Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation

Xiaoying Zhang[1*], Baolin Peng[2], Ye Tian[2], Jingyan Zhou[1], Lifeng Jin[2], Linfeng Song[2], Haitao Mi[2], Helen Meng[1,3]

[1]The Chinese University of Hong Kong, Hong Kong [2]Tencent AI Lab, Bellevue [3]Centre for Perceptual and Interactive Intelligence, Hong Kong *Work done during the internship at Tencent AI Lab

Introduction

Problem

- LLMs hold relevant knowledge ("knowing"), yet often struggle with factual inaccuracies, *i.e.,* "hallucinations" ("telling")

Limitations of Existing Approaches

- Necessitate high-quality human factuality annotations
- Employ consistency-based factuality signals, intrinsically linked to the LLM's generation ability

Motivation

- An LLM shows potential in "self-evaluation", *i.e.,* identifying factual inaccuracies within its generated responses, with a reasonable prediction confidence

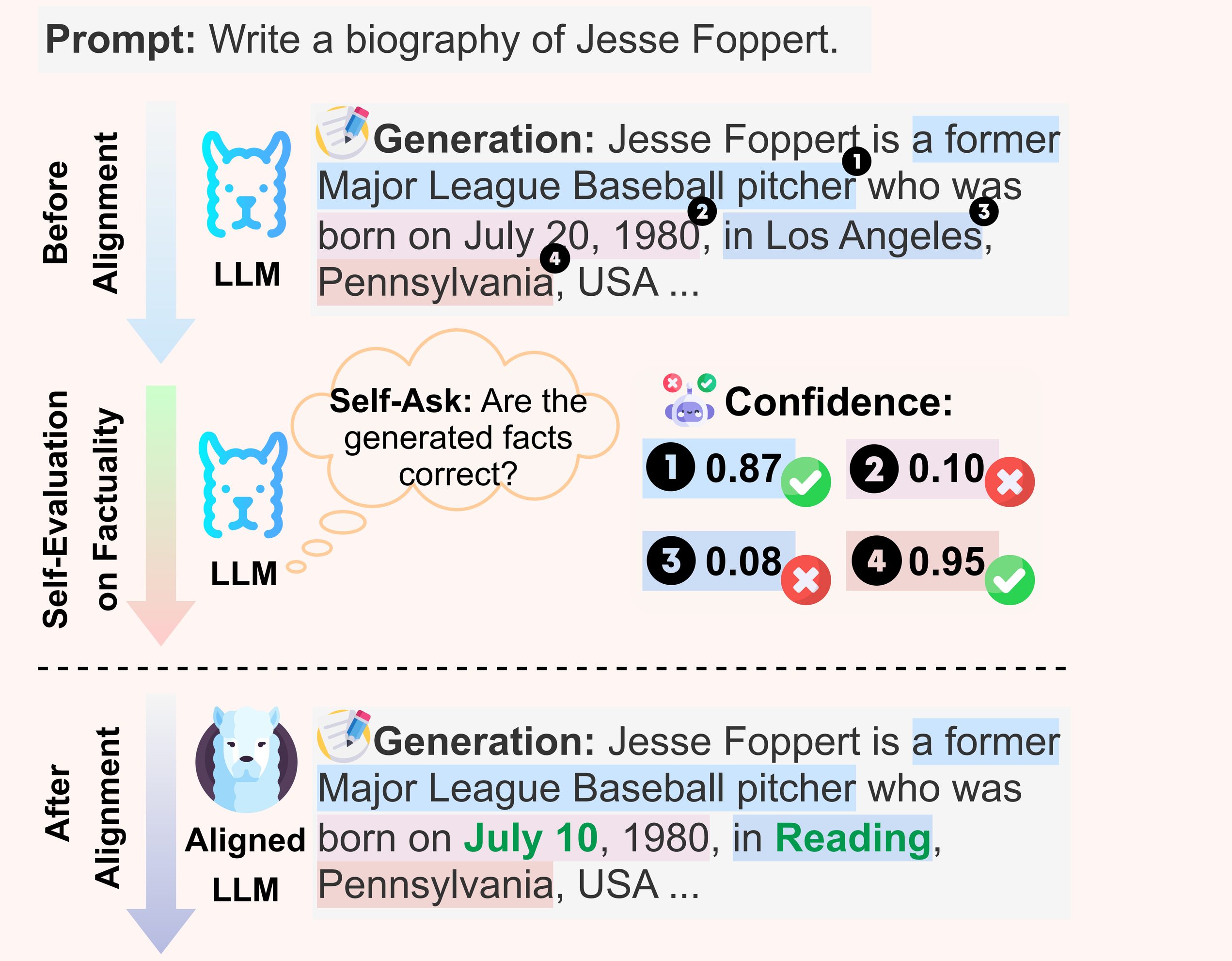

Figure 1. Illustration of *self-alignment for factuality*. Given a prompt to write a biography, before factuality alignment, the LLM generates some facts that are not accurate. Through self-evaluation, the LLM is capable of identifying these inaccurate facts. The feedback from the self-evaluation is used as a reward signal to align the LLM towards factuality. Each fact is highlighted in distinct colors, and the corrected facts are marked in green.

Contributions

- Propose *self-alignment for factuality* framework that leverages an LLM's self-evaluation capability to mitigate hallucinations
- Introduce SK-TUNING to improve an LLM's confidence estimation and calibration, boosting self-evaluation
- Show the efficacy of *Self-Alignment for Factuality* on three knowledge-intensive tasks

Self-Alignment for Factuality

Overview

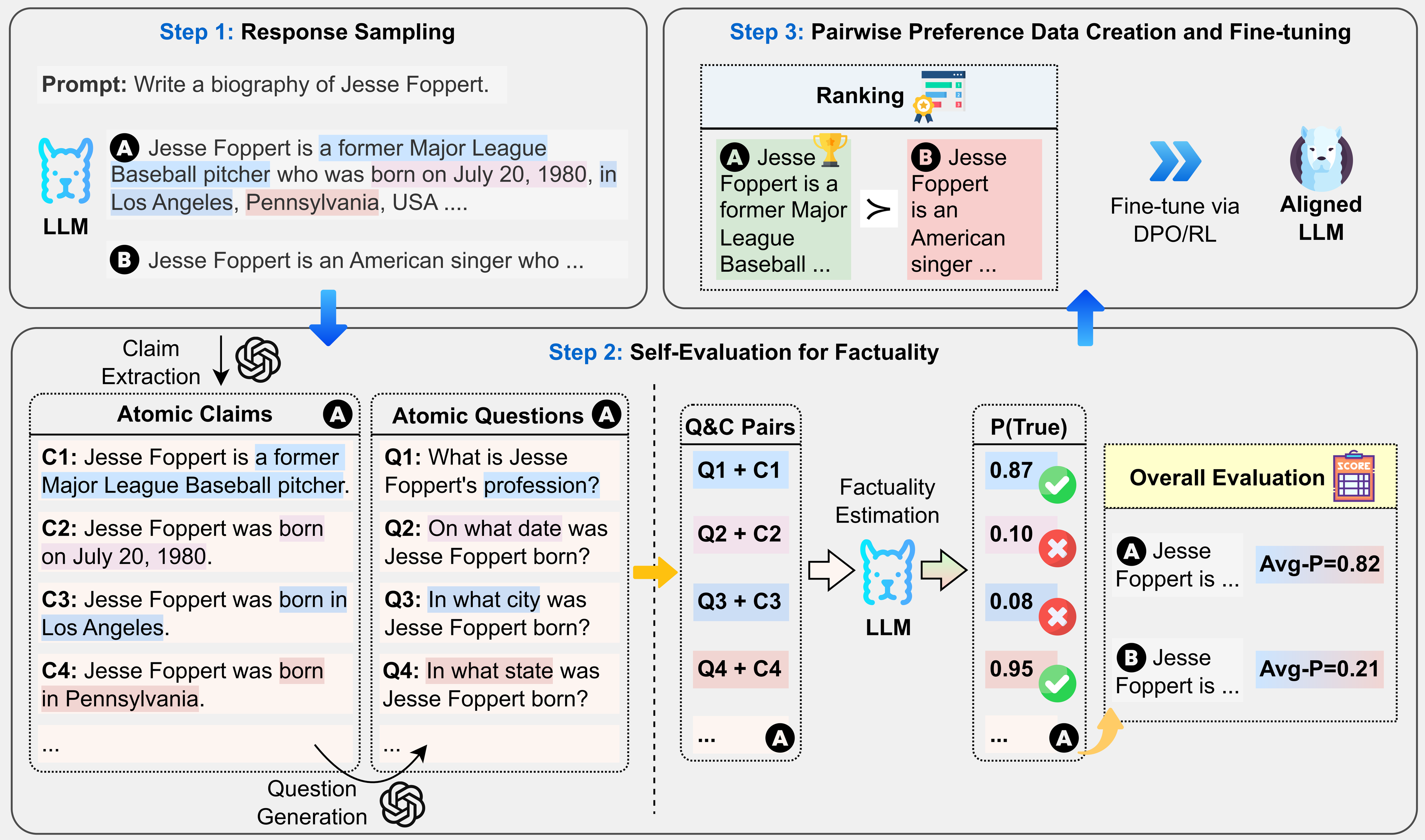

Figure 2. Illustration of *self-alignment for factuality* in the long-form generation task. **(i) Step 1**: Generate initial responses for preference data collection. **(ii) Step 2**: Estimate responses factuality via SELF-EVAL for preference labeling. **(iii) Step 3**: Create preference data and aligning the LLM with DPO.

Factuality Self-Evaluation

- **SELF-EVAL Component**, built on an LLM $\mathcal{M}$, is prompted to assess the validity of $\mathcal{M}$'s response a, using exclusively its internal knowledge, given a prompt q

$$p(\text{True}|q, a) = f_{\mathcal{M}}(q, a) \tag{1}$$

- **Self-Knowledge Tuning (SK-TUNING)** augments LLMs' self-evaluation ability

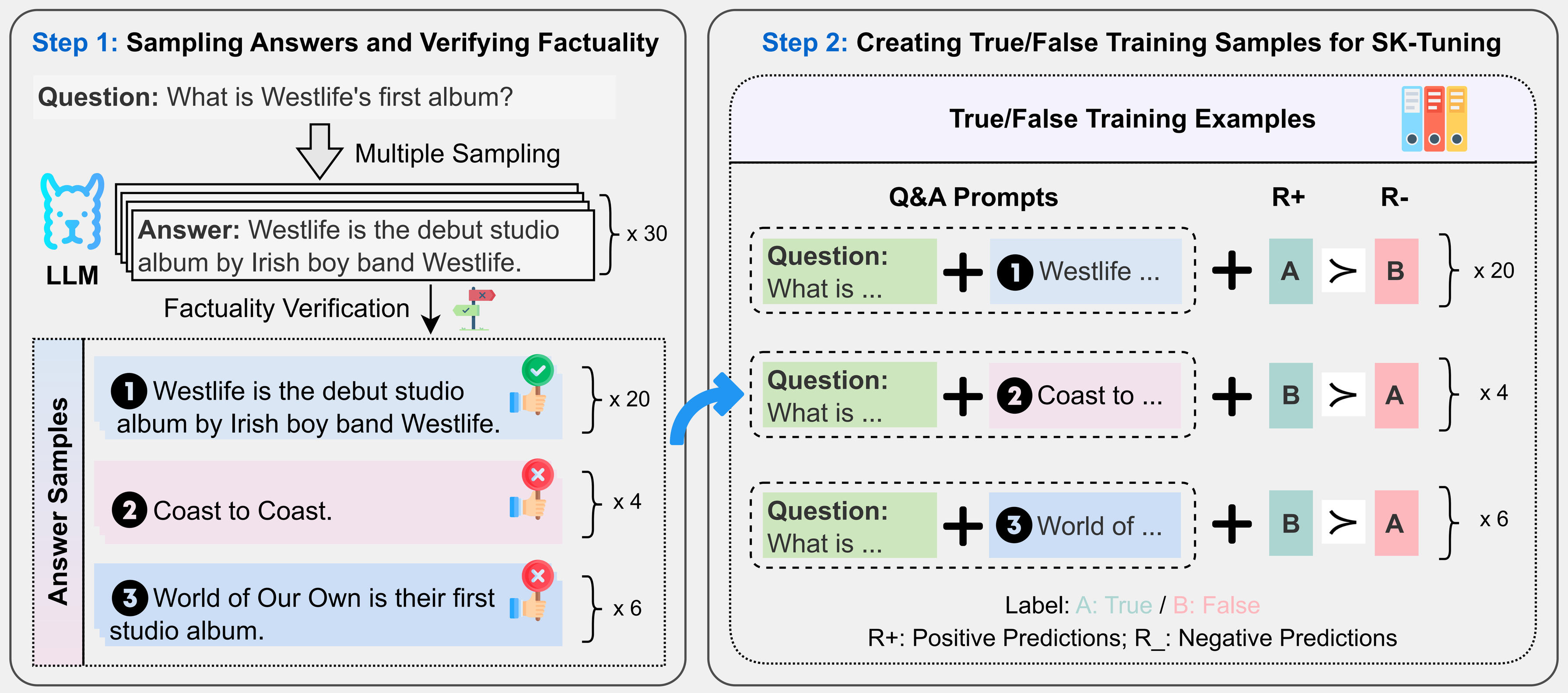

Figure 3. The process of constructing training data $\mathcal{D}_{\psi}$ for SK-TUNING.

$$\mathcal{L}_{\phi} = -\mathbb{E}_{(q,a,r_+,r_-)\sim\mathcal{D}_{\psi}}\left[\log \sigma\left(\log \pi_{\phi}\left(r_+ \mid q, a\right) - \log \pi_{\phi}\left(r_- \mid q, a\right)\right)\right] \tag{2}$$

Experiments

Main Results

Model	Labeled In-dom. Data	TruthfulQA % Acc.	TruthfulQA (Gen.) % True	TruthfulQA (Gen.) % Info	TruthfulQA (Gen.) % True*Info	BioGEN (Long-Form Gen.) # Cor.	BioGEN # Incor.	BioGEN % Res.	BioGEN % FActScore
LLAMA-7B*	-	25.60	30.40	96.30	26.90	7.70	16.92	98.00	30.72
+ SFT*	✓	24.20	47.10	-	36.10	8.52	16.52	98.00	32.17
+ ITI* (Li et al., 2023)	✓	25.90	**49.10**	-	43.50	-	-	-	-
+ DoLA* (Chuang et al., 2023)	✓	32.20	42.10	**98.30**	40.80	7.46	13.70	99.00	33.91
+ FACTTUNE-MC (Tian et al., 2023)		-	-	-	-	**10.98**	21.33	99.00	30.92
Self-Alignment for Factuality (Ours)									
w/ SELF-EVAL-P(TRUE)		36.59	42.88	97.81	41.51	6.21	13.19	100.00	31.33
w/ SELF-EVAL-SKT		**45.48**	47.40	97.26	**45.75**	8.54	**13.49**	**100.00**	**38.28**
LLAMA2-7B	-	28.90	50.41	88.22	39.04	8.84	12.65	99.00	40.54
+ DoLA (Chuang et al., 2023)	✓	31.10	47.53	94.66	42.60	8.74	11.85	72.00	38.99
+ FACTTUNE-MC (Tian et al., 2023)		-	-	-	-	**12.64**	16.16	100.00	42.71
Self-Alignment for Factuality (Ours)									
w/ SELF-EVAL-P(TRUE)		43.15	44.52	94.93	41.10	8.46	**11.17**	**100.00**	42.73
w/ SELF-EVAL-SKT		**44.10**	**55.07**	**98.08**	**53.42**	12.12	14.44	99.00	**46.50**

Table 1. Few-shot evaluation results on three distinct tasks: six-shot prompting results of the MCQA and short-form generation tasks on TruthfulQA, and five-shot prompting results of the long-form generation task on BioGEN.

- *Self-alignment for factuality* is effective on mitigating hallucinations.
- SK-TUNING is helpful to improve factuality estimation with LLM's inherent knowledge.

In-Depth Analysis of SELF-EVAL

Task	Model	Multi-choice QA Datasets: TruthfulQA (Full)	CommonSenseQA	OpenBookQA (Closed)	MedQA	MMLU
Selection (Metric: Acc.)	LLAMA2-7B	25.49	54.30	55.00	30.71	44.76
	SELF-EVAL-P(TRUE)	32.64	64.95	65.40	29.69	43.29
	SELF-EVAL-SKT	**43.97**	**70.43**	**67.40**	**36.37**	**49.88**
Discrimination (Metric: AUROC)	SELF-EVAL-P(TRUE)	51.33	79.76	71.66	52.75	59.52
	SELF-EVAL-SKT	**59.02**	**84.65**	**75.72**	**60.40**	**67.07**

Table 2. Five-shot results on MCQA tasks, following Singhal et al. (2023).

- SK-TUNING shows strong efficacy in improving the LLM's confidence estimation.
- Factuality evaluation is easier than factual generation.
- SK-TUNING improves the LLM's confidence calibration.

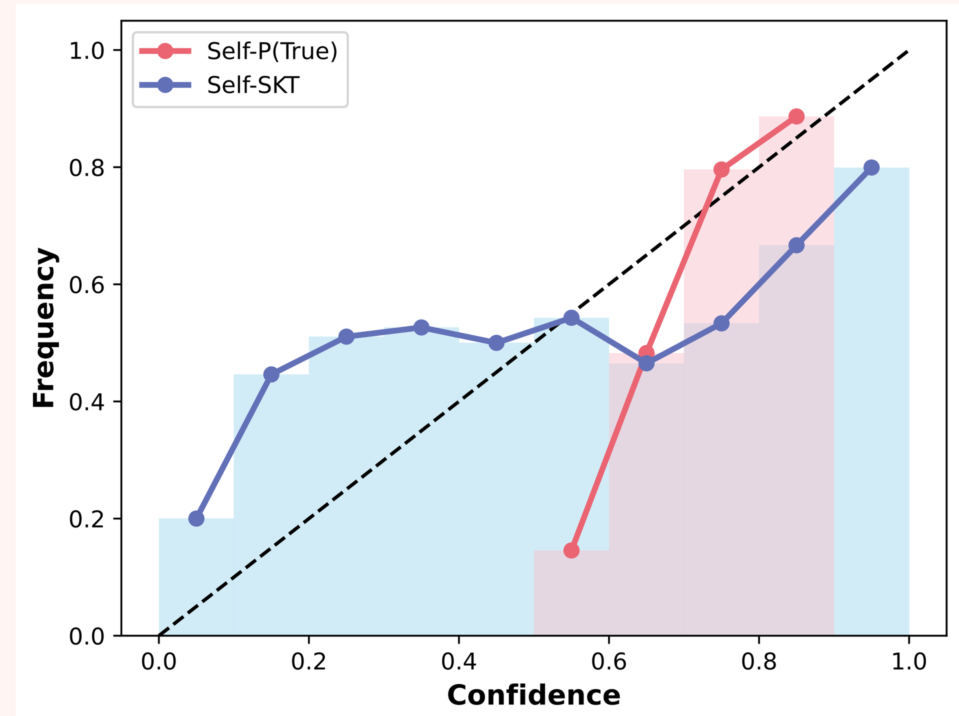

Figure 4. Calibration curves of utilizing SELF-EVAL-P(TRUE) and SELF-EVAL-SKT on LLAMA2-7B in the CommonsenseQA task. Following Kadavath et al. (2022), we plot confidence vs. frequency that a prediction is correct. The dashed line indicates perfect calibration.