Training-Inference Mismatch In LLM KD(2)

Alephia 25/10/15

BACKGROUND

模型q_θ依据前缀w^{t-1}生成文本的时候，loss可以表示为

$$l(q_\theta, w^{t-1}; o) = \mathop{\mathbb{E}}_{w_t \sim o(\cdot|w^{t-1})} \log \frac{o(w_t|w^{t-1})}{q_\theta(w_t|w^{t-1})}$$

由目标分布o采样下一个token w_t，再进行KLD对齐。

可以展开得到

$$L(q_\theta; o) \approx \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w^{t-1} \sim d_o^{t-1},\, w_t \sim o(\cdot|w^{t-1})} \log \frac{o(w_t|w^{t-1})}{q_\theta(w_t|w^{t-1})}$$

TRAIN-INFERENCE MISMATCH

由于模型能力不足，生成token的分布与目标分布存在差距，进而模型训练和推理时面对的前缀是不同的

- Distribution Mismatch (Exposure Bias)
- Error Accumulation

Distribution Mismatch会导致Error Accumulation

TRAIN-INFERENCE MISMATCH

模型推理时，与oracle model的总偏差表示为

$$
\begin{aligned}
L(q_\theta; o) &= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w^{t-1}\sim d_{q_\theta}^{t-1},\, w_t\sim o(\cdot|w^{t-1})} \log \frac{o(w_t|w^{t-1})}{q_\theta(w_t|w^{t-1})} \\
&= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w^{t-1}\sim d_{q_\theta}^{t-1}} D_{KL}(o(\cdot|w^{t-1})||q_\theta(\cdot|w^{t-1}))
\end{aligned}
$$

记生成第t个token的期望误差为

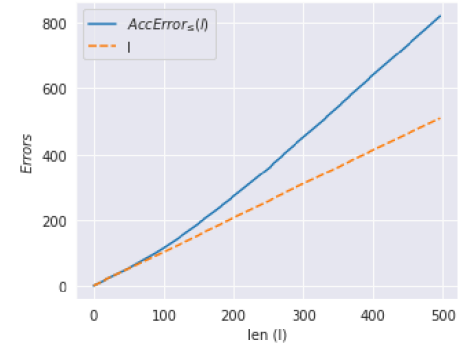
$$
\epsilon_t = \mathop{\mathbb{E}}_{w_0^{t-1}\sim d_o^{t},\, w_t\sim o(\cdot|w^{t-1})} \log \frac{o(w_t|w^{t-1})}{q_\theta(w_t|w^{t-1})}
$$

Arora, K., Asri, L.E., Bahuleyan, H., & Cheung, J.C. Why Exposure Bias Matters: An Imitation Learning Perspective of Error Accumulation in Language Generation. In ACL, 22

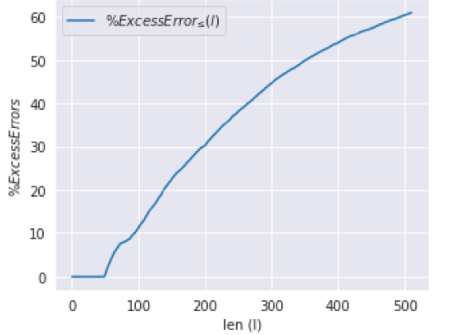
TRAIN-INFERENCE MISMATCH

$$l\epsilon_{\leq l} \leq L_{\leq l}(q_\theta) \leq l^2\epsilon_{\leq l}, \;\; \epsilon_{\leq l} = \frac{1}{l}\sum_{t=1}^{l}\epsilon_t$$

$$AccErr_{\leq}(l) = \frac{L_{\leq l}(q_\theta)}{\epsilon_{\leq l}}, \;\; ExAccErr_{\leq}(l) = \frac{L_{\leq l}(q_\theta) - l\epsilon_{\leq l}}{l\epsilon_{\leq l}} \cdot 100$$



(a) $AccErr_{\leq}(l)$ vs l

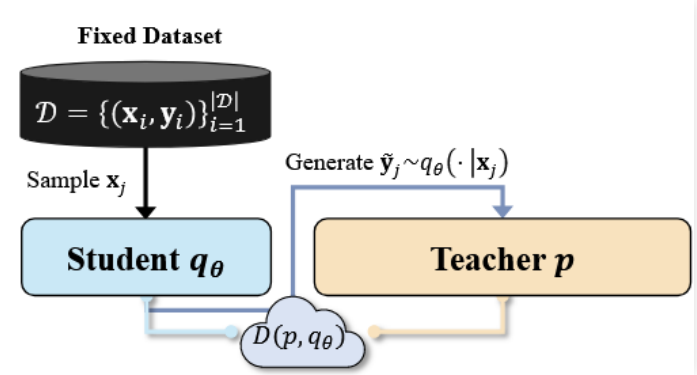


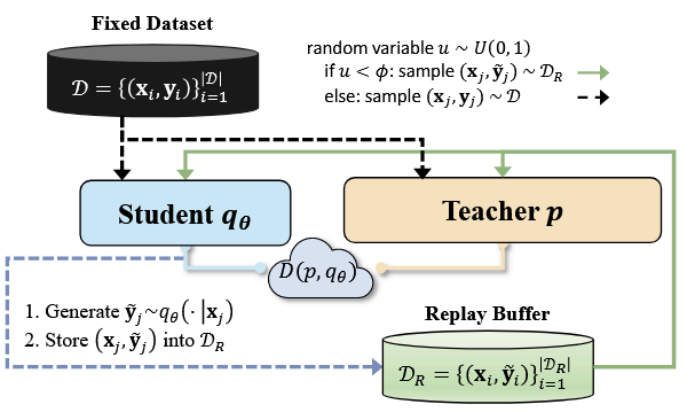
(b) $\%\text{ExError}_{\leq}(l)$ vs l.

THE UTILIZATION OF SGO

引入模型自己推理生成的内容用于训练(Student Generated Output)

$$L_{SGO}(q_\theta; o) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w^{t-1}\sim d_{q_\theta}^{t-1},\, w_t\sim o(\cdot|w^{t-1})} \log \frac{o(w_t|w^{t-1})}{q_\theta(w_t|w^{t-1})}$$





Agarwal, R., Vieillard, N., Zhou, Y., Stańczyk, P., Ramos, S., Geist, M., & Bachem, O. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In ICLR, 24

Ko, J., Kim, S., Chen, T., & Yun, S. DistiLLM: Towards Streamlined Distillation for Large Language Models. In ICML, 24

THE UTILIZATION OF SGO

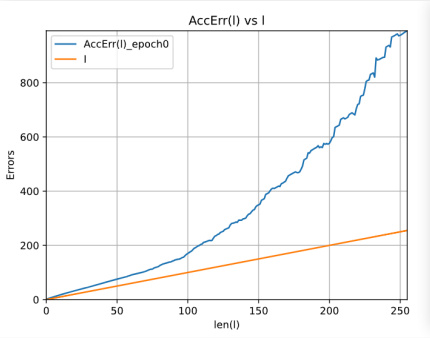
- prepare SGOs for each training sample before distillation
- **only use SGOs for distillation**
- update SGOs with latest student parameters during distillation(twice in huawei's work)

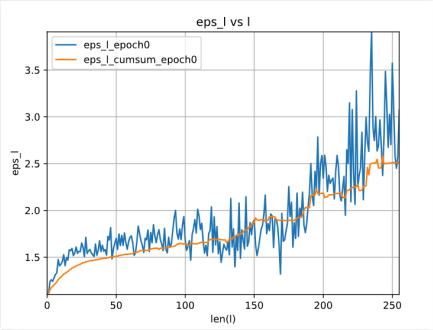
"we find that the performance of the ”Distillation by Student” strategy can be enhanced. By **periodically updating the student model with its latest parameters during the training process—in our case, twice—and re-generating responses for subsequent distillation**, the model’s accuracy is further improved to 63.43%. "

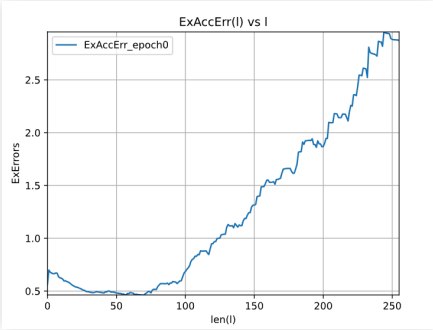
Rang, M., Bi, Z., Zhou, H., Chen, H., Xiao, A., Guo, T., Han, K., Chen, X., & Wang, Y. (2025). Revealing the Power of Post-Training for Small Language Models via Knowledge Distillation

CURVES BEFORE/AFTER DISTILLATION

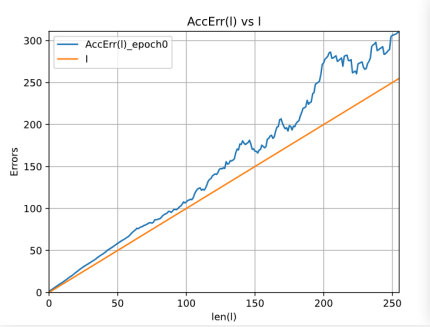
before distillation

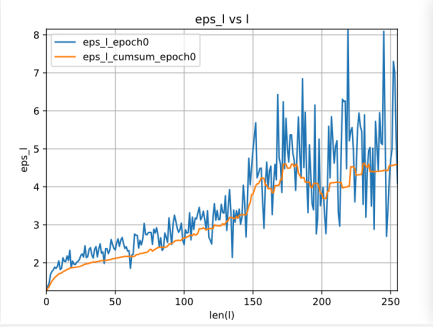


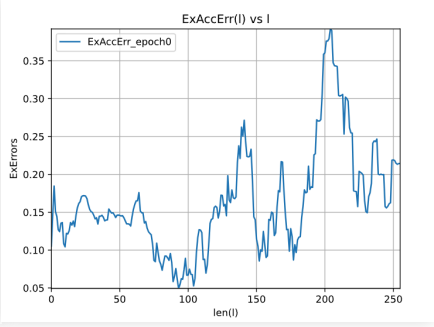




after distillation

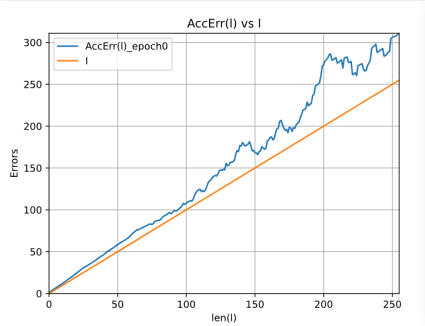


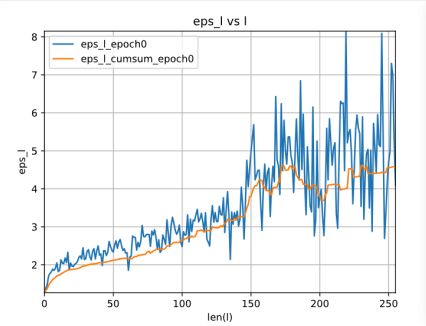


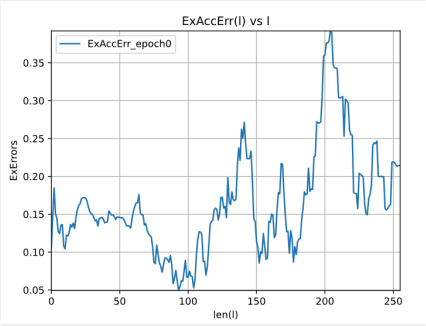


CURVES ON DIFFERENT GENERATION WAYS

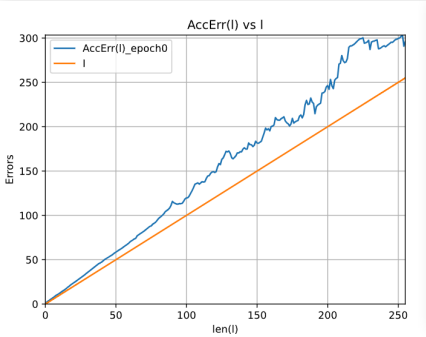
Sampling(T=1)

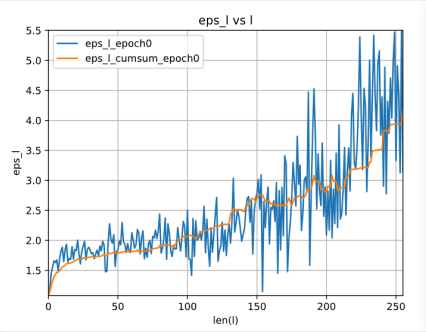


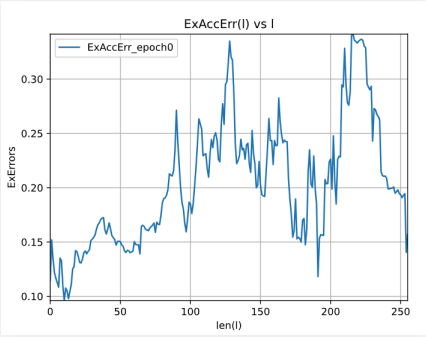




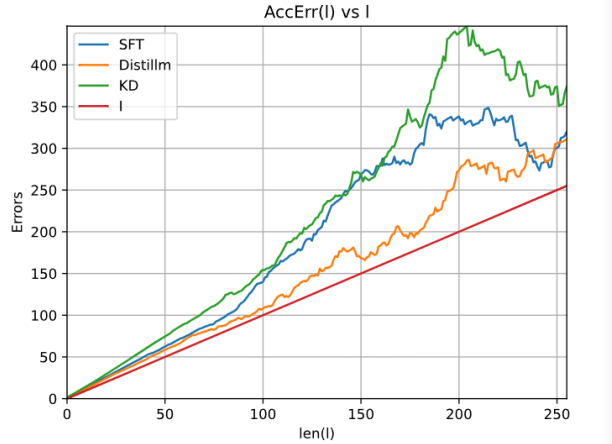
Greedy

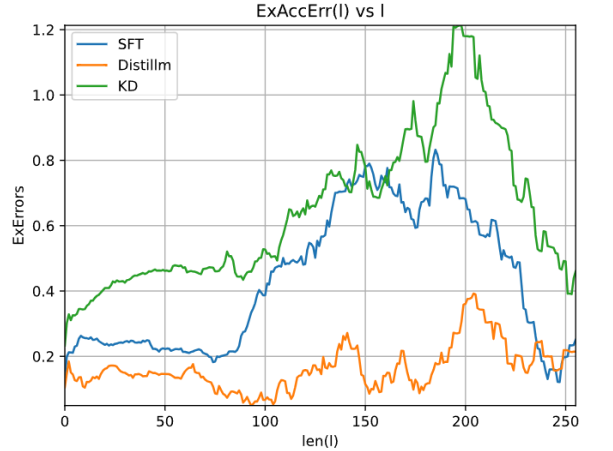


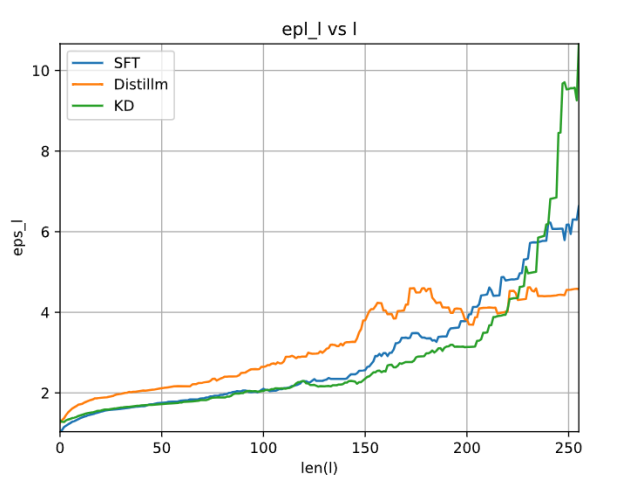




CURVES ON VARIOUS POST-TRAINING WAYS



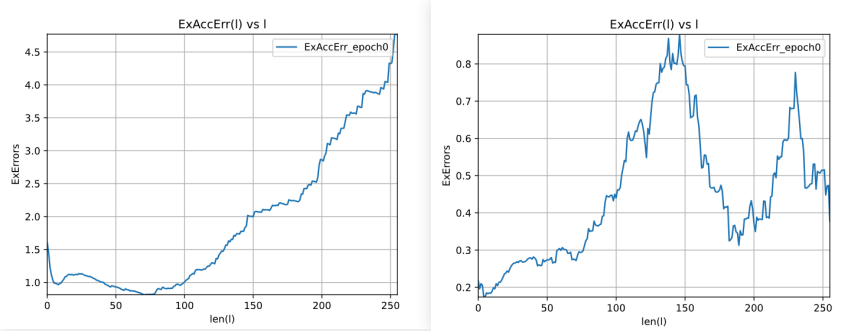




on opt-0.1b, dolly datast

QUESTIONS

- 🤔 The curves collapse in the late generation stage(occurs after post-training)



gpt-base before/after sft

A more powerful oracle model?
What will happen when model have more powerful long text generation ability?

- 🤔 A better model, a higher KLD?

SAMPLE WISE WEIGHT

$$L_{WSGO}(q_\theta; o) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w^{t-1} \sim d_{q_\theta}^{t-1}} \frac{1}{Q(o; w^{t-1})} D_{KL}(o(\cdot|w^{t-1}) || q_\theta(\cdot|w^{t-1}))$$

```
# sample_weights = torch.exp(avg_log_probs) # [batch_size]
sample_weights = (1 / (torch.exp(-avg_log_probs)+1e-6)).clamp(min=0.1) # [batch_size]
sample_weights = sample_weights / sample_weights.mean() # make the expection of weights to 1
```

Dataset	Model	Way	Loss(On validation set)	Exact_match	RougeL
Dolly	gpt2-base(124M)	distillm(new weights)	5.866	**2.002**	25.518
		distillm	**5.693**	1.9019	**26.292**
Dolly	opt-0.1b	distillm(new weights)	8.656	0.9	18.074
		distillm	**5.861**	**1.0**	**21.5209**

WEIGHT OF CE LOSS WITH SGO

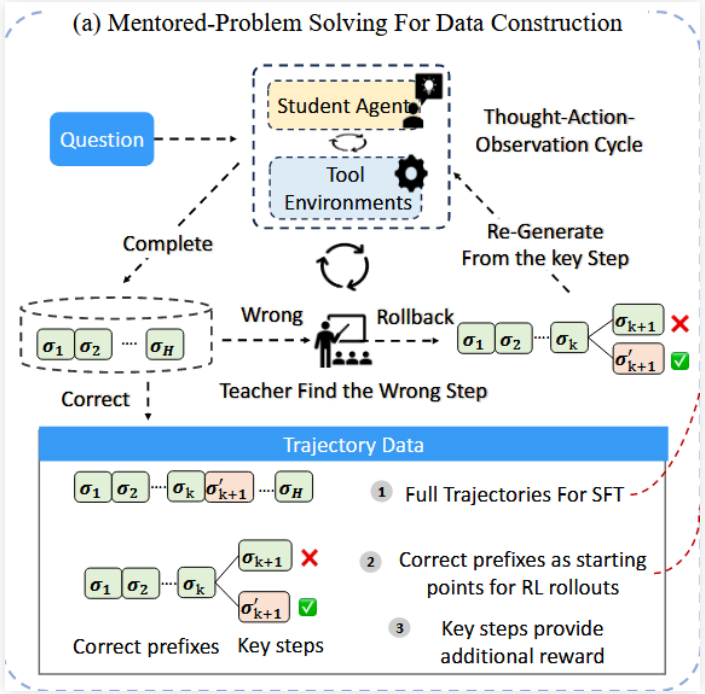
$$Loss = (1 - \lambda_{KD})L_{CE} + \lambda_{KD}L_{KD}$$

🤔 adaptive weight to λ_{KD}

Dataset	Model	weight_KD	Loss(On validation set)	Exact_match	RougeL
Dolly	gpt2-base(124M)	1	**5.693**	**1.9019**	**26.292**
		0.5	6.0615234375	0.0	9.5584
		0.3	6.26708984375	0.1001	2.7882
Dolly	opt-0.1b	1	5.861	1.0	**21.5209**
		0.9	**5.7586669921875**	**1.4**	19.6998
		0.5	5.9224853515625	1.1	9.5386
		0.3	5.9422607421875	1.0	8.1571

FURTURE

🤔 student explores, teacher corrects

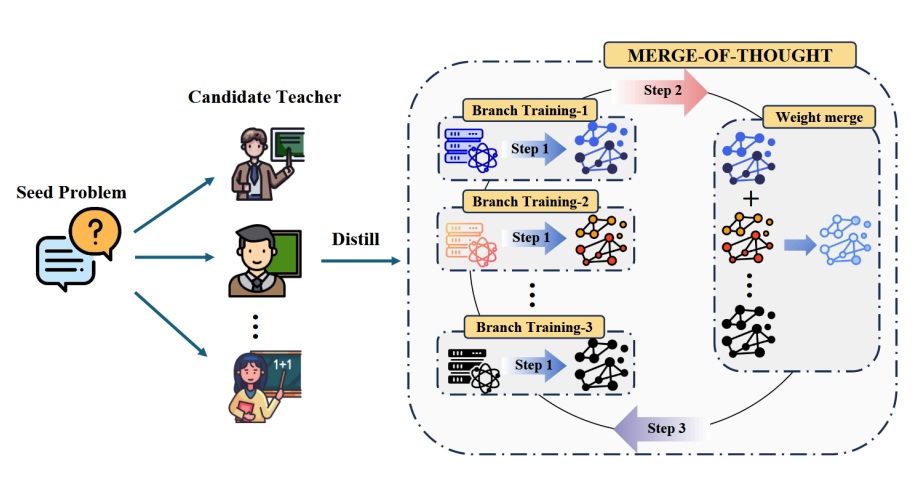


Lyu, Y., Wang, C., Huang, J., & Xu, T. (2025). From Correction to Mastery: Reinforced Distillation of Large Language Model Agents.

FURTURE

🤔 Merge-of-Teachers

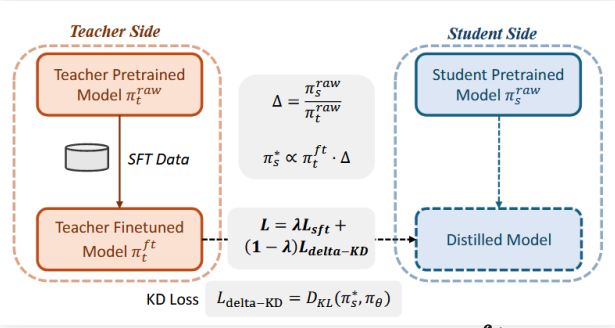
"different students have different 'best teachers', and even for the same student the best teacher can vary across datasets"



Shen, Z., Qin, Z., Huang, Z., Chen, H., Hu, J., Zhuang, Y., Lu, G., Chen, G., & Zhao, J. (2025). Merge-of-Thought Distillation.

FURTURE

🤔 align relative information



$$\Delta(p_1, p_2)(y|x) = \frac{p_1(y|x)}{p_2(y|x)}$$

$$\Delta(\pi_s^*, \pi_t^{ft})(y|x) \propto \Delta(\pi_s^{raw}, \pi_t^{raw})(y|x)^\alpha$$

$$\pi_s^*(y|x) = \frac{1}{Z(x,y)} \pi_t^{ft}(y|x) \left(\frac{\pi_s^{raw}(y|x)}{\pi_t^{raw}(y|x)} \right)^\alpha$$

FURTURE

🤔 **Explores on model structure**

MoE：训练推理时选择的expert可能不同，从而导致后续差异

目前大部分工作聚焦在算法层面

---

https://zhuanlan.zhihu.com/p/1959976628290590602

THANKS