(t)} \in \mathbb{R}^n$, $\sigma^{(t)} \in \mathbb{R}_+$, $\boldsymbol{C}^{(t)} \in \mathbb{R}^{n\times n}$

- $\boldsymbol{\mu}^{(t)} \to \boldsymbol{\mu}^{(t+1)}$: Maximum likelihood update, i.e. $P(\boldsymbol{x}^{(t)}_{\text{selected}} | \boldsymbol{\mu}^{(t+1)}) \to \max$
- $\boldsymbol{C}^{(t)} \to \boldsymbol{C}^{(t+1)}$: Maximum likelihood update, i.e. $P(\frac{\boldsymbol{x}^{(t)}_{\text{selected}} - \boldsymbol{\mu}^{(t)}}{\sigma^{(t)}} | \boldsymbol{C}^{(t+1)}) \to \max$, under consideration of prior $\boldsymbol{C}^{(t)}$ (otherwise $\boldsymbol{C}^{(t+1)}$ becomes singular).
- $\sigma^{(t)} \to \sigma^{(t+1)}$: Update to achieve conjugate perpendicularity, i.e. conceptually $(\boldsymbol{\mu}^{(t+2)} - \boldsymbol{\mu}^{(t+1)})^{\mathrm{T}} {\boldsymbol{C}^{(t)}}^{-1} (\boldsymbol{\mu}^{(t+1)} - \boldsymbol{\mu}^{(t)}) / {\sigma^{(t+1)}}^2 \to 0$

Black-Box Tuning

- **Experimental Setup**

Hyper-parameter	Default
Prompt length (L)	50
Subspace dimension (d)	500
Population size (λ)	20
Random projection ($\mathbf{A}$)	Uniform
Loss function $\mathcal{L}$	Cross Entropy
Budget (# of API calls)	8000

Category	Dataset	$\lvert\mathcal{Y}\rvert$	|Train|	|Test|	Type	Template	Label words
single-sentence	SST-2	2	67k	0.9k	sentiment	$\langle S\rangle$. It was `[MASK]`.	great, bad
	Yelp P.	2	560k	38k	sentiment	$\langle S\rangle$. It was `[MASK]`.	great, bad
	AG's News	4	120k	7.6k	topic	`[MASK]` News: $\langle S\rangle$	World, Sports, Business, Tech
	DBPedia	14	560k	70k	topic	[Category: `[MASK]`] $\langle S\rangle$	Company, Education, Artist, Athlete, Office, Transportation, Building, Natural, Village, Animal, Plant, Album, Film, Written
sentence-pair	MRPC	2	3.7k	0.4k	paraphrase	$\langle S_1\rangle$? `[MASK]` , $\langle S_2\rangle$	Yes, No
	RTE	2	2.5k	0.3k	NLI	$\langle S_1\rangle$? `[MASK]` , $\langle S_2\rangle$	Yes, No
	SNLI	3	549k	9.8k	NLI	$\langle S_1\rangle$? `[MASK]` , $\langle S_2\rangle$	Yes, Maybe, No

Black-Box Tuning

- **Experimental Results (16-shot)**

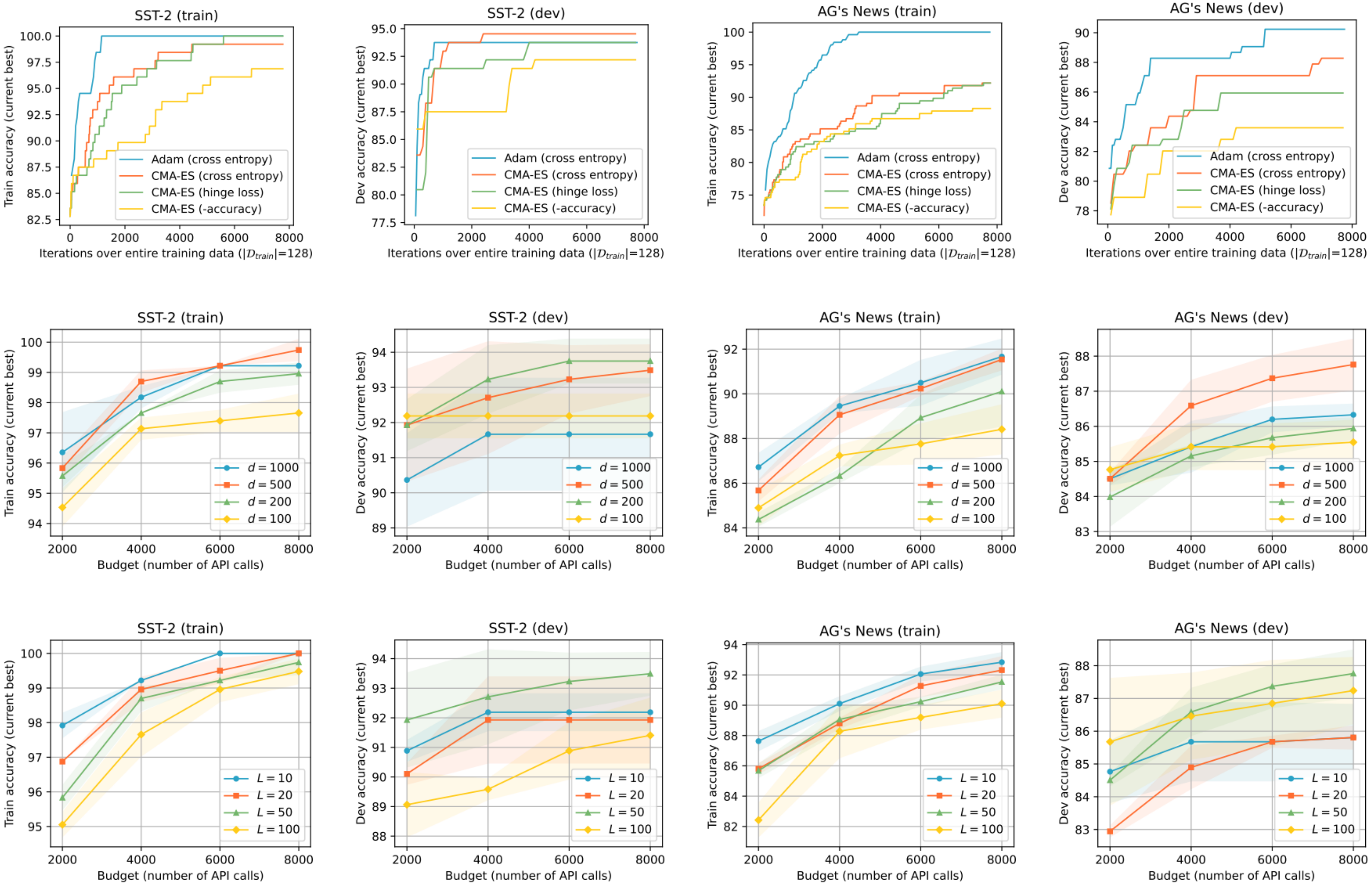
Method	**SST-2** acc	**Yelp P.** acc	**AG's News** acc	**DBPedia** acc	**MRPC** F1	**SNLI** acc	**RTE** acc	**Avg.**
				Gradient-Based Methods				
Prompt Tuning	68.23 ±3.78	61.02 ±6.65	84.81 ±0.66	87.75 ±1.48	77.48 ±4.85	64.55 ±2.43	77.13 ±0.83	74.42
Model Tuning	85.39 ±2.84	91.82 ±0.79	86.36 ±1.85	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-Linear	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	75.51 ±5.54	83.83 ±0.21	77.62 ±1.30	**83.90**

Black-Box Tuning

- **Experimental Results (16-shot)**

	Deployment-Efficient	As-A-Service	Test Accuracy	Training Time	Memory Footprint User	Memory Footprint 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-Linear	√	√	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 mins	9 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-Linear	√	√	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	23.3 mins	9 MB	4.6 GB	22 KB	1 KB

Black-Box Tuning



What's Next?

- **This paper is just a lower bound**
- **Optimization**
  - Pre-trained prompt embedding and projection matrix
  - Sequential random embedding
  - Better DFO algorithms
  - ...
- **Prompt**
  - Better prompt/verbalizer engineering
  - Prompt ensemble
  - ...

Thanks!

https://arxiv.org/abs/2201.03514

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