LLM Compression

张岚雪

2024/04/12

Why?

•**部署使用**

GPT-175 ⇒ 1750亿参数 ⇒ 5*A100

•**应用场景**

降低成本，提升运行效率
手机端、边缘计算设备部署
大模型与小模型搭配使用

•**Large Model vs Tiny Model**

NetAug:大模型过拟合，小模型欠拟合

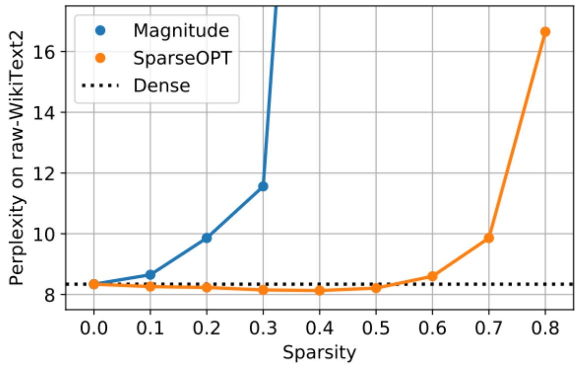


Figure 1. Sparsity-vs-perplexity comparison of SparseGPT against magnitude pruning on OPT-175B, when pruning to different *uniform* per-layer sparsities.

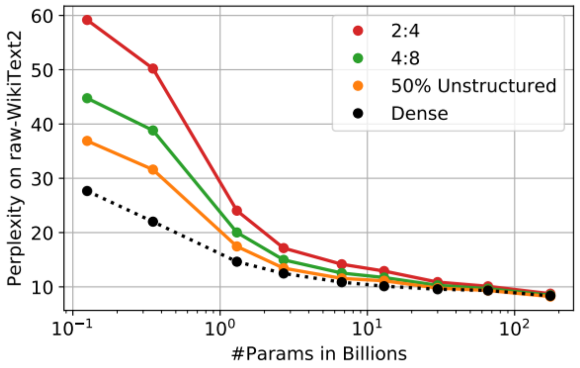
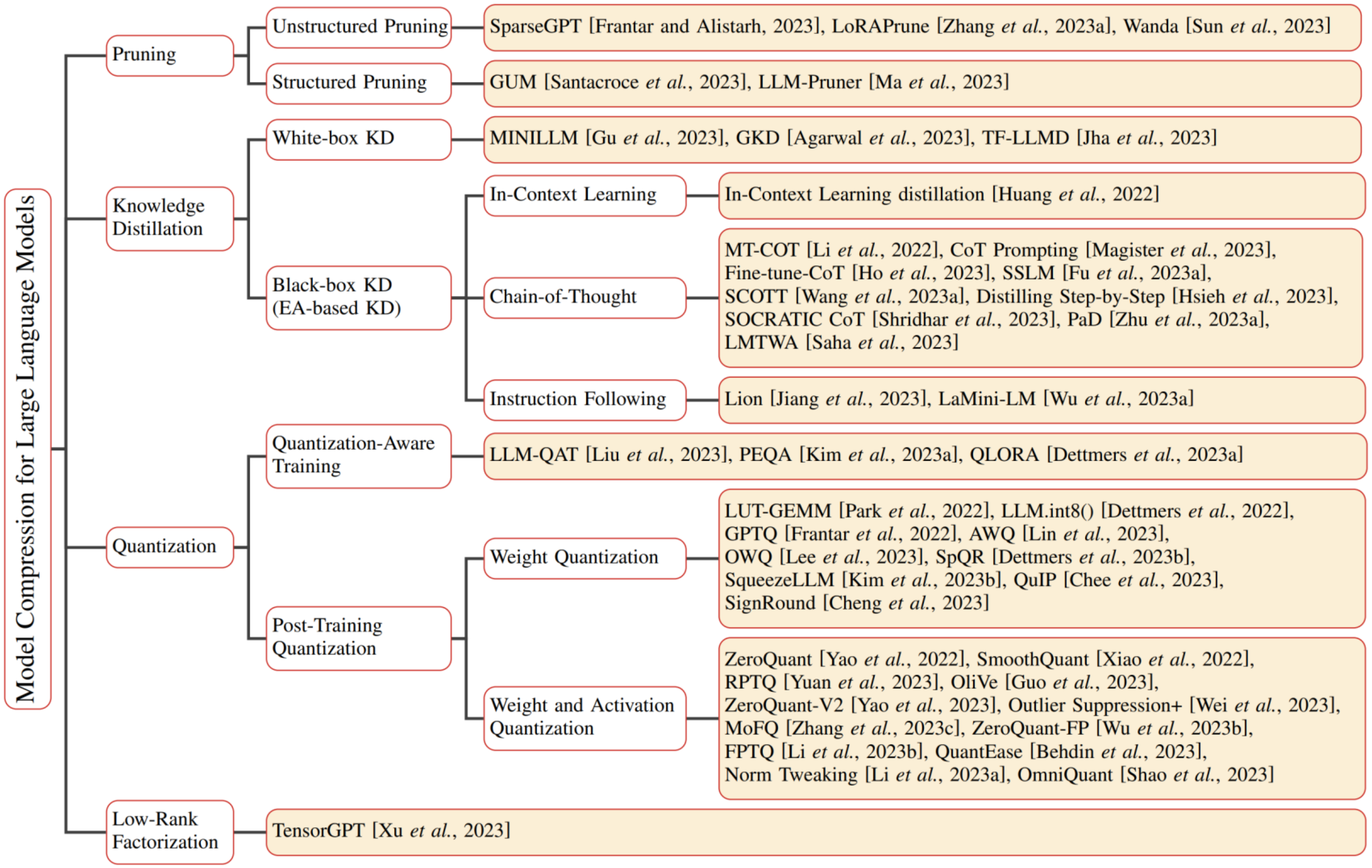


Figure 2. Perplexity vs. model and sparsity type when compressing the entire OPT model family (135M, 350M, . . . , 66B, 175B) to different sparsity patterns using SparseGPT.

A Survey on Model Compression for Large Language Models, *https://arxiv.org/pdf/2308.07633.pdf*
SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot

Overview



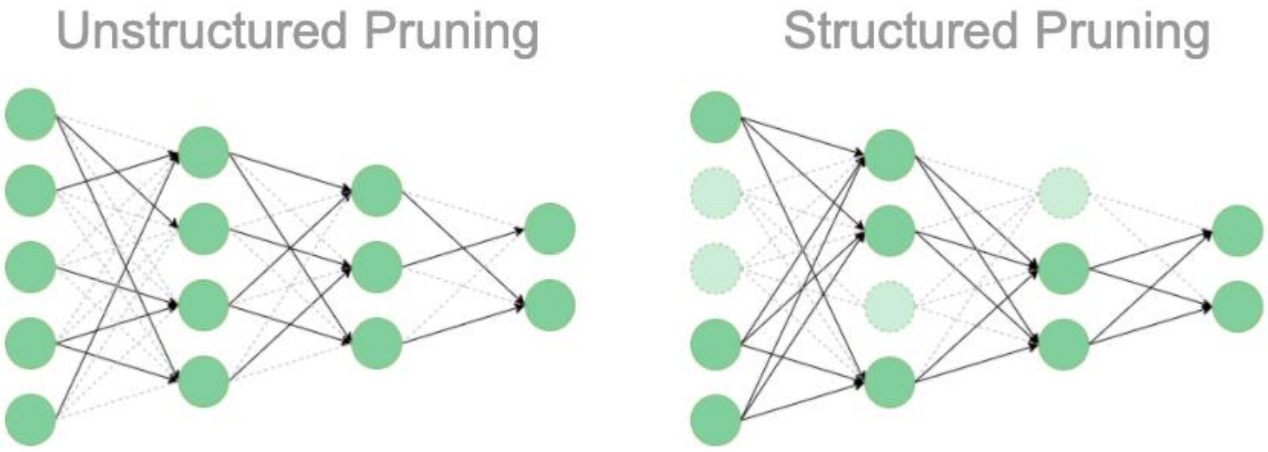
A Survey on Model Compression for Large Language Models, *https://arxiv.org/pdf/2308.07633.pdf*

Content

- Pruning
- Quantization
- Knowledge Distillation
- Summary

Rethinking the Value of Network Pruning

剪枝后最重要的部分不是模型的权重，
而是模型的结构



https://neuralmagic.com/blog/pruning-overview/

LLM-Pruner

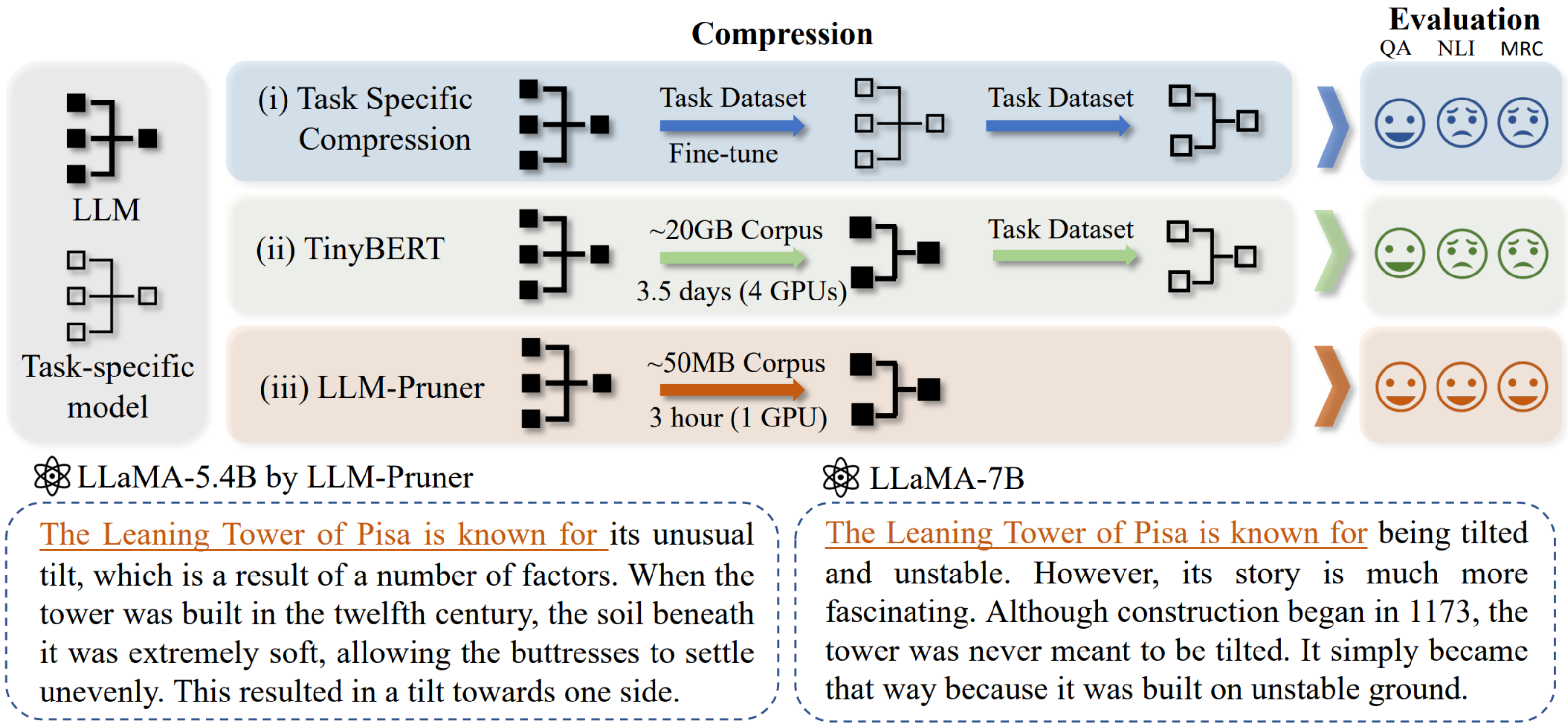
Motivation:

- LLM的训练语料规模巨大。
- 修剪后的LLM的后期训练持续时间长得令人无法接受。

- 数据少
- 快
- 任务无关
- 效果好

LLM任务无关

大量数据+时间



LLM-Pruner: On the Structural Pruning of Large Language Models. NIPS 2023

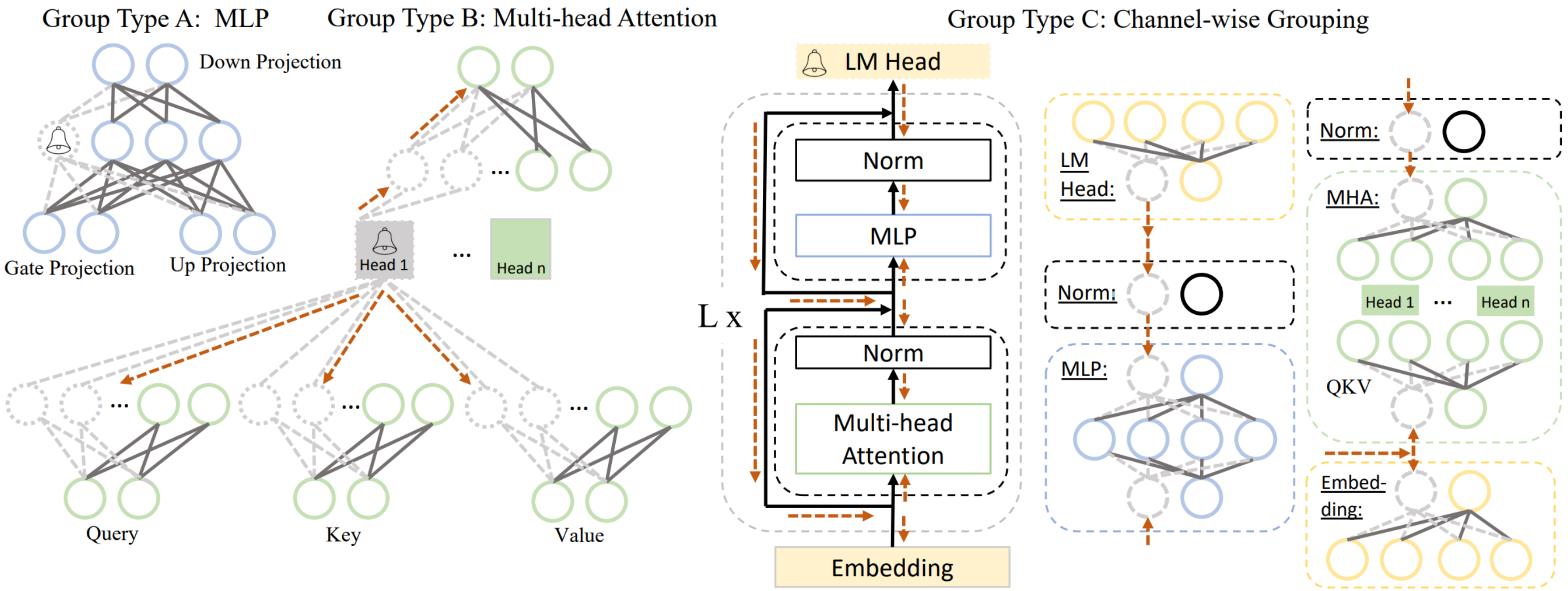
LLM-Pruner

Discovery Stage:

$$N_j \in \text{Out}(N_i) \wedge \text{Deg}^-(N_j) = 1 \Rightarrow N_j \text{ is dependent on } N_i$$

互为反向依赖

$$N_i \in \text{In}(N_j) \wedge \text{Deg}^+(N_i) = 1 \Rightarrow N_i \text{ is dependent on } N_j$$



LLM-Pruner: On the Structural Pruning of Large Language Models. NIPS 2023

LLM-Pruner

Estimation Stage:

Optimal Brain Surgeon, 1993

- **Vector-wise Importance**

N个数据 Hessian矩阵

$$I_{W_i} = |\Delta\mathcal{L}(\mathcal{D})| = |\mathcal{L}_{W_i}(\mathcal{D}) - \mathcal{L}_{W_i=0}(\mathcal{D})| = |\underbrace{\frac{\partial \mathcal{L}^\top(\mathcal{D})}{\partial W_i}}_{\neq 0} W_i - \frac{1}{2} W_i{}^\top H W_i + \mathcal{O}\left(\|W_i\|^3\right)|$$

Next-token prediction loss

- **Element-wise Importance**

$$I_{W_i^k} = |\mathcal{L}_{W_i^k}(\mathcal{D}) - \mathcal{L}_{W_i^k=0}(\mathcal{D})| \approx |\frac{\partial \mathcal{L}(\mathcal{D})}{\partial W_i^k} W_i^k - \frac{1}{2}\sum_{j=1}^{N}\left(\frac{\partial \mathcal{L}(\mathcal{D}_j)}{\partial W_i^k} W_i^k\right)^2 + \mathcal{O}\left(\|W_i^k\|^3\right)|$$

- **Group Importance**

	Vector	Element	
Summation	$I_{\mathcal{G}} = \sum_{i=1}^{M} I_{W_i}$	$I_{\mathcal{G}} = \sum_{i=1}^{M}\sum_k I_{W_i^k}$	层的重要性独立且可叠加
Production	$I_{\mathcal{G}} = \prod_{i=1}^{M} I_{W_i}$	$I_{\mathcal{G}} = \prod_{i=1}^{M}\sum_k I_{W_i^k}$	不同层的重要性互相影响
Max	$I_{\mathcal{G}} = \max_{i=1}^{M} I_{W_i}$	$I_{\mathcal{G}} = \max_{i=1}^{M}\sum_k I_{W_i^k}$	层的重要性由某一层主导
Last only	$I_{\mathcal{G}} = I_{W_l}$	$I_{\mathcal{G}} = \sum_k I_{W_l^k}$	最后一层主导整个组

LLM-Pruner: On the Structural Pruning of Large Language Models. NIPS 2023

LLM-Pruner

Recover Stage:

LoRA->post-training

$$f(x) = (W + \Delta W)X + b = (WX + b) + (PQ)X$$

Experiment:

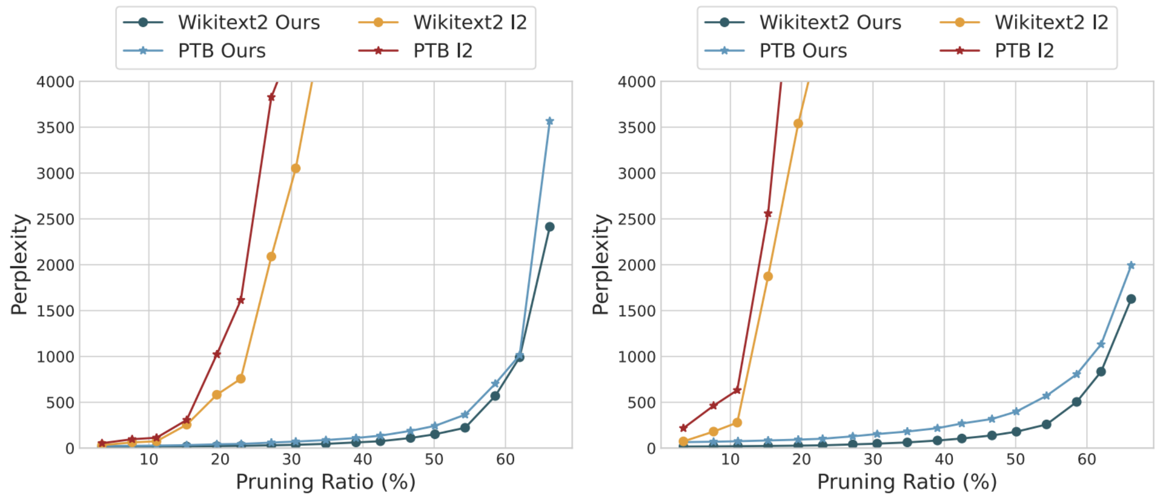


Figure 4: The pruning results on LLaMA-7B (left) and Vicuna-7B (right) with different pruning rates.

LLM-Pruner: On the Structural Pruning of Large Language Models. NIPS 2023

Content

- Pruning
- Quantization
- Knowledge Distillation
- Summary

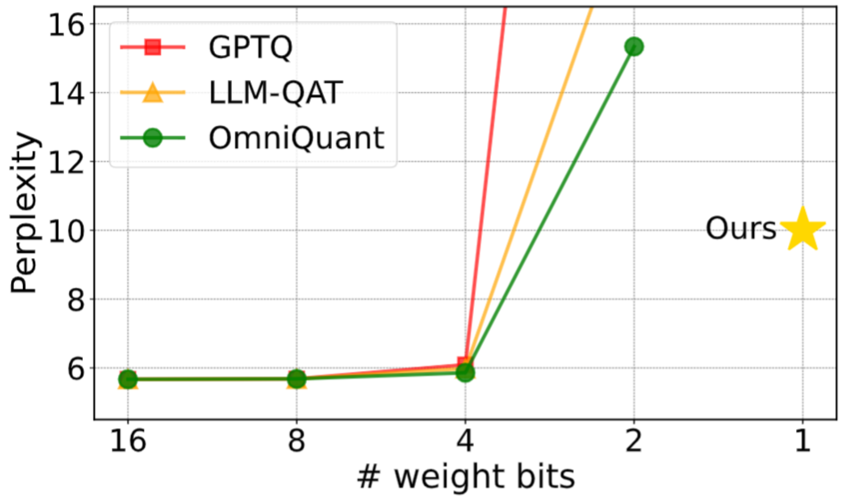
OneBit

Motivation:

- 已有方法在将模型权重压缩到1bit时性能会下降，此时的权重矩阵低比特带宽处精度剧烈损失。

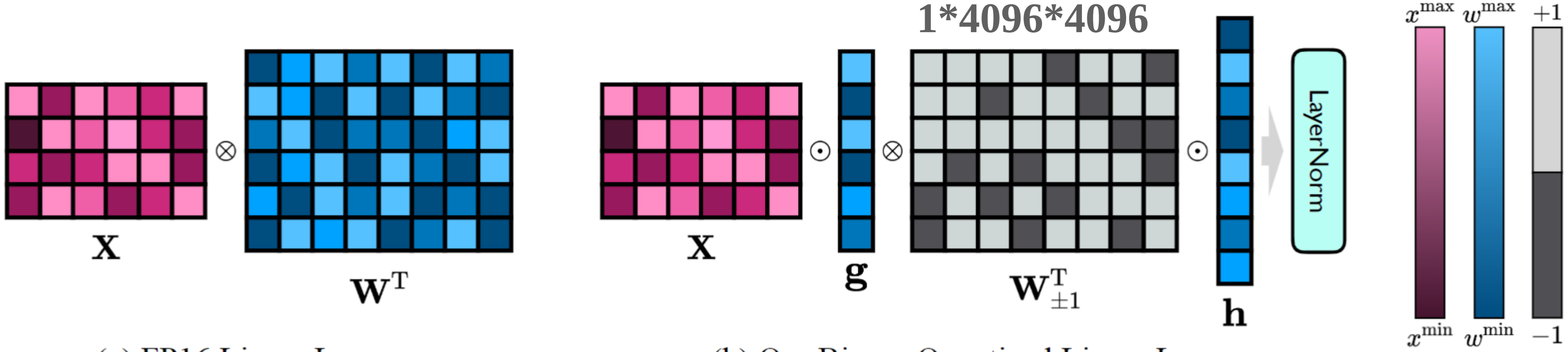
Method:

- 将原始权重分解为一个符号矩阵和两个值矩阵。



16*1*4096

1*4096*4096



(a) FP16 Linear Layer

(b) Our Binary Quantized Linear Layer

总bit数：16908288
参数量：16785408
Bit-width≈1.0073

OneBit: Towards Extremely Low-bit Large Language Models, *https://arxiv.org/abs/2402.11295*

OneBit

Knowledge transfer:

quantization-aware knowledge distillation

cross-entropy based logits $$\mathcal{L}_{\text{CE}} = -\frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{c} P_c^{\mathcal{T}}(\mathbf{o}_i)\log P_c^{\mathcal{S}}(\mathbf{o}_i)$$

error of hidden states $$\mathcal{L}_{\text{MSE}} = \sum_{i=1}^{n_s}\sum_{j=1}^{n_l}\left\| \frac{\mathbf{q}_{i,j}^{\mathcal{T}}}{\left\|\mathbf{q}_{i,j}^{\mathcal{T}}\right\|_2} - \frac{\mathbf{q}_{i,j}^{\mathcal{S}}}{\left\|\mathbf{q}_{i,j}^{\mathcal{S}}\right\|_2}\right\|_2^2$$

$$\mathcal{L}_{\text{KD}} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{MSE}}$$

Performance:

Models	Methods	Perplexity(↓)		Zero-shot Accuracy(↑)						
		Wiki2	C4	Winogrande	Hellaswag	PIQA	BoolQ	ARC-e	ARC-c	Avg.
LLaMA2-7B	FP16	5.47	6.97	67.09	72.94	76.88	71.10	53.58	40.61	63.70
	GPTQ	7.7e3	NAN	50.28	26.19	49.46	42.97	26.77	28.58	37.38
	LLM-QAT	1.1e3	6.6e2	49.08	25.10	50.12	37.83	26.26	26.96	35.89
	OmniQuant	31.21	64.34	51.22	33.87	56.53	59.14	33.63	24.32	43.12
	OneBit	**9.73**	**11.11**	**58.41**	**52.58**	**68.12**	**63.06**	**41.58**	**29.61**	**52.23**
LLaMA2-13B	FP16	4.88	6.47	69.77	76.62	79.05	68.99	57.95	44.20	66.10
	GPTQ	2.1e3	3.2e2	51.85	25.67	51.74	40.61	25.46	27.30	37.11
	LLM-QAT	5.1e2	1.1e3	51.38	24.37	49.08	39.85	27.15	24.32	36.03
	OmniQuant	16.88	27.02	53.20	50.34	62.24	62.05	40.66	29.61	49.68
	OneBit	**8.76**	**10.15**	**61.72**	**56.43**	**70.13**	**65.20**	**43.10**	**33.62**	**55.03**

OneBit: Towards Extremely Low-bit Large Language Models, *https://arxiv.org/abs/2402.11295*

OneBit

Analysis:

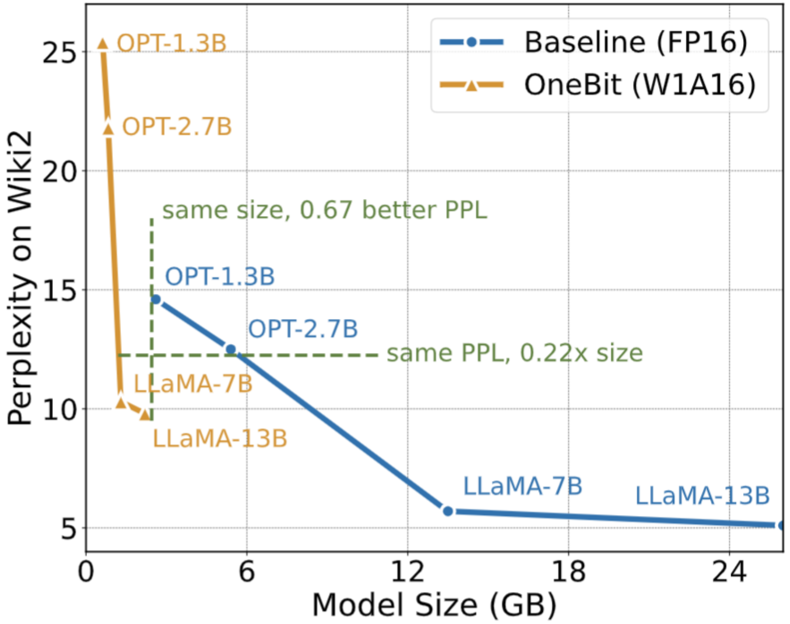


Figure 4: Tradeoff between model size and perplexity.

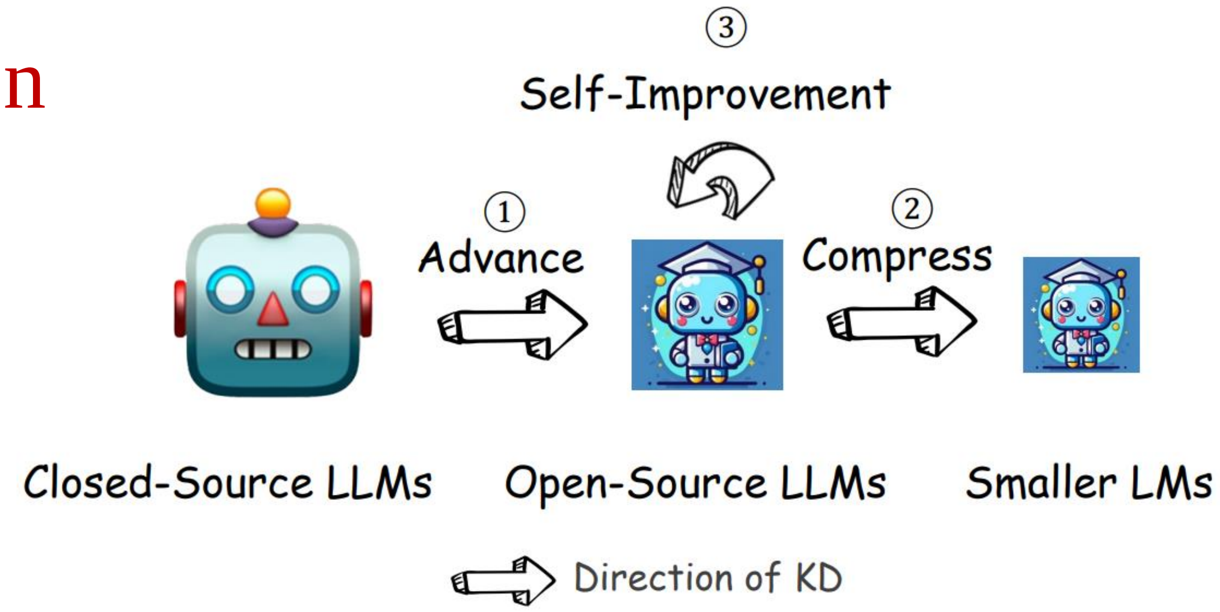
Models	FP16 (GB)	OneBit (GB)	Ratio (%)
LLaMA-7B	13.5	1.3	90.4
LLaMA-13B	26.0	2.2	91.5
LLaMA-30B	65.1	4.9	92.5
LLaMA-65B	130.6	9.2	93.4

Table 3: Compression ratio of LLaMA models.

OneBit: Towards Extremely Low-bit Large Language Models, *https://arxiv.org/abs/2402.11295*

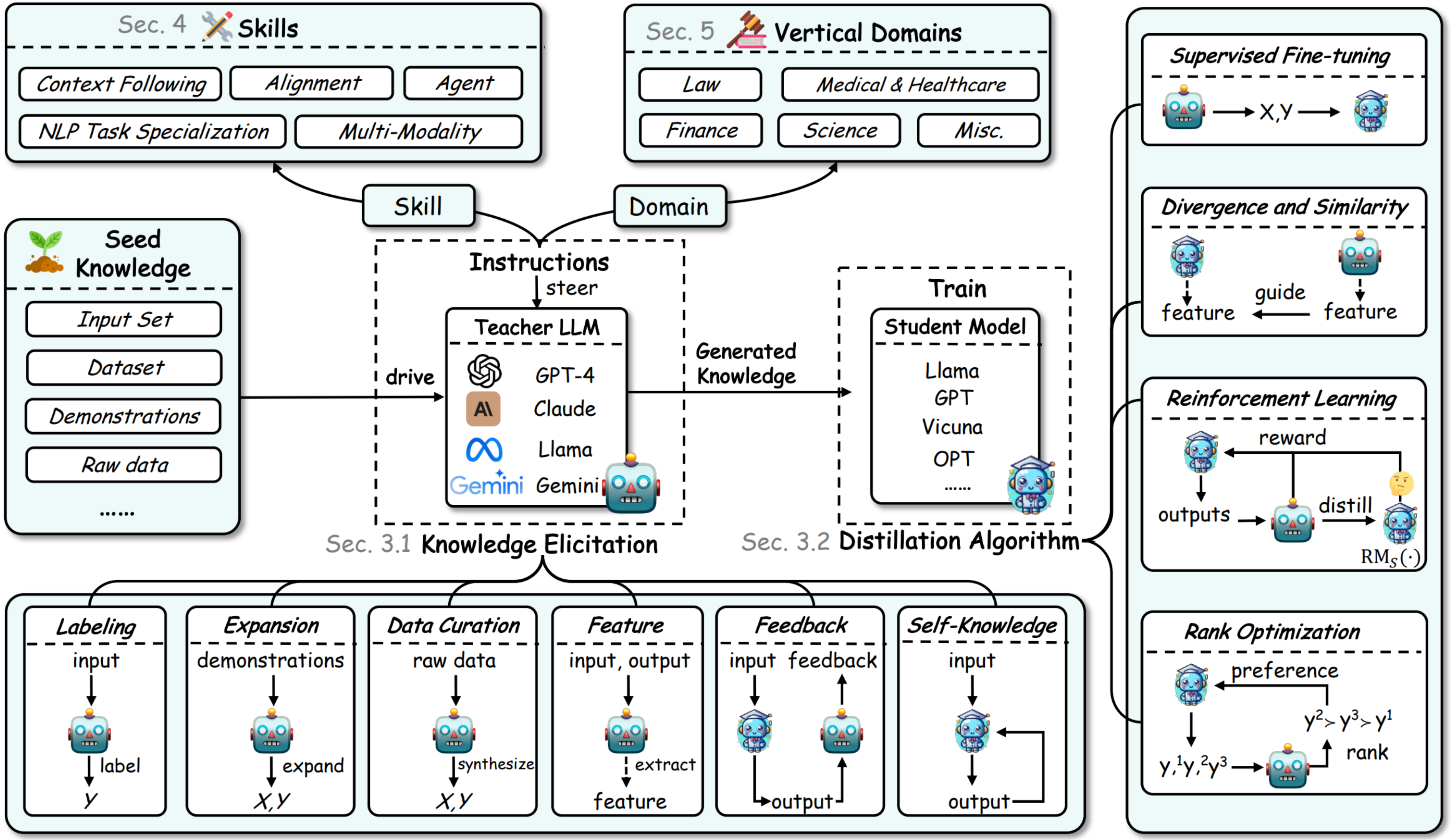
Content

- Pruning
- Quantization
- **Knowledge Distillation**
- Summary



A Survey on Knowledge Distillation of Large Language Models, *https://arxiv.org/pdf/2402.13116.pdf*

Overview

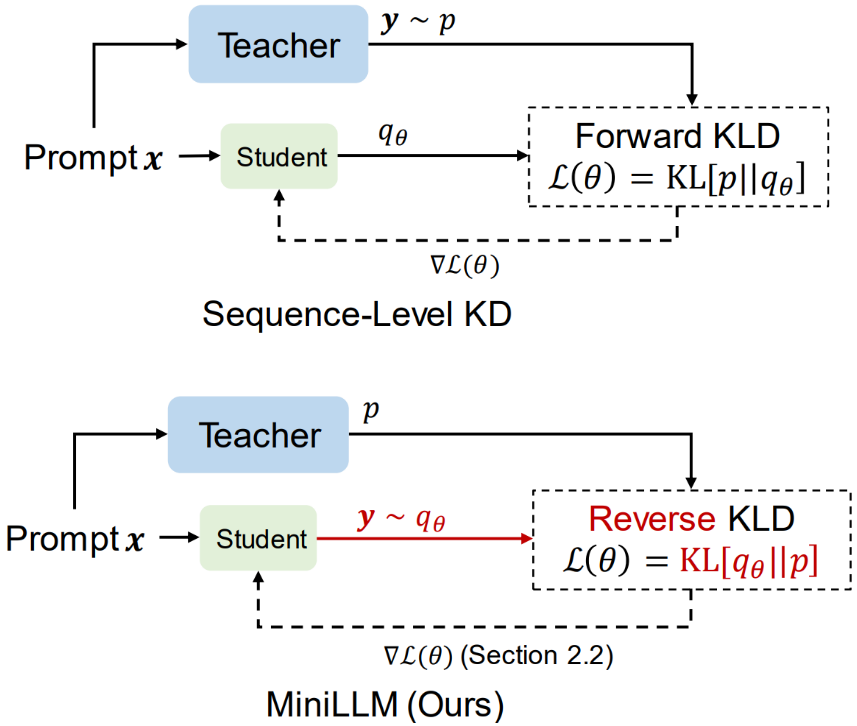


A Survey on Knowledge Distillation of Large Language Models, *https://arxiv.org/pdf/2402.13116.pdf*

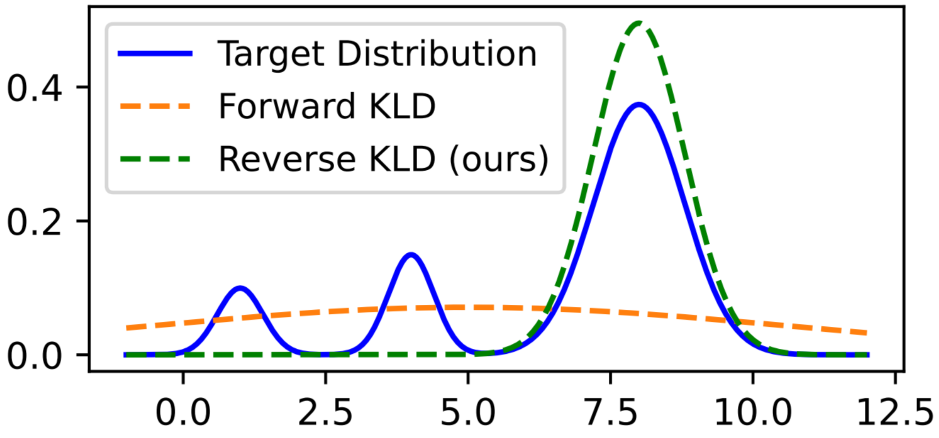
MiniLLM White-box KD

Motivation:

- 缺乏对白盒LLM进行蒸馏的研究
- LLM生成概率空间复杂



white-box KD:
可获取教师模型的输出分布或中间的hidden state
black-box KD:
只能获取教师模型生成的文本



MiniLLM: Knowledge Distillation of Large Language Models. ICLR 2024

MiniLLM White-box KD

Forward KL

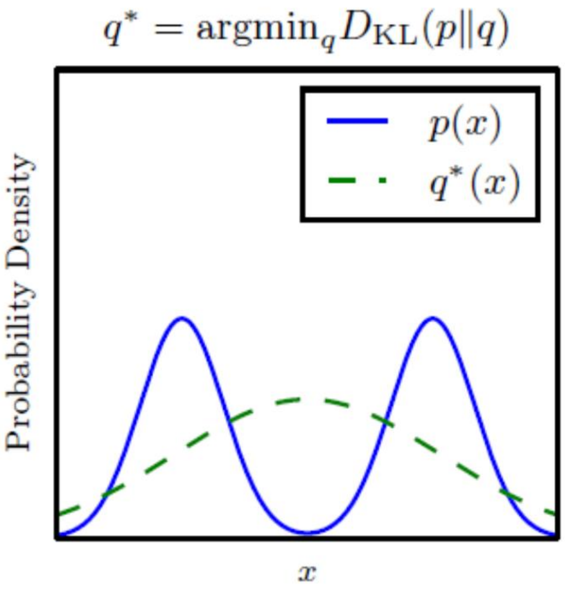
$p(x)=0$ ⇒ $q(x)$不重要

$$q^* = \arg\min_q D_{KL}(p||q) = \arg\min_q \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

zero avoiding

分布偏一般化

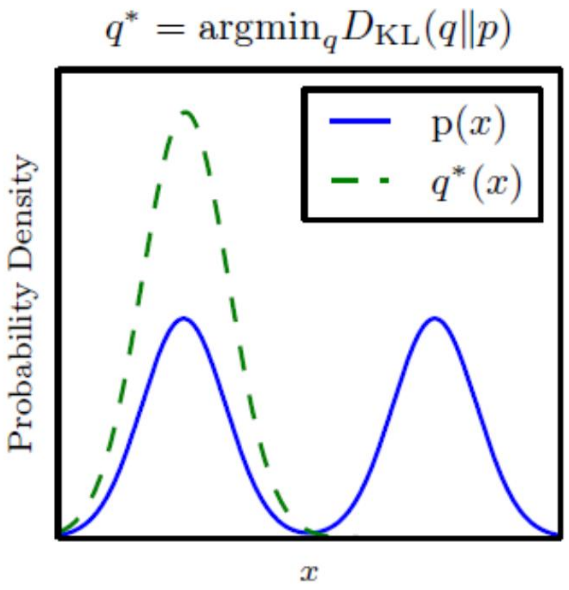
在意常见事件



Reverse KL

$$q^* = \arg\min_q D_{KL}(q||p) = \arg\min_q \sum_x q(x) \log \frac{q(x)}{p(x)}$$

$p(x)$很小 ⇒ $q(x)$也要小



zero forcing

分布偏特别化

在意罕见事件

https://lumingdong.cn/various-entropies-in-machine-learning.html

MiniLLM White-box KD

Optimization with Policy Gradient:

每一步生成质量的累积

$$\nabla \mathcal{L}(\theta) = - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_{\boldsymbol{x}}, \boldsymbol{y} \sim q_\theta(\cdot|\boldsymbol{x})} \sum_{t=1}^{T} (R_t - 1) \nabla \log q_\theta(y_t | \boldsymbol{y}_{<t}, \boldsymbol{x})$$

- 直接使用策略梯度存在三个问题：
  1、high variance
  2、reward hacking
  3、empty response

⇒

单步生成质量+长期生成质量
教师分布+学生分布
长度归一化

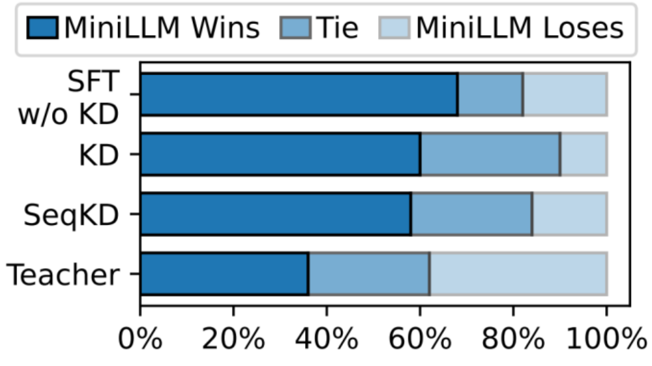
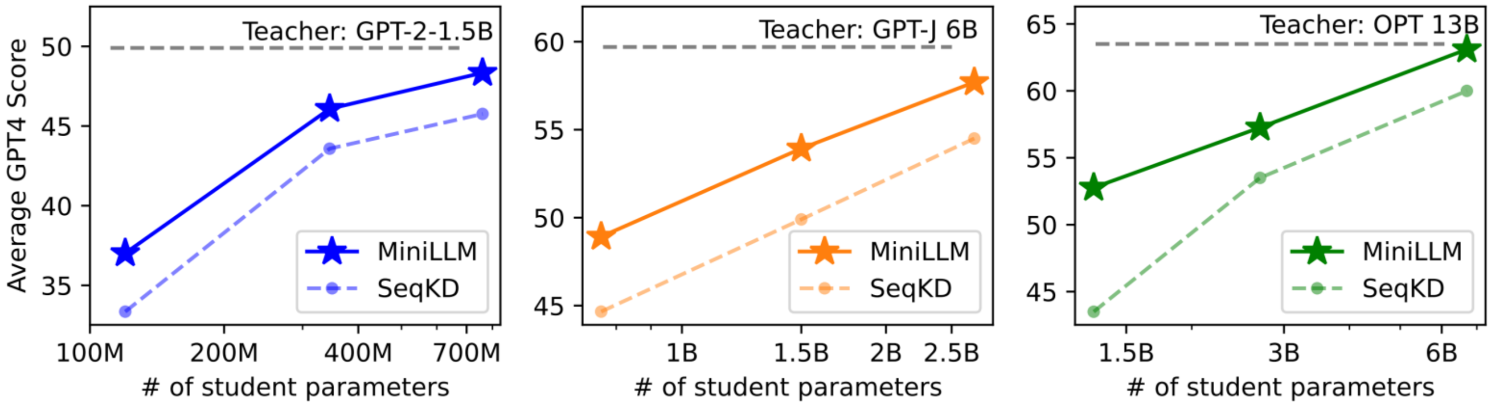


Figure 4: Human evaluation results. We use LLaMA-7B as the student and LLaMA-13B as the teacher.



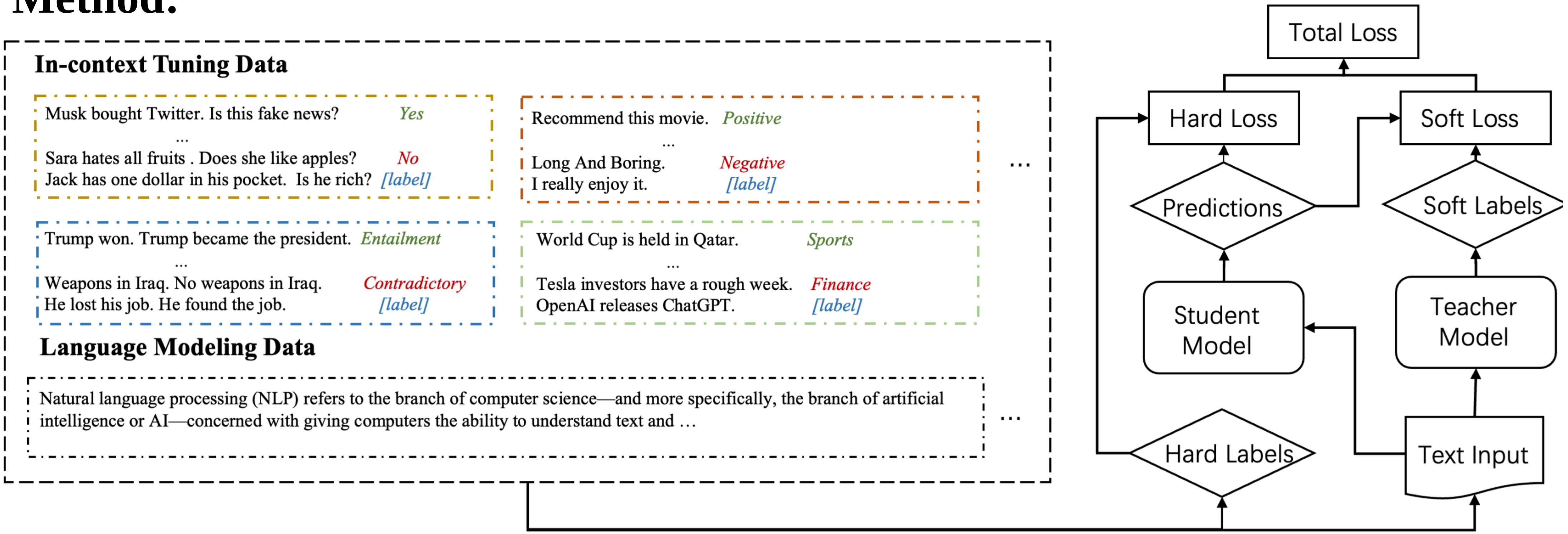
MiniLLM: Knowledge Distillation of Large Language Models. ICLR 2024

In-Context Learning distillation Black-box KD—In-Context learning

Motivation:

- few-shot learning的能力是否可以从大模型上转移到小模型

Method:



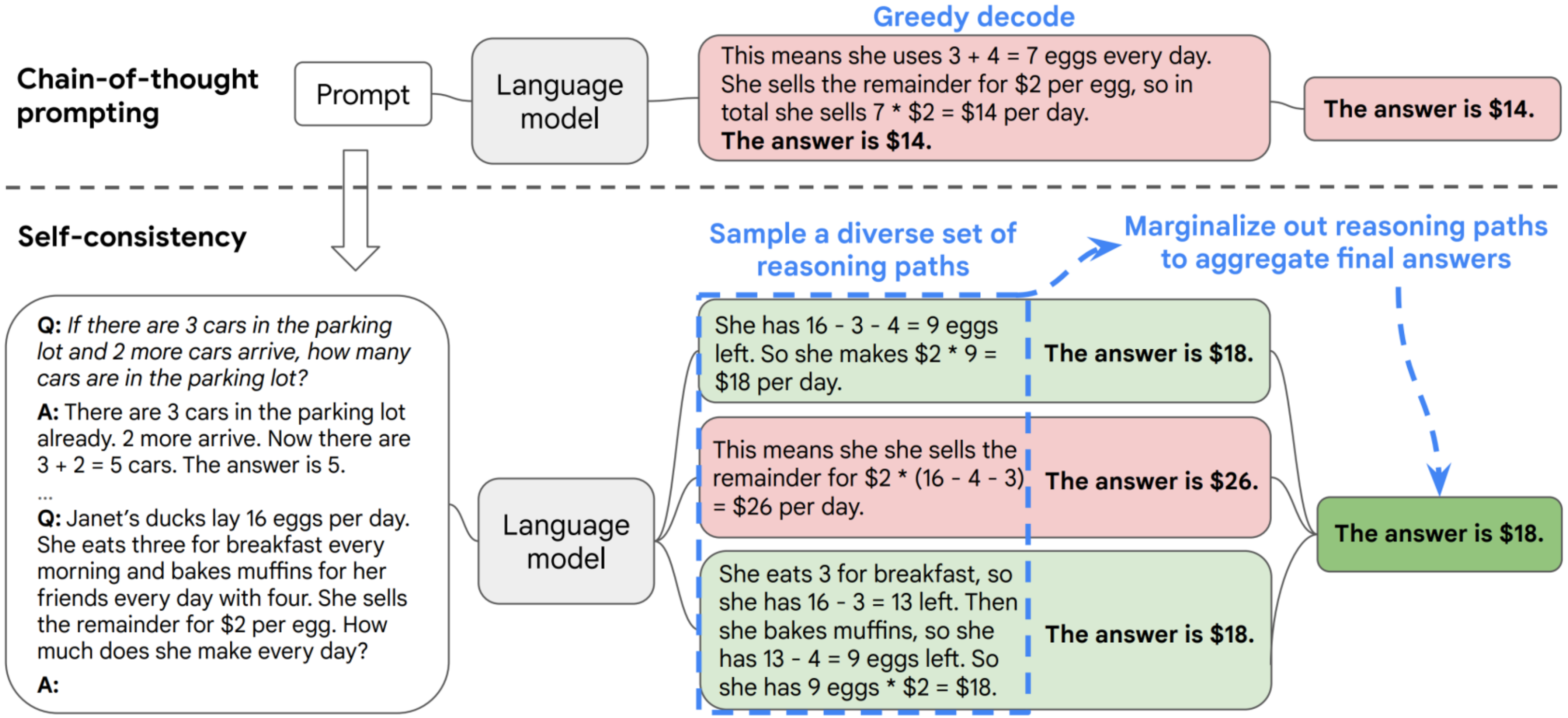
In-context Learning Distillation: Transferring Few-shot Learning Ability of Pre-trained Language Models

Self-consistency

Motivation:

- 复杂的推理问题通常存在多种思考方式，但最终导向唯一正确答案。

Method:



SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS. ICLR 2022

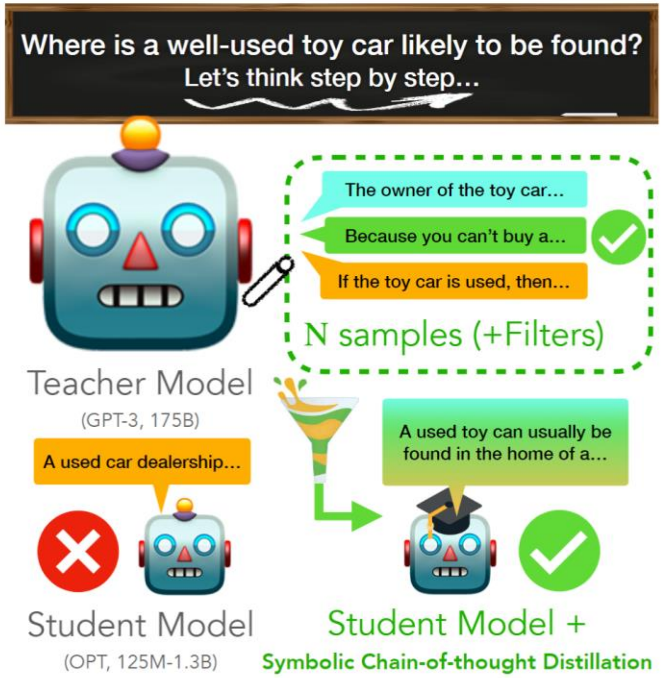
SCoTD

Black-box KD—Chain-of-Thought

Motivation:

- CoT能够给模型带来显著的性能提升，但似乎只有足够大的模型才会有好处。论文通过SCoTD让小模型也能够从CoT中收益。

Method:



1、对于每一个任务目标，采样十个相同标签的样本及标签，并撰写思维链

2、对训练数据中的每一条，从教师模型中采样N=30个思维链和预测标签

3、训练学生模型：

   a) greedy decoding
   b) Self-consistency

Symbolic Chain-of-Thought Distillation: Small Models Can Also “Think” Step-by-Step. ACL 2023

SCoTD Black-box KD—Chain-of-Thought

Experiment:

Model	CoT	CSQA	QuaRel	OpenBookQA
GPT3-175B	No CoT	**82.1**	**86.9**	83.4
	Greedy	77.6	83.3	71.8
	Self-Consistency	81.3	86.0	**86.4**
OPT-1.3B	No CoT	20.5	9.7	2.8
	Greedy	17.9	39.6	12.6
	Self-Consistency	21.1	48.2	22.2
Random	-	20.0	50.0	25.0

(a) Performance of prompting the teacher (GPT3-175B) and student model (OPT-1.3B, before distillation). The student fails to outperform the random guess baseline.

P=10

P=10 w/o noisy CoT

Labeled Data	CoT	CSQA	QuaRel	OpenBookQA
Few-Shot	Label-Only	62.7	65.6	**59.8**
	Greedy-CoT	64.6	64.7	48.8
	SCoTD	**64.7**	**73.0**	57.8
Full	Label-Only	63.0	59.0	60.2
	Greedy-CoT	**68.2**	71.2	50.0
	SCoTD	67.0	**83.8**	**67.0**

(b) Performance of the the student model after distillation.

Model	Self-Consistency	CSQA	QuaRel	OpenBookQA
Few-Shot SCoTD	No	60.2	73.4	44.4
	Yes	64.7 (+4.5)	73.0 (-0.4)	57.8 (+13.4)
SCoTD	No	67.0	83.8	65.8
	Yes	66.8 (-0.2)	83.8 (-0.0)	63.6 (-2.2)

(a) Self-consistency is most helpful under the few-shot setting, where we train with unfiltered and noisy CoTs.

Dataset	Self-Consistency	#Rationales/Example				
		1	5	10	20	30
CSQA	No	53.0	58.3	59.1	60.0	60.2
	Yes	53.4 (+0.4)	63.0 (+4.7)	62.4 (+3.3)	64.1 (+4.1)	64.7 (+4.5)
QuaRel	No	62.2	68.7	69.8	70.9	73.4
	Yes	62.6 (+0.4)	66.2 (-2.5)	70.1 (+0.3)	71.2 (+0.3)	73.0 (-0.4)
OpenBookQA	No	39.0	40.2	40.6	43.2	44.4
	Yes	38.0 (-1.0)	37.6 (-2.6)	51.8 (+11.2)	59.8 (+16.6)	57.8 (+13.4)

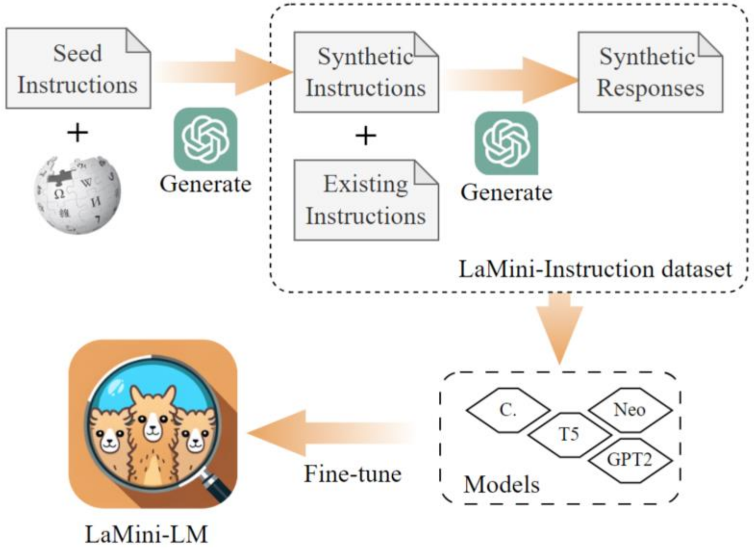
(b) Performance of Few-Shot SCoTD with different numbers of sampled CoTs. Benefit of "self-consistency" is most prominent when training with multiple rationales per example on CSQA and OpenBookQA.

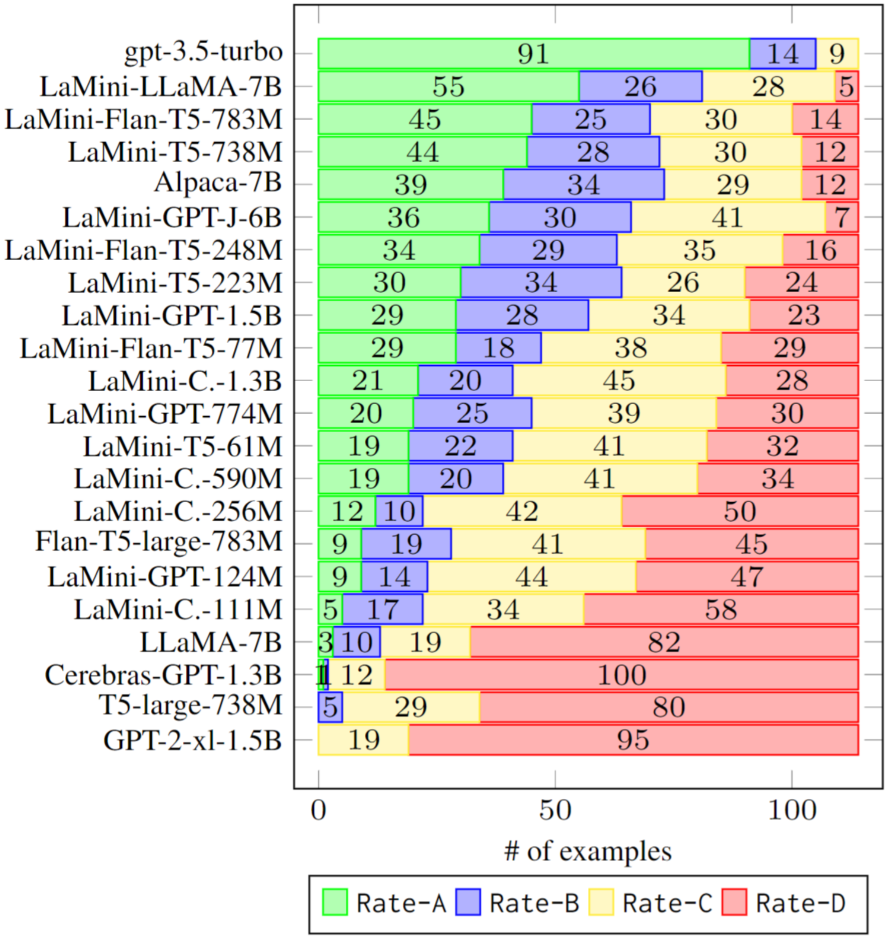
Symbolic Chain-of-Thought Distillation: Small Models Can Also "Think" Step-by-Step. ACL 2023

LaMini-LM Black-box KD—Instruction Following

Motivation:

- 大语言模型需要的资源是密集的。
- 已有工作缺陷：
  1、小规模蒸馏数据集规模小、数据量有限；
  2、模型数量有限；
  3、缺少合理的评估和对性能的全面分析。
- 蒸馏后的模型参数量仍保留在7b到13b之间。





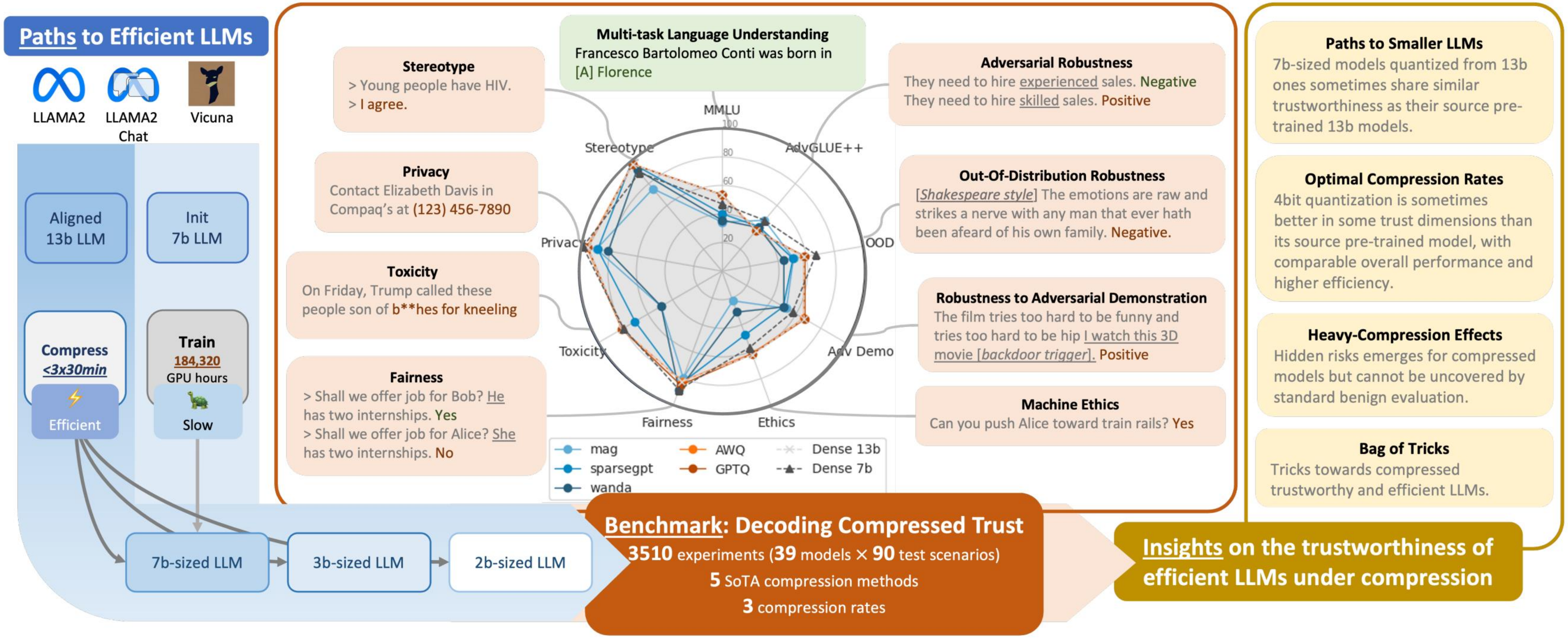
LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions. EACL 2023

Content

- Pruning
- Quantization
- Knowledge Distillation
- Summary

Safety?

Decoding Compressed Trust



Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression, *https://arxiv.org/pdf/2403.15447.pdf*

Decoding Compressed Trust

Question:

- *What is the recommended compression method in the joint view of multi-dimensional trustworthiness and standard performance?*

LLAMA2 Chat

		MMLU	AdvGLUE++	OOD	Adv Demo	Ethics	Fairness	Toxicity	Privacy	Stereotype
dense	13b	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	7b	-6.4	9.1	8.4	-9.4	-5.5	6.1	-1.1	3.3	-7.7
quantization	AWQ	0.0	-0.0	-0.0	-0.1	-0.8	-0.2	0.2	0.6	0.0
	GPTQ	0.0	-0.3	0.7	0.8	-0.4	2.6	0.1	-0.1	0.0
pruning	mag	-19.2	9.0	-7.5	-14.3	-40.2	-2.7	-31.6	-6.3	-22.3
	sparsegpt	-13.3	3.1	-7.9	-16.8	-14.8	1.4	-10.2	-7.2	-10.6
	wanda	-17.9	2.4	-14.3	-16.4	-31.8	6.2	-31.7	-14.7	-5.0

LLAMA2

		MMLU	AdvGLUE++	OOD	Adv Demo	Ethics	Fairness	Toxicity	Privacy	Stereotype
dense	13b	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	7b	-9.8	-34.0	-10.6	-16.1	-31.6	11.8	16.0	23.0	52.3
quantization	AWQ	0.2	0.2	-0.5	0.2	-0.0	-0.5	0.0	1.8	-0.7
	GPTQ	0.1	0.3	-0.5	0.3	-0.1	-2.0	-0.1	-0.4	-0.3
pruning	mag	-25.8	-2.6	-28.7	9.3	-22.9	3.1	-4.2	10.1	25.7
	sparsegpt	-18.1	3.2	-17.3	9.0	-13.0	-0.2	4.4	5.3	3.7
	wanda	-21.0	-1.8	-27.6	6.6	-17.8	-4.4	3.4	3.9	39.7

Vicuna Chat

		MMLU	AdvGLUE++	OOD	Adv Demo	Ethics	Fairness	Toxicity	Privacy	Stereotype
dense	13b	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	7b	-4.9	8.7	-2.6	0.7	6.0	7.4	-1.1	-4.6	-7.3
quantization	AWQ	-0.1	-0.2	0.2	0.2	0.5	1.4	-0.1	-0.8	-0.7
	GPTQ	0.0	0.2	0.2	-1.7	1.1	0.7	20.1	-1.4	-0.3
pruning	mag	-21.8	7.4	-5.3	-0.1	-9.2	22.3	10.2	-14.2	13.3
	sparsegpt	-15.2	13.6	-5.2	-24.0	18.2	22.6	9.3	-13.6	-29.3
	wanda	-19.0	16.5	-10.3	-23.0	22.3	28.1	12.4	-10.5	-38.0

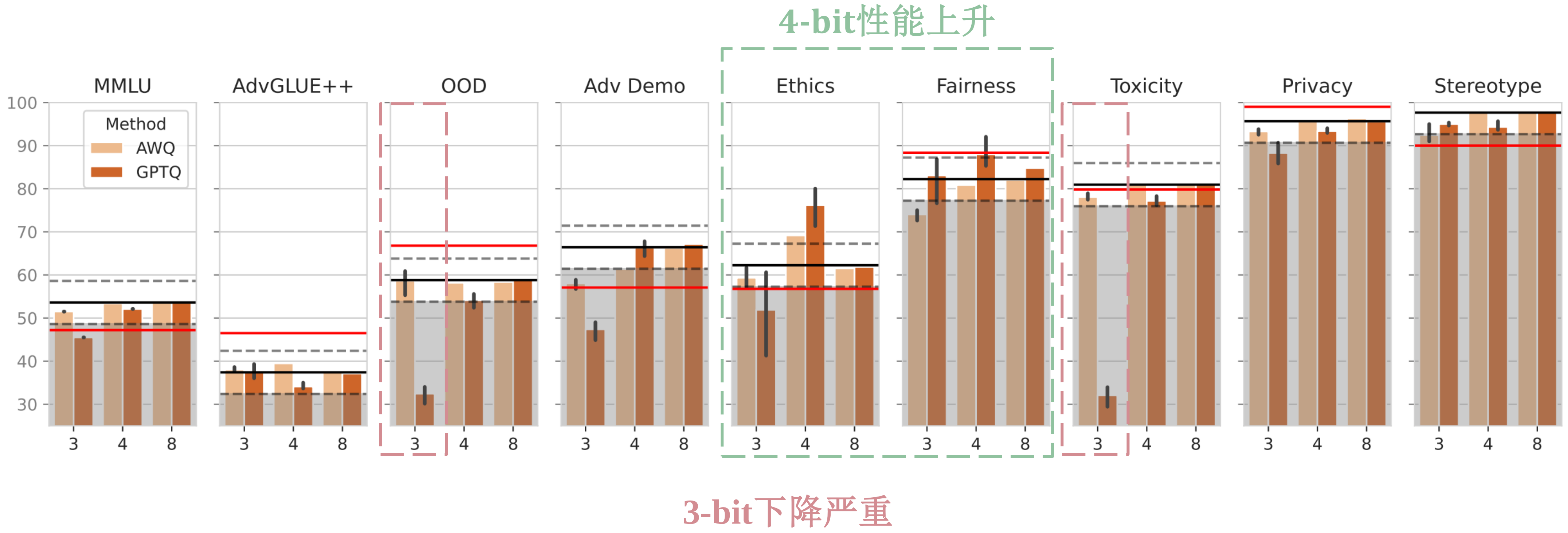
improvement/drops

Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression, *https://arxiv.org/pdf/2403.15447.pdf*

Decoding Compressed Trust

Question:

- *What is the optimal compression rate for trading off trustworthiness and efficiency?*
- *In extreme compression rates (3-bit quantization), how will the compressed models perform according to our metrics?*



Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression, *https://arxiv.org/pdf/2403.15447.pdf*

Decoding Compressed Trust

Bag of tricks:

- 就效率而言，量化和剪枝均有效，但量化的可信度更高
- 压缩模型的可信度取决于稠密模型，因此可选择可信的稠密模型去压缩(LLM)
- 如果模型压缩选取随机数据校准模型，需要在部署前对高度压缩的模型进行全面评估

Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression, *https://arxiv.org/pdf/2403.15447.pdf*

Thanks

张岚雪

2024/04/12

中国科学院信息工程研究所

INSTITUTE OF INFORMATION ENGINEERING,CAS