In-Context Editing: Learning Knowledge from Self-Induced Distributions

Siyuan Qi [1], Bangcheng Yang [1], Kailin Jiang [1,2], Xiaobo Wang [1], Jiaqi Li [1], Yifan Zhong [1,3], Yaodong Yang [1,3], Zilong Zheng [1]

[1] State Key Laboratory of General Artificial Intelligence, BIGAI

[2] University of Science and Technology of China [3] Peking University

Knowledge Editing: Setup

Objective: incorporate new facts into a language model M_θ

Notation: query-response pairs $\{q_i, x_i^*\}$, original response x

Example: q : ”Did Messi win the FIFA world cup?”

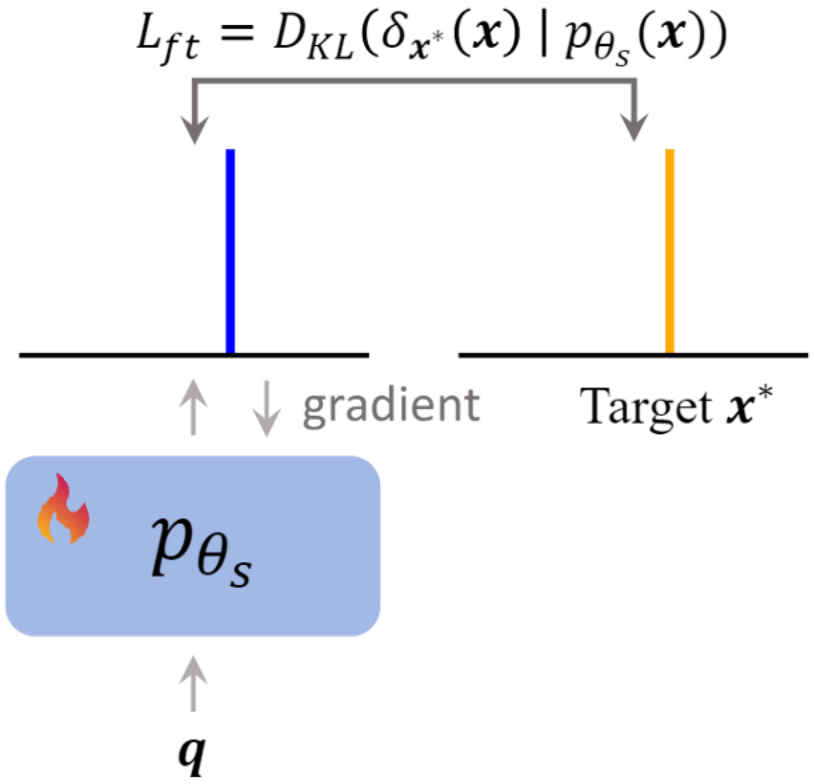
x: ”No".

x^*: ”Yes, in 2022".

Typically done by maximizing the probability $p_\theta(x^*|q)$.

Problem with Fine-Tuning

Problem: overfitting



In deep learning, a large batch size helps mitigate overfitting.

Can we find a target **distribution** that is close to x^*, and contains other useful information?

Fine-Tuning with Sampling

We can sample a distribution!

Employ a softer distribution generated by the model itself in a bootstrapping manner:

$$\mathcal{L}^*_{ft} = D_{KL}(\delta_{\mathbf{x}^*}(x_{1:m})p_\theta(x_{>m}|[\mathbf{q}, \mathbf{x}^*]) \;||\; p_\theta(\mathbf{x}|\mathbf{q})).$$

Observation 1. The objective of fine-tuning with samples **is equivalent to** the objective of traditional fine-tuning, i.e.,

$$\mathcal{L}^*_{ft} = \mathcal{L}_{ft}$$

This implies that the model cannot learn and improve on its own without external information!

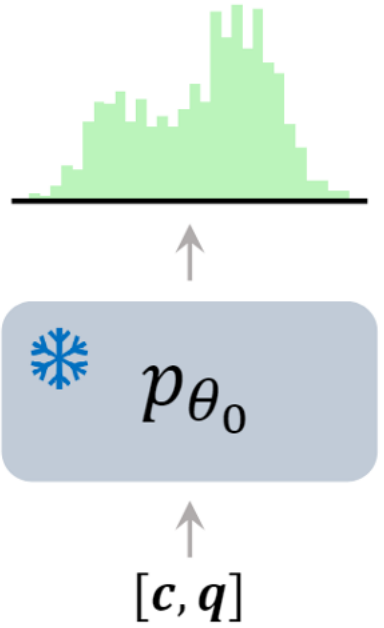
In-Context Learning: Language Models are Few Shot Learners

LLM can perform a task just by conditioning on input-output examples, without optimizing any parameters.

Example: c : ”Messi won the world cup in 2022. ”

q : ”Has Messi won the world cup?”

x: ”Yes, in 2022".



In-Context Editing (ICE)

We need to introduce extra information that

- Guides the model towards a new distribution that aligns with the target
- Maintains similarity to its original distribution

Create the target distribution by adding a context prompt **c**:

$$\mathcal{L}_{\text{ICE}} = D_{\text{KL}}(p_\theta(\mathbf{x}|[\mathbf{c}, \mathbf{q}]) \,||\, p_\theta(\mathbf{x}|\mathbf{q})).$$

The combined objective is:

$$\mathcal{L} = \mathcal{L}_{\text{FT}} + \lambda \mathcal{L}_{\text{ICE}},$$

| Framework

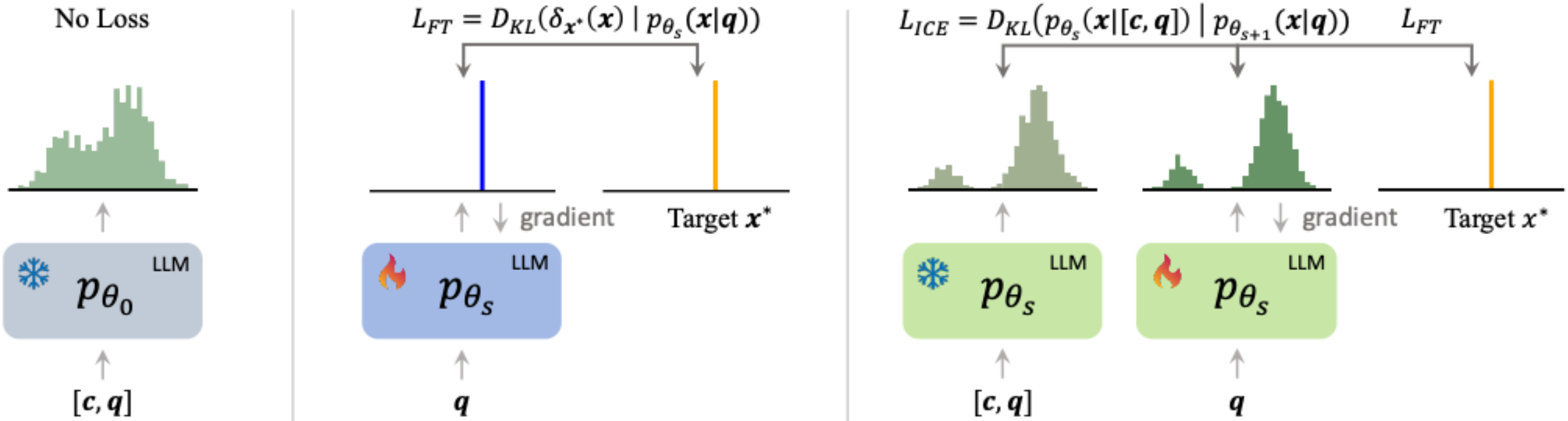
Frozen Tuned c: context q: query x: model output x^*: target output s: step θ_0: initial params θ_s: optimized params

Example query q : Has Messi won the world cup?

Context c : Messi won the world cup in 2022.

Target output x^* : Yes, in 2022.

Original model output x : No.



(a) In-Context Learning (b) Fine-Tuning (c) Consistent In-Context Editing (ICE)

Experiments: Single Edits

Table 1: Main results on knowledge insertion and question-answering datasets of Llama2-7b-chat.

	WikiData_{recent}					ZsRE				
	Edit Succ. ↑	**Portability** ↑	**Locality** ↑	**Fluency** ↑	**PPL**$_r$ ↓	**Edit Succ.** ↑	**Portability** ↑	**Locality** ↑	**Fluency** ↑	**PPL**$_r$ ↓
ROME	97.25	36.58	30.40	581.00	107.47	96.66	52.90	26.61	573.02	53.88
MEMIT	97.03	37.00	29.28	573.06	87.17	95.61	52.73	24.79	563.42	38.67
FT-L	45.63	34.73	34.80	558.91	68.92	43.60	43.90	51.38	560.94	30.36
FT-M	**100.00**	59.28	41.54	**587.17**	70.64	**100.00**	54.47	53.84	**580.10**	27.33
ICE	**100.00**	**61.02**	**46.39**	585.58	**34.08**	**100.00**	**55.52**	**56.97**	562.70	**15.50**

Table 2: Main results on knowledge modification datasets of Llama2-7b-chat.

	WikiBio				$\text{WikiData}_{counterfact}$				
	Edit Succ. ↑	**Locality** ↑	**Fluency** ↑	**PPL**$_r$ ↓	**Edit Succ.** ↑	**Portability** ↑	**Locality** ↑	**Fluency** ↑	**PPL**$_r$ ↓
ROME	95.83	68.38	617.67	3.70	98.68	42.45	21.13	585.40	109.97
MEMIT	94.54	69.96	616.65	3.51	98.13	44.16	19.48	576.26	122.48
FT-L	59.41	28.94	615.50	**1.89**	36.13	29.37	38.37	566.55	89.24
FT-M	**100.00**	35.34	**618.12**	3.67	**100.00**	72.39	40.76	**586.80**	54.71
ICE	99.88	**70.60**	617.88	2.15	**100.00**	**73.49**	**45.88**	583.29	**18.95**

Thanks!

Takeaway:

ICE provides a framework for knowledge editing in language models. It leverages in-context learning to induce target distributions that ensure competitive accuracy and adaptability, all while preserving model integrity through continuous updates.