Training Large Language Models for Retrieval-Augmented Question Answering through Backtracking Correction

Huawen Feng, Zekun Yao, Junhao Zheng, Qianli Ma

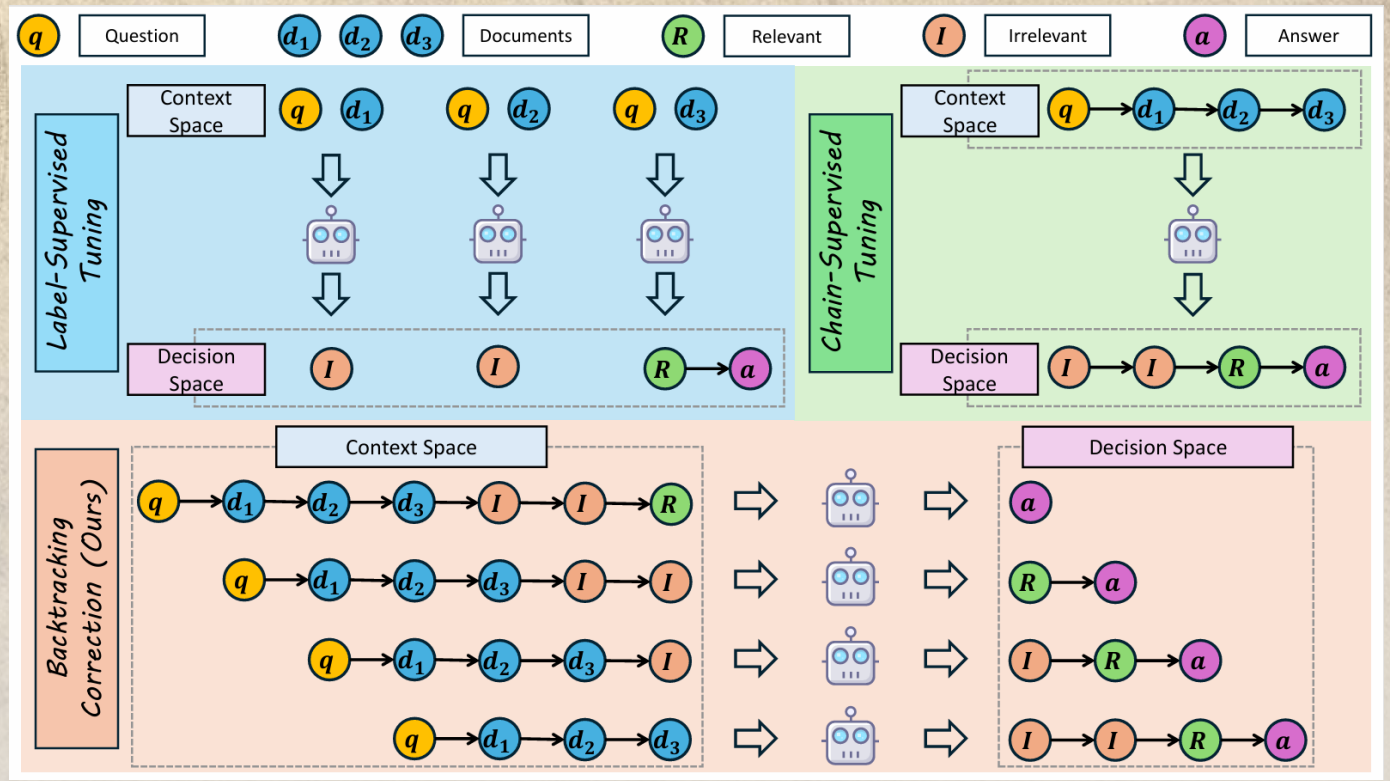
School of Computer Science and Engineering,

South China University of Technology,

Guangzhou, Guangdong, China

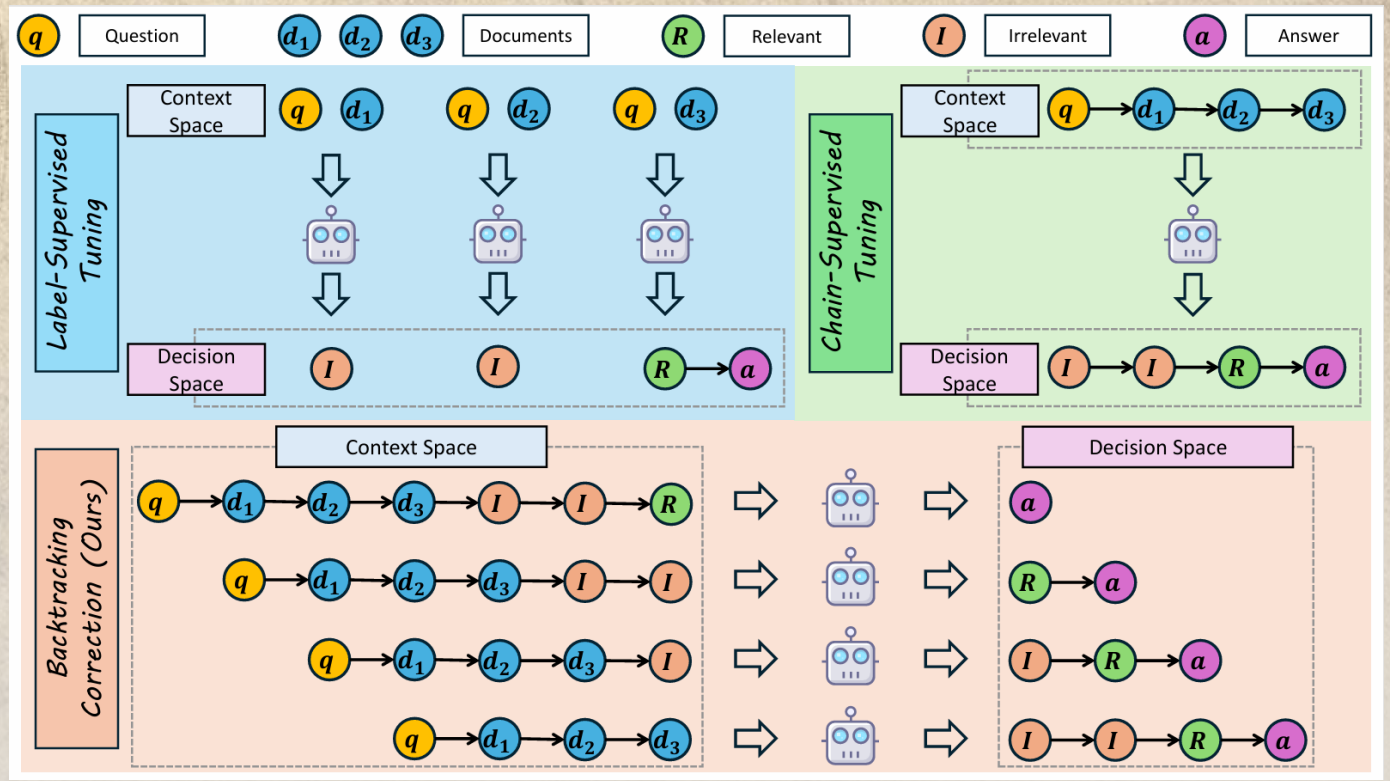
BACKGROUND

- Despite recent progress in Retrieval-Augmented Generation (RAG) achieved by large language models (LLMs), retrievers often recall uncorrelated documents, regarded as "noise" during subsequent text generation. To address this, current methods train LLMs to distinguish between relevant and irrelevant documents using labeled data, enabling them to select the most likely relevant ones as context.
  - **Label-Supervised Tuning**
  - **Chain-Supervised Tuning**



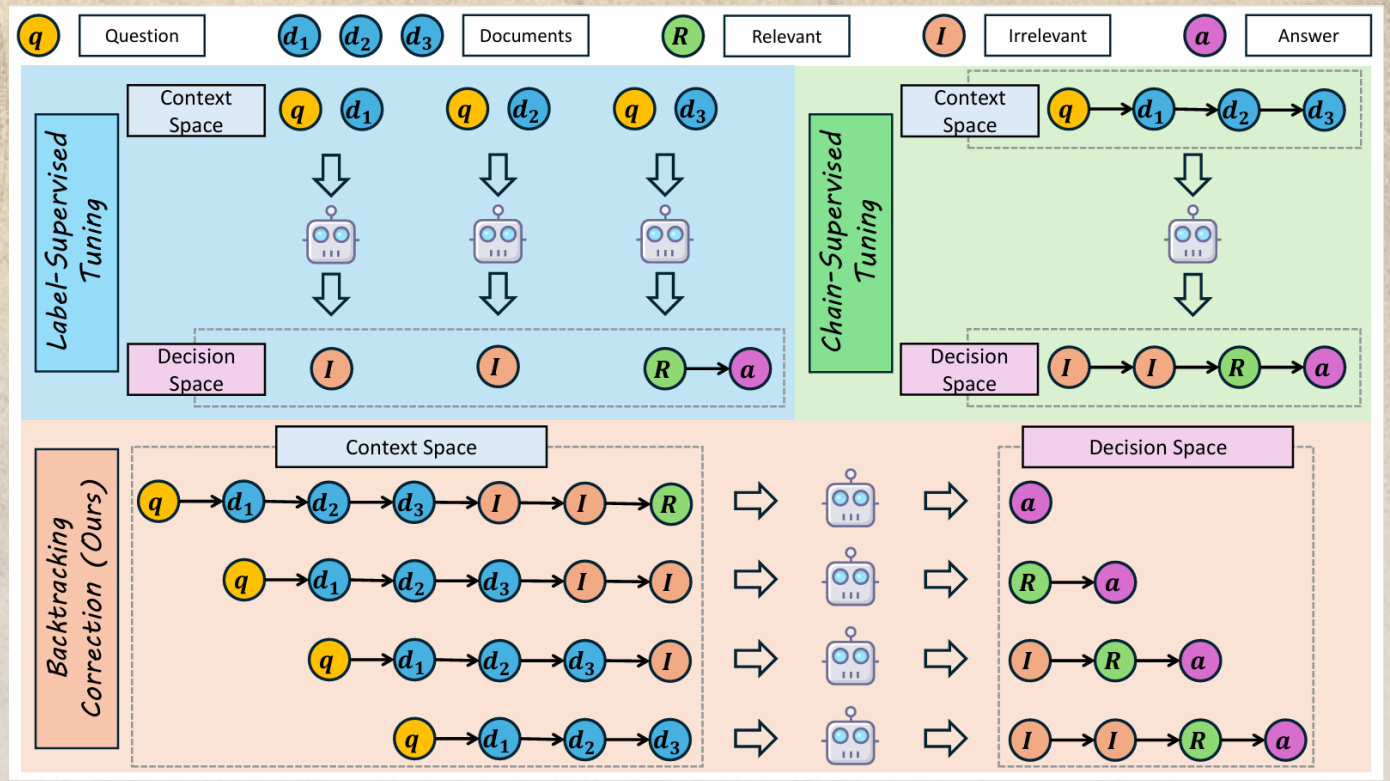
LIMITS

- **Label-Supervised Tuning**—The sparse supervision in the decision-making process limits the model's ability to understand why a retrieved document is relevant or irrelevant.
- **Chain-Supervised Tuning**—Stepwise-annotated data is challenging to collect, and much of it depends on proprietary language models (e.g., ChatGPT, GPT-4).
- Both of them requires **extensive training data with high-quality annotations**, significantly increasing costs.

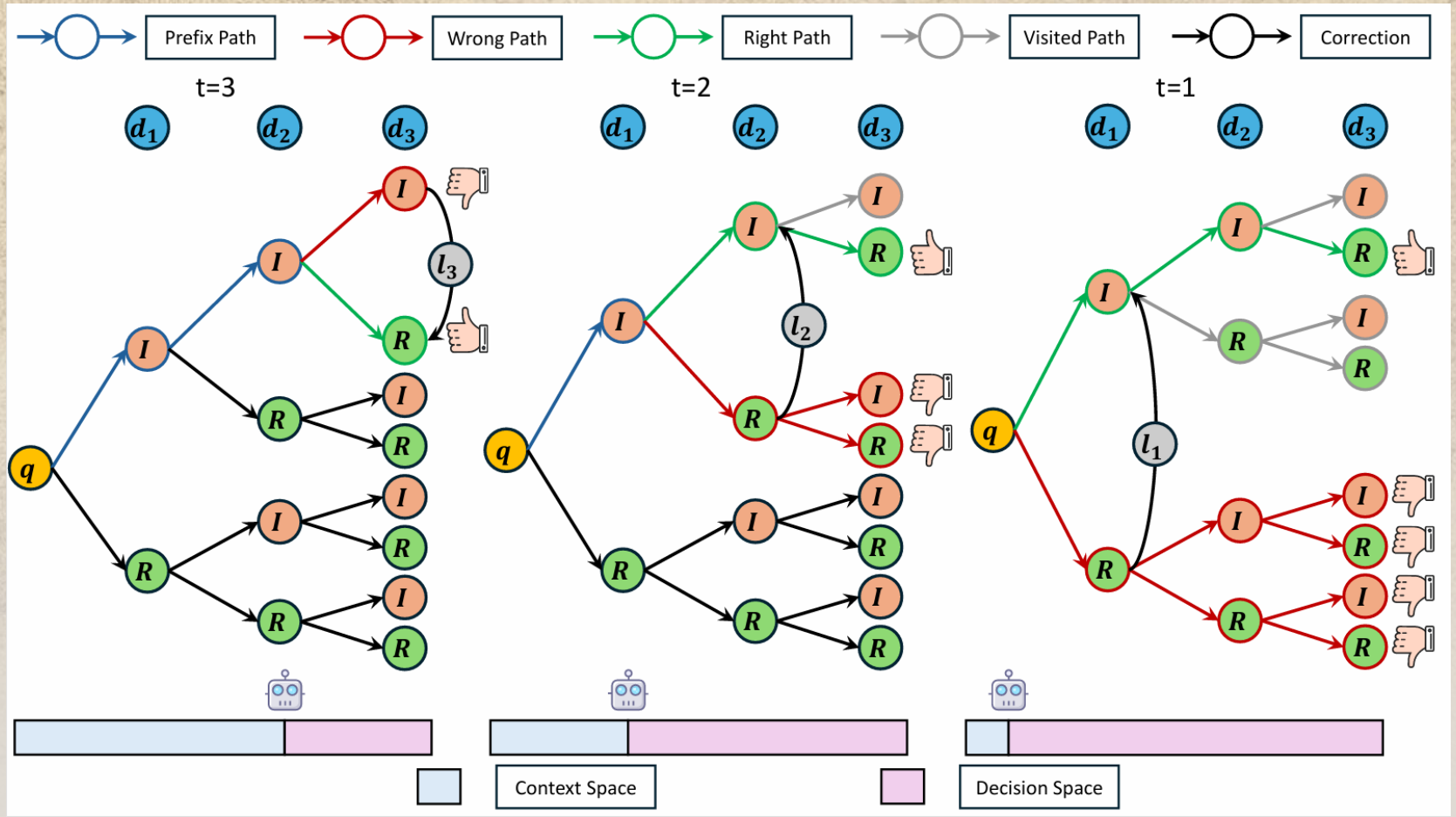


MOTIVATION

We apply **stepwise preference optimization** to reasoning chains. Starting from the final state, we sample model-generated errors and employ self-correction to optimize the current state through Reinforcement Learning (RL), then backtrack to the previous state. Unlike Chain-Supervised Tuning which learns the entire reasoning path, our method optimizes each step by focusing on decision-making based on the current state, as the remaining steps have already been learned. This approach gradually reduces the context space while expanding the decision space, following an easy-to-hard progression.



BACKTRACKING CORRECTION



DERIVATION

Player A:

$$\arg\max_{\theta} \mathbb{E}_{s_{t-1}^{+}, o_t \sim \Pi_\theta(s_{t-1}^{+}), o_t^{+} \sim \Pi_\theta(s_{t-1}^{+}, o_t, L)} \left[r(s_{t-1}^{+}, o_t^{+}) - r(s_{t-1}^{+}, o_t) \right]$$

$$\arg\min_{\theta} -\mathbb{E}_{s_{t-1}^{+}, o_t \sim \Pi_\theta(s_{t-1}^{+}), o_t^{+} \sim \Pi_\theta(s_{t-1}^{+}, o_t, L)} \left[\log \sigma(r(s_{t-1}^{+}, o_t^{+}) - r(s_{t-1}^{+}, o_t)) \right]$$

Player B:

$$\arg\max_{\theta} \mathbb{E}_{s_{t-1}^{+}, o_t \sim \Pi_\theta(s_{t-1}^{+}), o_t^{+} \sim \Pi_\theta(s_{t-1}^{+}, o_t, L)} \left[r\left(s_{t-1}^{+}, o_t^{+}\right) \right] - \beta \mathbb{D}_{KL} \left[\Pi_\theta(o_t^{+} | s_{t-1}^{+}) || \Pi_{ref}(o_t^{+} | s_{t-1}^{+}) \right]$$

$$\Pi_\theta(o_t^{+} | s_{t-1}^{+}) = \frac{1}{S(s_{t-1}^{+})} \Pi_{ref}(o_t^{+} | s_{t-1}^{+}) \exp(\frac{1}{\beta} r(s_{t-1}^{+}, o_t^{+}))$$

$$\Pi_\theta(o_t | s_{t-1}^{+}) = \frac{1}{S(s_{t-1}^{+})} \Pi_{ref}(o_t | s_{t-1}^{+}) \exp(\frac{1}{\beta} r(s_{t-1}^{+}, o_t))$$

DERIVATION

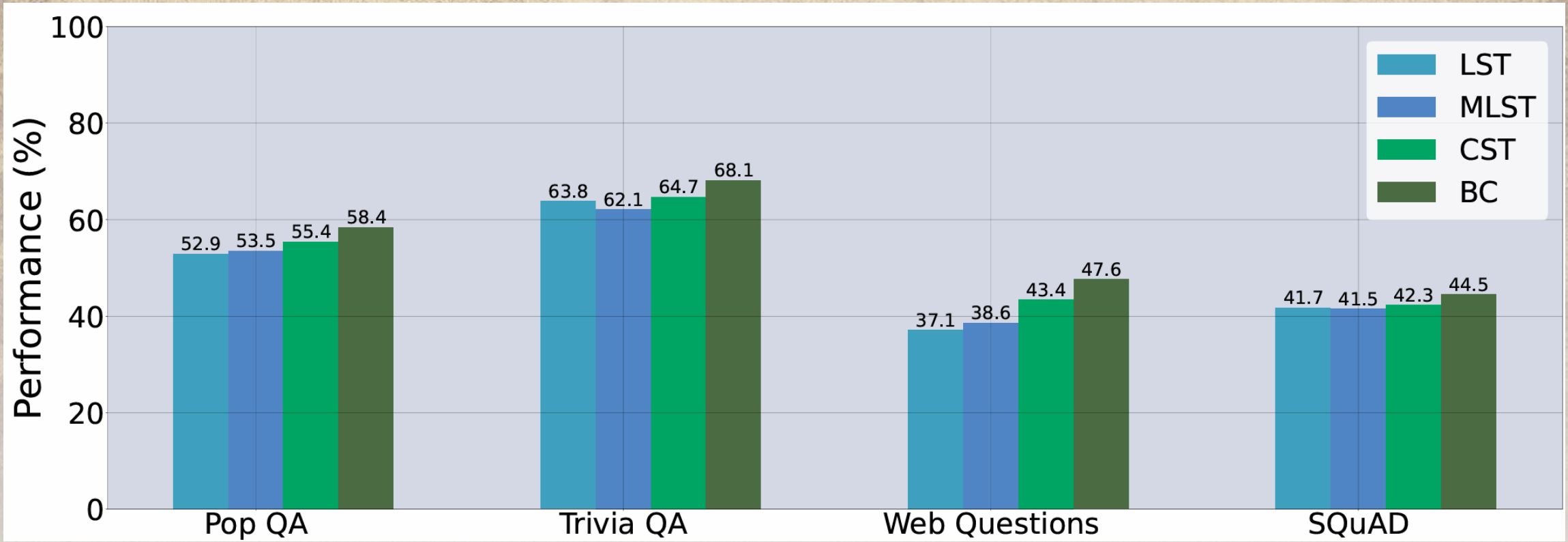
$$r(s_{t-1}^{+}, o_t^{+}) = \beta \log \frac{\Pi_\theta(o_t^{+}|s_{t-1}^{+})}{\Pi_{ref}(o_t^{+}|s_{t-1}^{+})} \qquad r(s_{t-1}^{+}, o_t) = \beta \log \frac{\Pi_\theta(o_t|s_{t-1}^{+})}{\Pi_{ref}(o_t|s_{t-1}^{+})}$$

$$\mathcal{L}_{BC} = \arg\min_\theta -\mathbb{E}\left[\log \sigma(\beta \log \frac{\Pi_\theta(o_t^{+}|s_{t-1}^{+})}{\Pi_{ref}(o_t^{+}|s_{t-1}^{+})} - \beta \log \frac{\Pi_\theta(o_t|s_{t-1}^{+})}{\Pi_{ref}(o_t|s_{t-1}^{+})}\right]$$

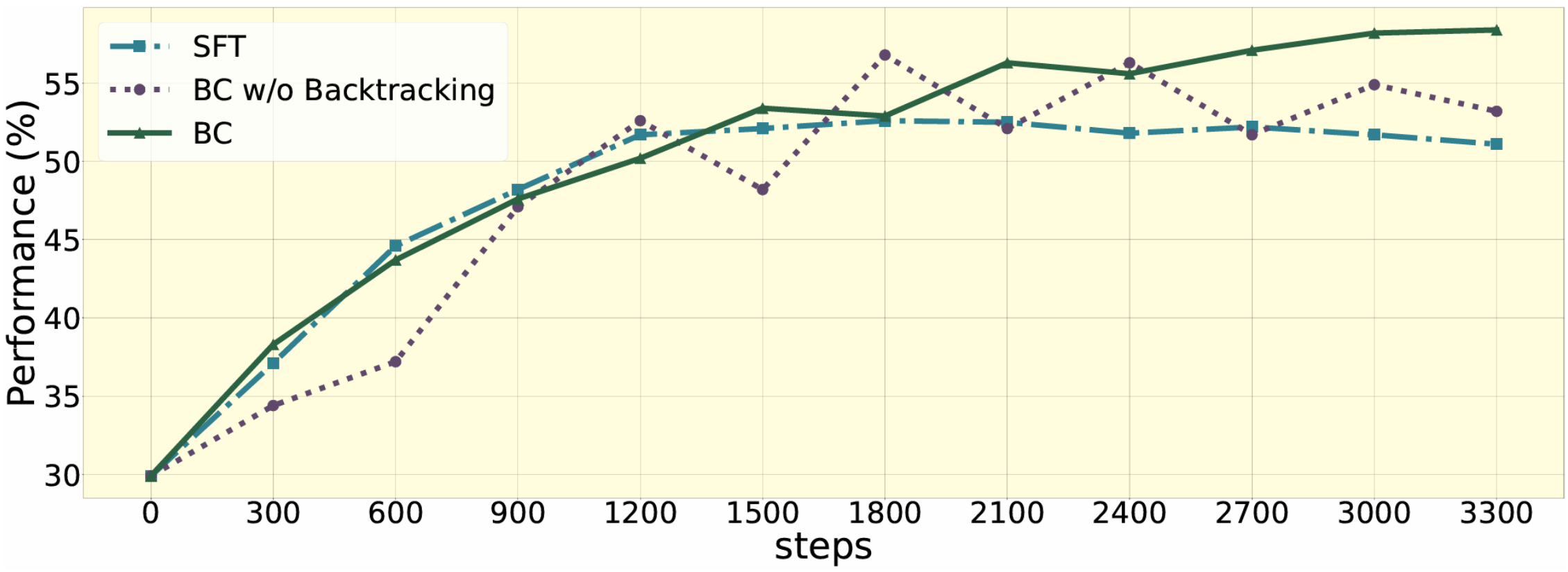
EXPERIMENT

		Pop QA	Trivia QA	Web Questions	SQuAD
Proprietary	ChatGPT	24.7	78.2	57.0	28.5
	GPT-4	38.7	83.4	61.6	35.1
Base	LLaMA2-7B-base	4.5	10.9	6.1	3.7
	LLaMA3-8B-base	5.7	10.8	4.4	3.8
Fine-tuned w/o Retrieval	ChatGLM3-6B	9.5	19.6	14.9	4.3
	Mistral-7B	25.2	66.1	56.8	25.9
	BaiChuan2-7B-chat	25.7	40.7	38.6	13.1
	LLaMA2-7B-chat	25.1	58.7	48.6	19.1
	Alpaca-7B	25.6	49.4	39.6	14.1
Fine-tuned w/ Retrieval	ChatGLM3-6B	42.6	35.7	24.3	19.5
	ToolFormer-6B	-	48.8	26.3	33.8
	Mistral-7B	59.1	71.2	57.6	42.0
	BaiChuan2-7B-chat	56.2	63.3	50.0	38.7
	LLaMA2-7B-chat	54.0	63.1	45.3	37.2
	Alpaca-7B	54.2	58.6	47.3	32.6
	Self-RAG-7B	53.8	62.6	32.4	26.5
	REAR	52.9	71.6	38.4	43.8
	RobustLM-13B	*49.1*	*62.0*	*27.3*	*27.4*
Backtracking Correction	+LLaMA2-7B-base	58.4	68.1	47.6	44.5
	+LLaMA3-8B-base	59.3	70.1	49.3	44.9

EXPERIMENT


Performance (%)
100
80
60
40
20
0
LST
MLST
CST
BC
52.9
53.5
55.4
58.4
63.8
62.1
64.7
68.1
37.1
38.6
43.4
47.6
41.7
41.5
42.3
44.5
Pop QA
Trivia QA
Web Questions
SQuAD

EXPERIMENT


SFT
BC w/o Backtracking
BC
Performance (%)
55
50
45
40
35
30
0
300
600
900
1200
1500
1800
2100
2400
2700
3000
3300
steps

Thanks for your listening!