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

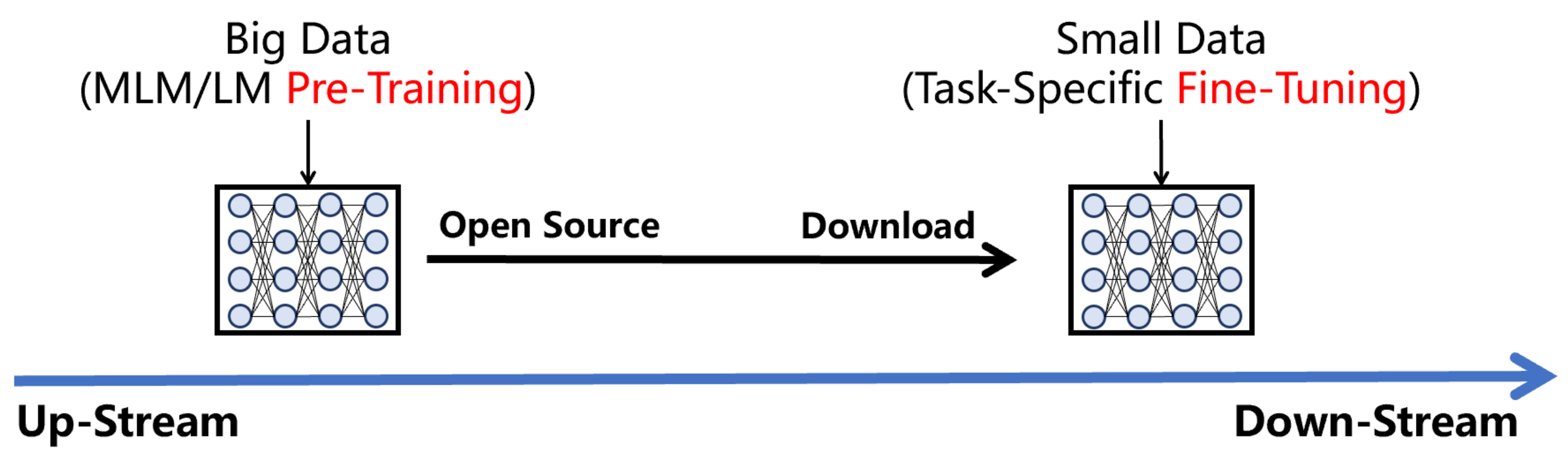
Tianxiang Sun[1], Yunfan Shao[1], Hong Qian[2], Xuanjing Huang[1], Xipeng Qiu[1,3]

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

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

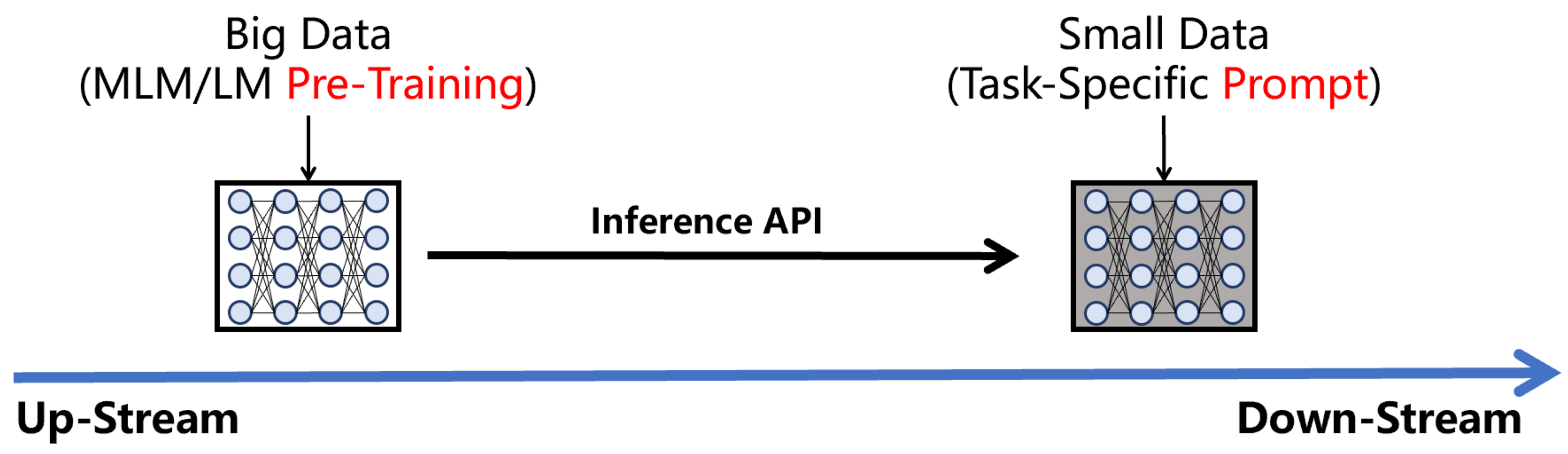
☐ Pre-training, then fine-tuning

Pre-training then fine-tuning is a promising paradigm to utilize the power of small/normal size pre-trained language models, achieving state-of-the-art performance on a wide range of downstream tasks.



☐ Language-Model-as-a-Service (LMaaS)

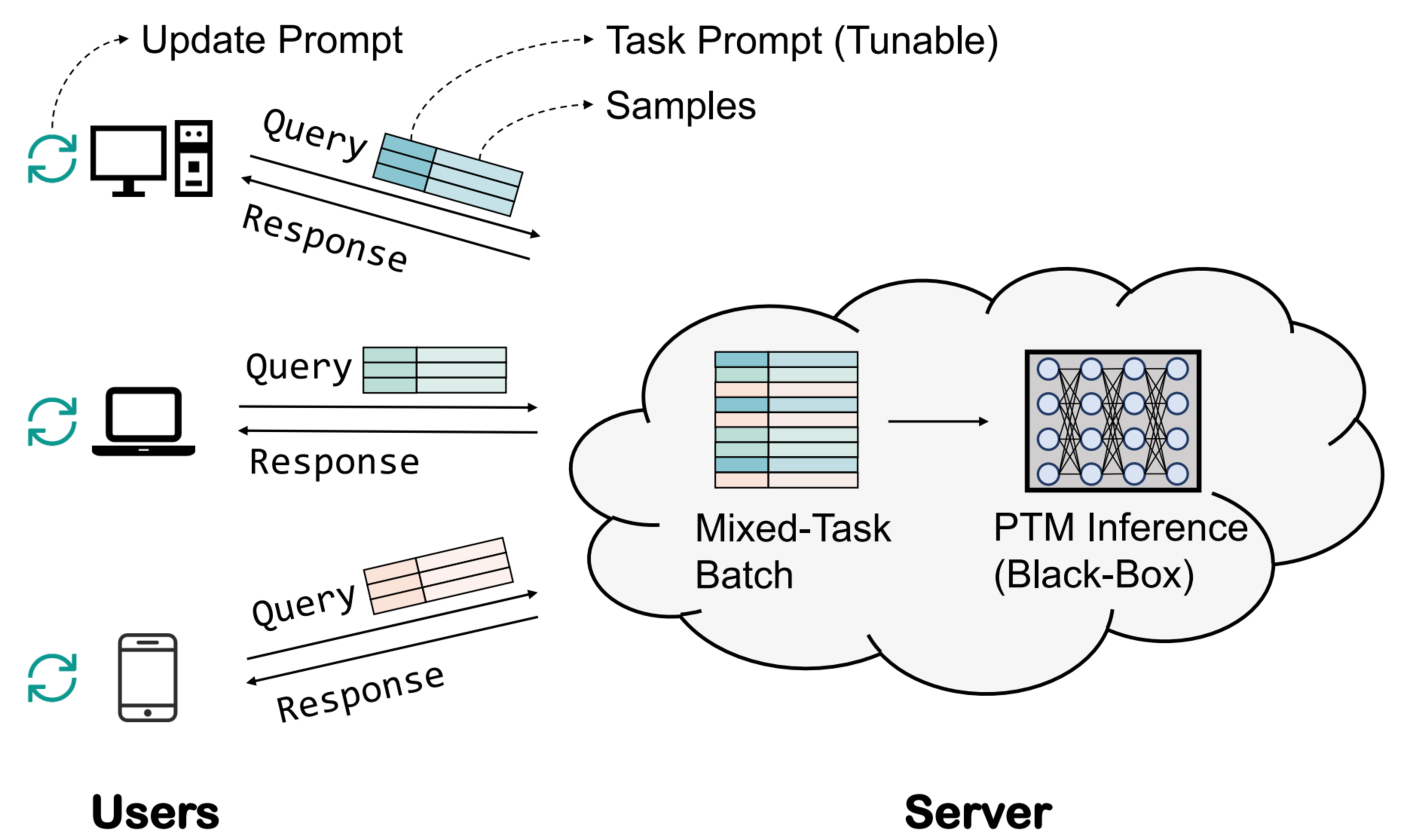
Due to commercial concerns and expensive tuning cost, large language models (LLMs) such as GPT-3 are usually released as a service instead of open-sourcing model weights. Users can only access their inference APIs.



How to make LLMs benefit more people when we only have access to model inference API?

Black-Box Tuning

☐ Can we optimize task-specific prompts by only accessing model output probability?



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

☐ Challenge of high dimensionality

The continuous prompt to be optimized contains **tens of thousands** of parameters, posing a challenge for **derivative-free optimization (DFO)**.

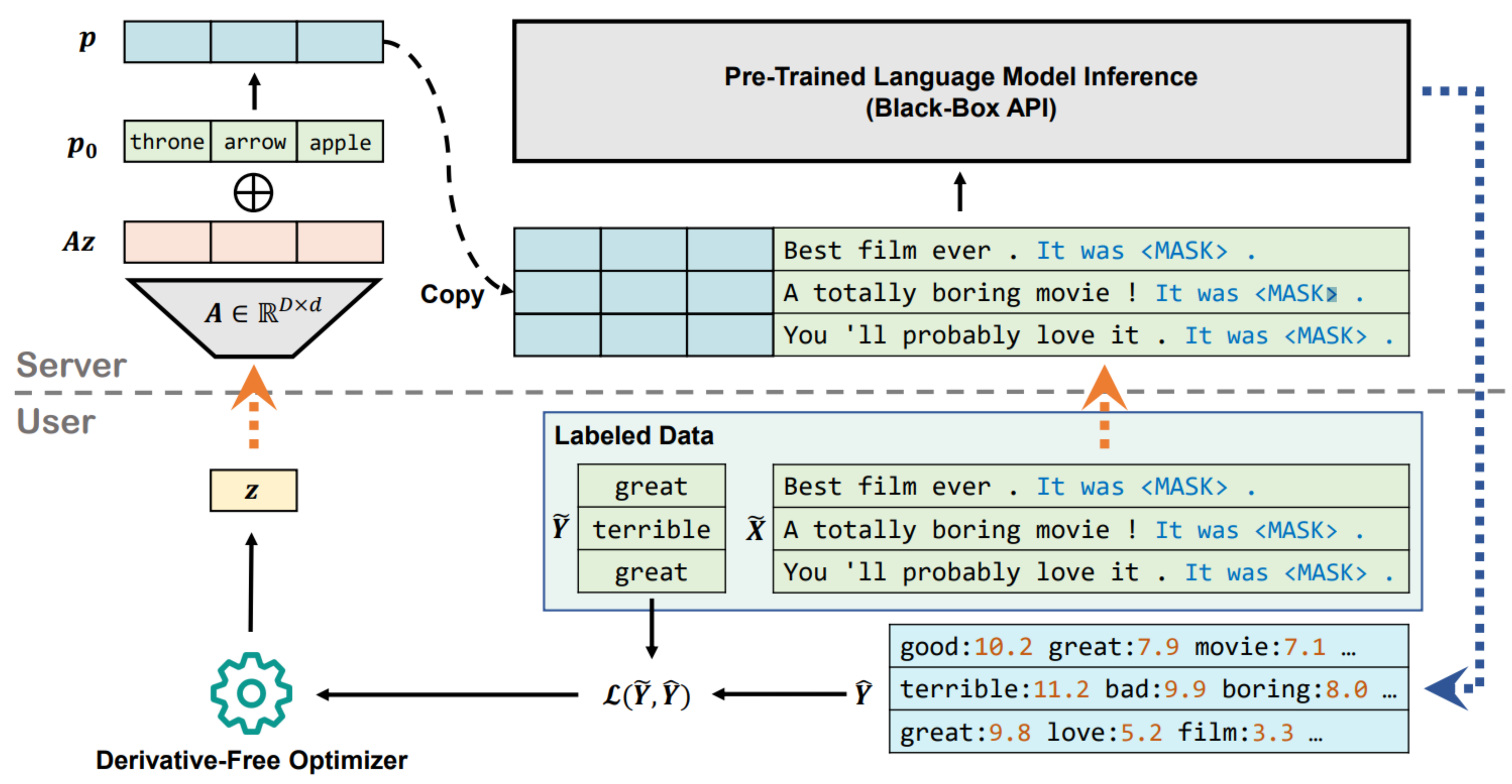
☐ Low intrinsic dimensionality of LLMs

Fortunately, it has been demonstrated that LLMs have a very low intrinsic dimensionality, and therefore we can perform DFO in a low-dimensional subspace via random embedding.

Thus, we recast the objective as:

$$\mathbf{z}^\star = \arg\min_{\mathbf{z}\in\mathcal{Z}} \mathcal{L}(f(\mathbf{A}\mathbf{z} + \mathbf{p}_0; \tilde{X}), \tilde{Y})$$

Overview of Approach



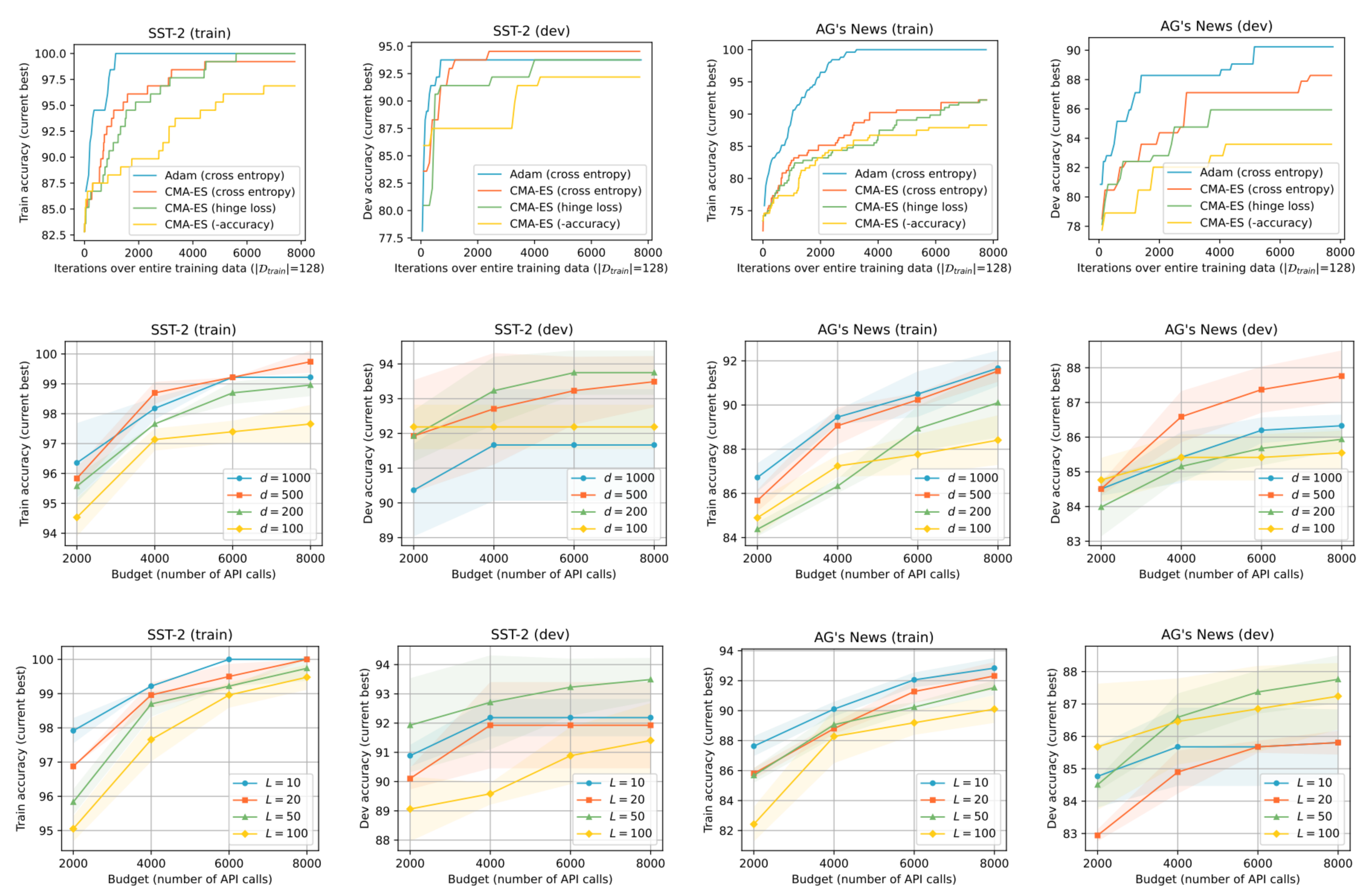
Experiments

☐ Main results under true few-shot setting

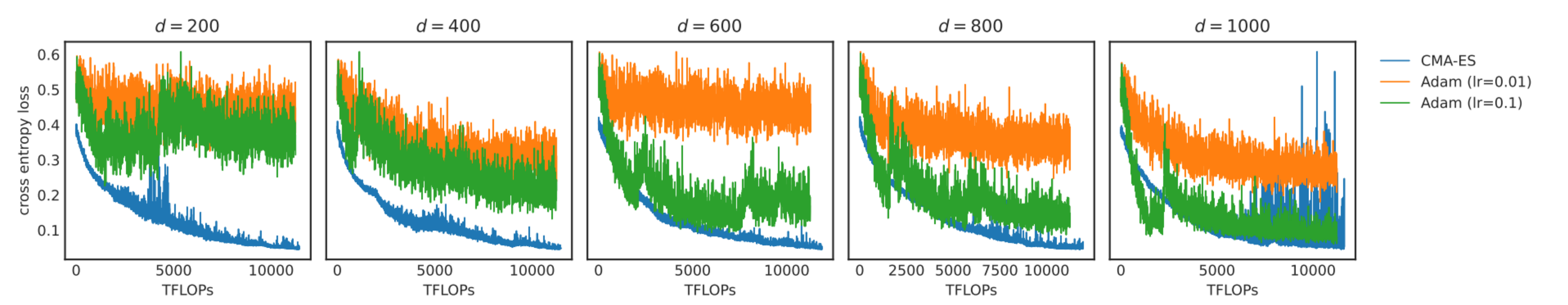
Method	SST-2 acc	Yelp P. acc	AG's News acc	DBPedia acc	MRPC F1	SNLI acc	RTE acc	Avg.
Gradient-Based Methods								
Prompt Tuning	68.23 ±3.78	61.02 ±6.65	84.81 ±0.66	87.75 ±1.48	51.61 ±8.67	36.13 ±1.51	54.69 ±3.79	63.46
+ Pre-trained prompt	/	/	/	/	77.48 ±4.85	64.55 ±2.43	77.13 ±0.83	74.42
P-Tuning v2	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**

Black-box tuning is more favorable than gradient descent in the scenario of parameter-efficient few-shot learning.

Ablations



CMA-ES vs. Adam



CMA-ES outperforms Adam when subspace dim is low

Paper Code Slides