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

Anonymous ACL submission

Abstract

Despite showing impressive abilities, large language models (LLMs) often struggle with factual inaccuracies, *i.e.,* "hallucinations", even when they hold relevant knowledge. To mitigate these hallucinations, current approaches typically necessitate high-quality human factuality annotations. In this work, we explore *Self-Alignment for Factuality*, where we leverage the self-evaluation capability of an LLM to provide training signals that steer the model towards factuality. Specifically, we incorporate SELF-EVAL, a self-evaluation component, to prompt an LLM to validate the factuality of its own generated responses solely based on its internal knowledge. Additionally, we design *Self-Knowledge Tuning* (SK-TUNING) to augment the LLM's self-evaluation ability by improving the model's confidence estimation and calibration. We then utilize these self-annotated responses to fine-tune the model via Direct Preference Optimization algorithm. We show that the proposed self-alignment approach substantially enhances factual accuracy over LLAMA family models across three key knowledge-intensive tasks on TruthfulQA and BioGEN. We will release our code and data upon acceptance.

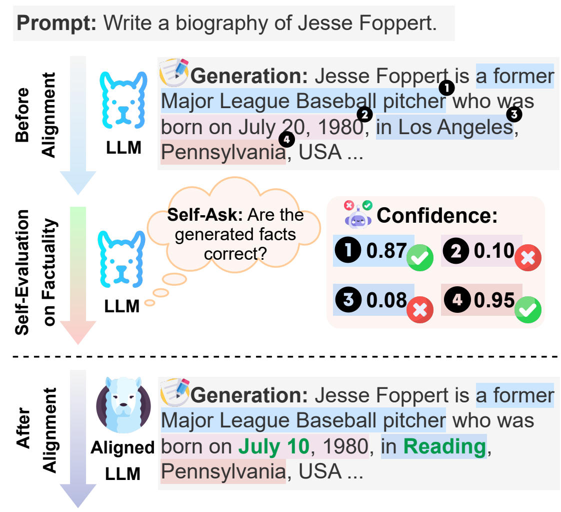


Figure 1: Illustration of *Self-Alignment for Factuality*. Given a prompt to write a biography, before factuality alignment, the LLM generates some facts that are not accurate. Through self-evaluation, the LLM is capable of identifying these inaccurate facts. The feedback from the self-evaluation is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

1 Introduction

Despite exhibiting remarkable proficiency in a diverse range of NLP tasks (Wei et al., 2022; Liu et al., 2023c; Chang et al., 2023), LLMs (OpenAI, 2022, 2023; Touvron et al., 2023b) occasionally generate seemingly plausible yet factually incorrect statements, *i.e.,* "hallucinations" (Huang et al., 2023; Ji et al., 2023; Zhang et al., 2023b; Tonmoy et al., 2024). Such hallucinations can undermine the trustworthiness and practical applicability of LLMs in real-world scenarios, particularly when employed in high-stakes tasks (Liu et al., 2023b).

In this paper, we focus on mitigating a noteworthy type of hallucination, where an LLM holds relevant knowledge in response to a query (*i.e.,* "knows"), yet occasionally falters in conveying accurate information (*i.e.,* "tells") (Li et al., 2023b, 2024). For instance, an LLM might generate an inaccurate response during one inference time but can provide a correct response at another time (Wang et al., 2023a; Manakul et al., 2023; Dhuliawala et al., 2023). This gap between "knowing" and "telling" (Saunders et al., 2022; Kadavath et al., 2022; Chen et al., 2023a) significantly undermines the potential of LLMs to accurately convey the knowledge acquired during the pre-training phase.

A few studies (Li et al., 2023b; Chuang et al., 2023; Zhang et al., 2023a) edit the model's internal representations towards "factuality" directions, using domain-specific annotated data. Meanwhile, acknowledging the inadequacy of the training objective—maximum likelihood estimation (MLE)—in accurately capturing factuality (Ouyang et al., 2022; Allen-Zhu and Li, 2023; Azaria and Mitchell, 2023; Tian et al., 2023a), a recent study (Tian et al., 2023a) introduces the LLM's internal factuality signals as training rewards to guide the models towards factuality. Given that the origin of a LLM's hallucinations is intrinsically linked to its confidence[1] (Huang et al., 2023), Tian et al. (2023a) employs consistency-based confidence regarding the factual correctness over the generate responses (Kuhn et al., 2023; Manakul et al., 2023) as the factuality signals. Nevertheless, such consistency-based confidence remains rely on the model's generation ability, which might be non-reflective on model's internal knowledge. Despite the challenges faced by an LLM in directly "telling" the correct response, it has showed potential in "evaluating" its generated responses (Kadavath et al., 2022; Saunders et al., 2022). As depicted in Figure 1, the LLM is capable of identifying factual inaccuracies within the responses it generates, with a reasonable prediction confidence. Such self-evaluation, *i.e.,* directly prompting the model itself about internal knowledge awareness, might be a more effective approach to factuality estimation.

In this paper, we introduce a self-alignment framework, *Self-Alignment for Factuality*, which harnesses an LLM's self-evaluation capability to mitigate hallucinations. Our approach encourages an LLM to generate prediction confidence scores pertaining to the factuality of its own generated responses through self-asking. Subsequently, these scores are utilized as reward signals to fine-tune the model using the Direct Preference Optimization (DPO) algorithm (Rafailov et al., 2023). Specifically, we incorporate a factuality self-evaluation component, SELF-EVAL, which prompts the LLM to directly validate its responses based on its internal knowledge. To bolster the LLM's universal self-evaluation ability, we introduce SK-TUNING to enhance the LLM's internal knowledge awareness, *i.e.,* prediction confidence estimation and calibration[2] (Guo et al., 2017; Tian et al., 2023b), through sufficient tuning across heterogeneous knowledge-oriented tasks.

We assess the effectiveness of the proposed *Self-Alignment for Factuality* framework on three crucial knowledge-extensive tasks for LLMs, namely Multi-Choice Question-Answering (MCQA), short-form open-ended generation, and long-form open-ended generation, using two benchmark datasets: TruthfulQA (Lin et al., 2022) and BioGEN (Min et al., 2023a). The results show that, **solely relying on the model's internal knowledge**, *Self-Alignment for Factuality* significantly enhances the factual accuracy of LLAMA family models (Touvron et al., 2023a,b) across all three tasks, notably surpassing the representation-editing methods (Chuang et al., 2023; Li et al., 2023c) and the recent work with consistency-based confidence (Tian et al., 2023a).

In summary, our contributions are three-fold:

- We propose *Self-Alignment for Factuality*, a self-alignment strategy that leverages an LLM's self-evaluation capability to mitigate the model's hallucinations.
- We introduce SK-TUNING to improve an LLM's confidence estimation and calibration, thereby enhancing its self-evaluation ability.
- We show the efficacy of *Self-Alignment for Factuality* on three crucial tasks using TruthfulQA and BioGEN, significantly improving factual precision over all compared methods.

2 Related work

Hallucinations in LLMs. Hallucinations in LLMs occur when generated content, is seemingly plausible, however deviates from actual world knowledge (Chen et al., 2023b; Li et al., 2023a; Zhang et al., 2023b; Tonmoy et al., 2024). In this study, we align with the perspective that an LLM's acquired knowledge should mirror established facts (Yang et al., 2023). We focus on a specific type of "unfaithful hallucination" where LLMs produce factually incorrect statements, even when possessing relevant knowledge (Evans et al., 2021; Park et al., 2023; Li et al., 2023b). Rather than broadly targeting the enhancement of LLMs' factuality (Sun et al., 2023; Zhou et al., 2023a; Lightman et al., 2023; Peng et al., 2023; Li et al., 2023d; Mallen et al., 2023; Varshney et al., 2023), our goal is to align LLMs to reliably convey accurate information when they have sufficient knowledge.

Hallucination Mitigation. Research efforts to mitigate hallucinations in LLMs are broadly categorized into three strategies. (*i*) In post-hoc correction, recent works have explored self-consistency techniques for model refinement (Kadavath et al.,

[1]A lower confidence score corresponds to a greater likelihood of hallucinated facts.

[2]The confidence in a prediction is expected to accurately reflect the probability that the prediction is correct.

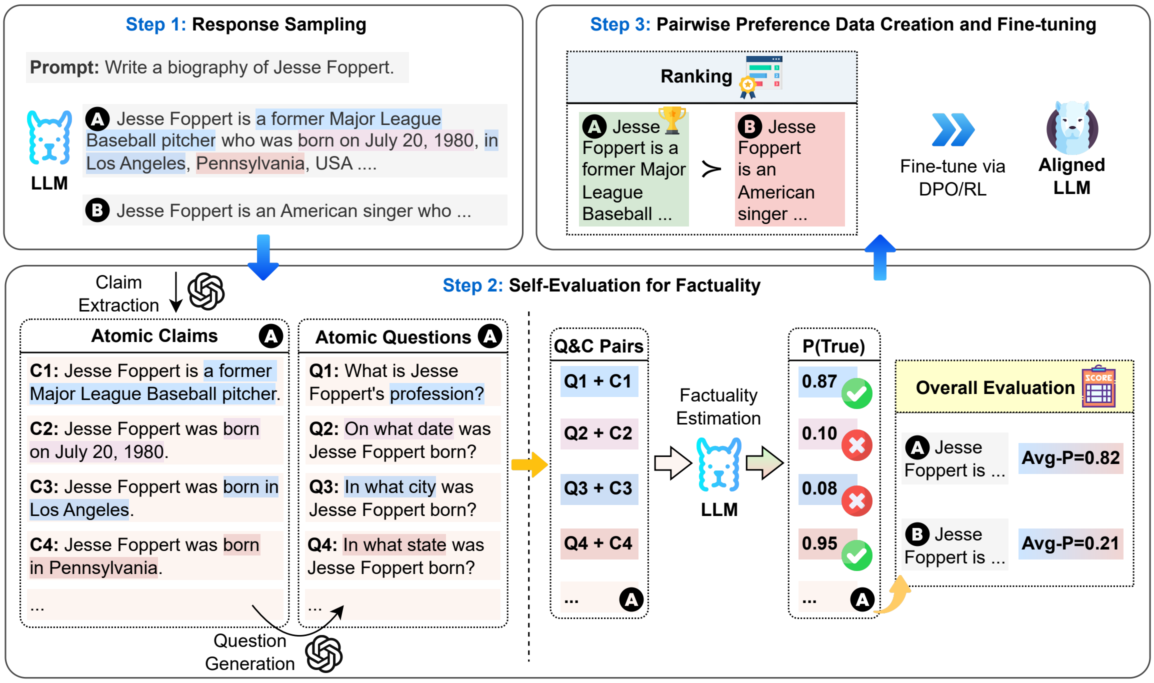


Figure 2: A diagram illustrating the three steps of our Self-Alignment for Factuality in long-form generation task.

2022; Ren et al., 2023; Tian et al., 2023b; Madaan et al., 2023; Dhuliawala et al., 2023; Wang et al., 2023a). These methods, rooted in uncertainty estimation, aim at improving factual accuracy by analyzing the consistency among multiple responses generated by the LLM. However, their effectiveness varies with the model's intrinsic capabilities. (ii) Inference-time intervention approaches focus on manipulating LLMs' internal representations to guide them towards factuality (Li et al., 2023b; Chuang et al., 2023; Li et al., 2023c; Zhang et al., 2023a). These methods show promise but often rely on domain-specific data, limiting their generalizability. (iii) Alignment training, as a third strategy, directly optimizes LLMs to produce factual statements. This involves either supervised fine-tuning with high-quality datasets (Wang et al., 2023b) or reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). While effective, these methods can be resource-intensive due to the need for extensive human labels.

Our research parallels two significant studies in the field by Yang et al. (2023) and Tian et al. (2023a). While Yang et al. (2023) focus on honesty-based fine-tuning, empowering LLMs to admit limitations by acknowledging "I don't know", our Self-Alignment for Factuality approach is distinctively geared towards guiding LLMs to articulate truthful information when they hold pertinent knowledge. In contrast to Tian et al. (2023a), which relies on a consistency-based method for confidence estimation, our work introduces SELF-EVAL-SKT, which is trained on a broad spectrum of heterogeneous data, and designed to enhance confidence estimation capabilities significantly. Experimental results from our study demonstrate a notable improvement in the accuracy and reliability of factual information presented by LLMs. We provide a brief summary in Appendix A.

3 Self-Alignment for Factuality

In this section, we introduce the proposed framework. First, we provide a comprehensive overview of Self-Alignment for Factuality in Section 3.1. Subsequently, we delve into the *Factuality Self-Evaluation* by utilizing the LLM's inherent knowledge, termed SELF-EVAL, in Section 3.2. Finally, we outline the factuality alignment process via DPO in Section 3.3.

3.1 Overview

Self-Alignment for Factuality generally operates in the following three steps, as depicted in Figure 2:

Step 1: Generating Initial Responses for Preference Data Collection. For a given prompt x, we generate multiple candidate responses $\{y_m\}_{m=1}^{M}$, where M represents the sample size. These are produced from a base LLM guided by a policy $\pi_{\text{ref}}(y \mid x)$. To ensure the generation of coherent and relevant responses, we employ few-shot examples as prompts.

Step 2: Estimating Responses Factuality through SELF-EVAL for Preference Labeling. In this step, we evaluate the factuality of generated candidate responses $\{y_m\}_{m=1}^{M}$ for a given prompt x by leveraging the intrinsic knowledge of LLMs. In long-form response generation tasks, *e.g.,* crafting a biography in Figure 2, a response often contains a mix of factually accurate and inaccurate information. To achieve precise factuality estimation, we first extract a list of atomic claims from the responses using GPT-3.5-turbo (OpenAI, 2022; Min et al., 2023a), with each claim representing a distinct piece of information (Liu et al., 2023d). Subsequently, we employ GPT-3.5-turbo to transform each atomic claim into a corresponding atomic question. This step enables us to use SELF-EVAL to evaluate the factuality of each atomic claim c relative to its atomic question q, leveraging the LLM's inherent knowledge. This process is denoted as $p(\text{True}|q, c)$. Finally, we calculate the average of the obtained factuality scores for individual claims, resulting in a final factuality score, Avg-p(True), for the candidate response.

Step 3: Creating Preference Data and Aligning LLM with DPO. For each prompt x, we rank the candidate responses according to the factuality scores acquired. Then, we select the top α responses as the preferred responses y_w and the remaining responses as the dis-preferred ones y_l, resulting in a set of preference pairs $\mathcal{D} = \{(x, y_w, y_l)\}$. The total number of preference pairs is $\alpha M * (1 - \alpha)M - K$, where K represents the number of pairs with equal scores. Finally, we align the LLM with these preference data via DPO.

3.2 Factuality Self-Evaluation

At the core of *Self-Alignment for Factuality* is the design of a factuality self-evaluation component, denoted as SELF-EVAL. Given a prompt q and a LLM $\mathcal{M}$, SELF-EVAL, built on $\mathcal{M}$, is tasked with assessing the validity of $\mathcal{M}$'s response a, leveraging exclusively its own internal knowledge. This process is quantified as the probability $p(\text{True}|q, a)$, which is formulated as follows:

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

There exist various methodologies to parameterize $f_{\mathcal{M}}(q, a)$. For instance, LLMs have demonstrated capabilities in discerning the extent of their knowledge (Kadavath et al., 2022). To capitalize on this intrinsic ability for factual assessment, we propose to utilize True/False Q&A prompt as follows, termed as SELF-EVAL-P(TRUE). This prompt facilitates the LLM's self-evaluation of factuality based on its inherent knowledge

```
Instruction: Please evaluate the truthfulness
of the proposed answer based on the given
question and internal knowledge.
<Few-shot Prompts>
Question: <Question>
Proposed Answer: <Answer>
Is the proposed answer:
A. True
B. False
The proposed answer is:
```

where we anticipate either "A" or "B" as an answer. The probability p(True) signifies the extent to which an LLM deems a generated answer (claim) valid. In line with Kadavath et al. (2022), we prepend few-shot prompts to encourage well-structured answers.

Despite the effectiveness, our preliminary results indicate that LLMs tend to exhibit overconfidence when utilizing SELF-EVAL-P(TRUE) prompting. This observation is in line with the findings presented by Tian et al. (2023b). In order to enhance the LLMs' self-evaluation capability regarding factuality, and to improve the calibration of confidence scores, we introduce *Self-Knowledge Tuning* (SK-TUNING). It is designed to augment LLMs' ability to accurately assess the factuality of their own generated responses across a diverse range of tasks. Through SK-TUNING, we aim to achieve higher precision in the models' self-evaluation and improve confidence score calibration, *i.e.,* assigning higher confidence scores to responses with a greater likelihood of being factually correct. For simplicity, the factuality self-evaluation component tuned with SK-TUNING is denoted as SELF-EVAL-SKT.

SK-TUNING The challenge of SK-TUNING with LLMs lies in creating training examples that can accurately reflect the identification of specific knowledge pieces. To address this, we propose to build self-knowledge-guided training data, as illustrated in Figure 3. Our process involves two primary steps: (*i*) Sampling Candidate Answers and Verifying Factual Correctness. For each question q, we generate a set of candidate answers $\{a_k\}_{k=1}^{K}$ using few-shot prompting. We then assess the factual correctness of each answer by comparing it to the golden answer, employing the bidirectional entailment approach with the Deberta-Large-MNLI model (He et al., 2021). Answers that are semantically equivalent to the golden answer are labeled as

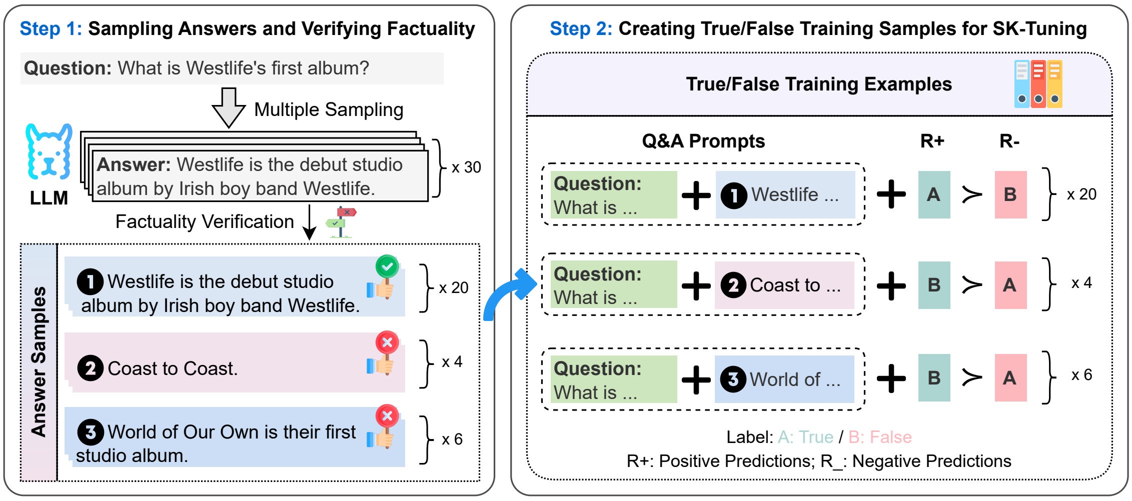


Figure 3: The process of constructing training data for SK-TUNING.

factually correct a_c, while others are deemed incorrect a_i. (*ii*) Creating True/False Training Examples. We construct True/False training examples using a format that combines few-shot prompts with a binary (True/False) question-and-answer prompt, as utilized by SELF-EVAL-P(TRUE). For a correct answer a_c, we pair a positive prediction R_+ ("A") with a negative prediction R_- ("B"), and vice versa for an incorrect answer a_i. This approach results in a dataset $\mathcal{D}_\psi$ comprising prediction pairs, with duplicates maintained to approximate the model's knowledge over the question, which helps improving the confidence calibration (Appendix G).

Following the assembly of $\mathcal{D}_\psi$, we proceed to fine-tune the LLM on this pairwise prediction data. The fine-tuning aims to minimize a loss function specifically designed to enhance the model's ability to leverage its inherent knowledge for accurate self-knowledge evaluation, as follows:

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

where π_ϕ is the LLM trained for factuality estimation and σ denotes the logistic function.

3.3 Alignment Tuning with DPO

After obtaining the preference data over candidate responses $\mathcal{D} = \{(x, y_w, y_l)\}$, where each tuple represents a choice preference between winning and losing responses to few-shot prompts, we proceed to the stage of alignment tuning for improving factuality. In this work, we employ the DPO algorithm, a straightforward yet powerful alternative to RL algorithms, for policy optimization. Specifically, DPO employs a standard cross-entropy objective for direct policy optimization, as follows:

$$\begin{aligned}\mathcal{L}_\theta = -\,\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\log\sigma\Bigg(\beta\log\frac{\pi_\theta\left(y_w\mid x\right)}{\pi_{\text{ref}}\left(y_w\mid x\right)} \\ -\beta\log\frac{\pi_\theta\left(y_l\mid x\right)}{\pi_{\text{ref}}\left(y_l\mid x\right)}\Bigg)\Bigg],\end{aligned} \tag{3}$$

where the model policy π_θ is initialized from the base reference policy π_{ref}, β is a parameter controlling the deviation from π_{ref}, and σ denotes the logistic function.

4 Experiments

In this section, we evaluate the efficacy of our proposed framework across three knowledge-intensive tasks: MCQA, short-form and long-form open-ended generation, following Li et al. (2023b).

4.1 Setup

Datasets and Evaluation Metrics. For the MCQA task, we utilize the TruthfulQA dataset (Lin et al., 2022). For short-form generation task, we use generation formulation of TruthfulQA, and BioGEN for the long-form one (Min et al., 2023b). In evaluating performance on TruthfulQA, we report Accuracy for the MCQA task, alongside metrics of truthfulness (True), informativeness (Info), and a composite True*Info score, all evaluated using a fine-tuned GPT-3 model (Lin et al., 2022). For assessments on BioGEN, we present the FActScore percentage and the Respond ratio. Moreover, we quantify the correctness of generated content by reporting the number of accurate (cor) and inaccurate facts (incor) per response, following Tian et al. (2023a). Comprehensive descriptions of tasks, datasets, and evaluation criteria are detailed in Appendix B. Note that for open-ended text generation tasks, self-alignment approaches only use the prompts provided in the datasets.

Baselines. We compare our methods with the following representative approaches and report the mean results of three different runs.

- **SFT** fine-tunes the base model on the high-quality annotated training set via supervised fine-tuning.
- **ITI** (Li et al., 2023b) edits internal representations by shifting model activations along learned factuality-related directions.
- **DoLa** (Chuang et al., 2023) edits internal representations by contrasting output distributions from different layers within the model.
- **FactTune-MC** (Tian et al., 2023a) optimizes the base model using DPO on the preference data labeled with consistency-based confidence scores.

Implementation Details. (*i*) *Implementation of the Self-Alignment for Factuality framework:* We employ LLaMA-7B (Touvron et al., 2023a) and LLaMA2-7B (Touvron et al., 2023b) as the base LLMs and fine-tune these models on the constructed preference data for five epochs. More details are shown in Appendix C. (*ii*) *Implementation of* SK-Tuning*:* We utilize Wikipedia, which is a frequently employed pre-training data source for LLMs (Touvron et al., 2023b; Shi et al., 2023), and the BIG-bench dataset (Srivastava et al., 2023) in our study. Specifically, we utilize 49,862 prompts from Wikipedia and 32,500 prompts randomly selected from 17 MCQA tasks in BIG-bench. More fine-tuning details are provided in Appendix C.

4.2 Main Results

Table 1 presents the main evaluation results across three tasks. We have the following observations:

Self-alignment for factuality is effective on mitigating hallucinations. Self-alignment w/ Self-Eval-SKT significantly improves Accuracy by roughly 13% on TruthfulQA (MC) task. Moreover, self-alignment w/ Self-Eval-SKT attains the highest True*Info (45.75% for LLaMA-7B and 53.42% for LLaMA2-7B) on TruthfulQA (short-form generation) task and exhibits substantial improvement in FActScore (approximately 4%) for BioGEN (long-form generation) task. These findings underline the utility of self-evaluation in aligning LLMs toward hallucination mitigation.

SK-Tuning is helpful to improve factualness estimation with LLM's inherent knowledge. Enhancing self-evaluation capabilities through SK-Tuning enables self-alignment with Self-Eval-SKT to achieve higher factual accuracy compared to w/ Self-Eval-P(True). In addition, self-alignment w/ Self-Eval-SKT considerably outperforms w/ Self-Eval-P(True) regarding True*Info (surpassing by 12%) and FActScore (exceeding by 4%). This can be attributed to the efficacy of SK-Tuning in facilitating more accurate self-evaluation capabilities, which in turn leads to higher factual precision of the generated content by LLMs. We provide an in-depth analysis in Section D. Moreover, self-alignment w/ Self-Eval-SKT evidently surpasses FactTune-MC[4], emphasizing the advantages of our proposed Self-Eval-SKT for confidence estimation over the consistency-based approach. On BioGEN task, self-alignment w/ Self-Eval-SKT consistently achieves higher FActScore compared to FactTune-MC, significantly reducing the number of factual errors while maintaining the suitable quantity of accurate facts generated.

Self-alignment w/ Self-Eval-SKT considerably surpasses the representation-editing methods – ITI and DoLa, which necessitate labeled in-domain data, by obtaining the highest True*Info while exhibiting remarkable True and Info scores on TruthfulQA. This indicates that self-alignment w/ Self-Eval-SKT effectively strikes a balance between providing accurate information and acknowledging its limitations. Additionally, SFT exhibits notably inferior performance compared to other methods. This observation aligns with the findings in Li et al. (2023b); Tian et al. (2023a).

4.3 Pairwise Evaluation

We conduct pairwise comparisons on 100 generated biographies in Section 4.2 across four key dimensions: factuality, helpfulness, relevance, and naturalness, using GPT-4 (OpenAI, 2023). The prompt employed can be found in Appendix E. In Table 2, we find that self-alignment w/ Self-Eval-SKT significantly outperforms FactTune-MC

[3]We use the default QA prompt as in Lin et al. (2022); Li et al. (2023b); Chuang et al. (2023) on `TruthfulQA` and the prompt generated by GPT-4 (OpenAI, 2023) on `BioGEN` (Table 10 in Appendix C).

[4]It is worth noting that the discrepancy between the reported results of FactTune-MC and the results presented in Tian et al. (2023a) may be attributed to the considerably small number of training prompts in this study.

Model	Labeled In-dom. Data	TruthfulQA	TruthfulQA (Gen.)			BioGEN (Long-Form Gen.)			
		% Acc.	% True	% Info	% True*Info	# Cor.	# Incor.	% Res.	% 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., 2023b)	✓	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., 2023a)		-	-	-	-	**10.98**	21.33	99.00	30.92
Self-Alignment for Factuality (Ours)									
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., 2023a)		-	-	-	-	**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: 6-shot prompting results of the MCQA and short-form generation tasks on TruthfulQA, and 5-shot prompting results of the long-form generation task on BioGEN.[3]Results on TruthfulQA marked with an asterisk are cited from Li et al. (2023b) and Chuang et al. (2023). The remaining results of DOLA and FACTTUNE-MC are reproduced following Chuang et al. (2023) and Tian et al. (2023a).

Pairwise Comparisons	BioGEN (% Win Rates)			
	Fact.	Help.	Rele.	Natu.
vs. FACTTUNE-MC	72.00	66.00	68.00	67.00
vs. w/ SELF-EVAL-P(TRUE)	65.00	68.00	62.00	51.00

Table 2: Results of pairwise comparisons on BioGEN across four dimensions: factuality, helpfulness, relevance and naturalness. The first and second row present the win rates of self-alignment w/ SELF-EVAL-SKT against FACTTUNE-MC and self-alignment w/ SELF-EVAL-P(TRUE), respectively.

Model	TruthfulQA			
	% MC acc.	% True	% Info	% True*Info
LLAMA-7B	25.60	30.40	96.30	26.90
w/ SE	37.26	33.29	**98.22**	31.78
w/ USC	38.63	41.92	96.16	38.77
w/ SELF-EVAL-SKT	**45.48**	**47.40**	97.26	**45.75**
LLAMA2-7B	28.90	50.41	88.22	39.04
w/ SE	42.47	44.38	97.81	42.33
w/ USC	40.55	44.66	**98.77**	43.84
w/ SELF-EVAL-SKT	**44.10**	**55.07**	98.08	**53.42**

Table 3: Results of *Self-Alignment for Factuality* that employ various approaches for confidence estimation.

and self-alignment w/ SELF-EVAL-P(TRUE) (with LLAMA2-7B as the base model) with considerable winning rates across all dimensions. Examples of qualitative studies are shown in Appendix F.

4.4 Self-Alignment with Varying Factuality Estimation Methods

Setup. To further examine the effectiveness of Self-Alignment for Factuality, we introduce two variants, *i.e.,* self-alignment w/ SE and w/ USC, which adopt Semantic Equivalence (Kuhn et al., 2023) and Universal Self-Consistency (Chen et al., 2023c) for confidence estimation, respectively. In particular, (i) *self-alignment w/* SE clusters the initial responses based on semantic equivalence and then uses the largest cluster of semantically equivalent responses as the preferred responses, while treating the remaining responses as dis-preferred ones. (ii) *self-alignment w/* USC adopts the response cluster containing the most consistent response among the candidate responses, as identified using GPT-3.5-turbo, as the preferred responses.

Results. Despite exhibiting lower performance than self-alignment with SELF-EVAL-SKT, both variants consistently improve factuality over the base models in the MCQA and open-ended generation tasks, which further reveals the effectiveness of SK-TUNING on improving factuality estimation. These promising results suggest a potential groundwork for investigations into the area of self-alignment for enhancing factuality.

5 In-dpeth Analysis of SELF-EVAL

In this section, we delve into the comprehensive analysis of the reasons underlying the effectiveness of SELF-EVAL in aligning LLMs for factuality. Specifically, following Kadavath et al. (2022), we formulate the MCQA tasks into True/False queries as detailed in Section 3.2, each accompanied by a mix of the correct answer and several incorrect answers. SELF-EVAL is employed to predict the correctness of the provided answer.

5.1 Setup

Datasets. We employ five well-studied MCQA datasets: TruthfulQA, CommonSenseQA (Talmor et al., 2019), OpenBookQA (Closed-Form) (Mihaylov et al., 2018), MedQA (USMLE) (Pal et al.,

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 4: Following Taylor et al. (2022); Singhal et al. (2023), we report the 5-shot results on MCQA tasks. Note that the results of Llama2-7B are reported using the lettered choices format (examples are provided in Appendix D Table 7), as Kadavath et al. (2022); Rae et al. (2022) suggest that models are well-calibrated in this format.

2022), and Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021).

Evaluation Metrics. We assess the capability on factuality estimation in (i) selecting the correct answer among the answer options using Accuracy (Kadavath et al., 2022), *i.e.,* the probability that the correct answer has the highest confidence score among all answer options; (ii) distinguishing between the correct answer and a randomly sampled incorrect answer using Area Under the Receiver Operating Characteristic curve (AUROC) (Kuhn et al., 2023), *i.e.,* the probability that the correct answer has a higher confidence score than a randomly chosen incorrect answer.

5.2 Results

SK-Tuning shows strong efficacy in improving the model's confidence estimation. We present the results in Table 4. Through SK-Tuning, Self-Eval-SKT consistently outperforms Self-Eval-P(True) by a substantial margin in terms of Accuracy for the selection task and AUROC for the discrimination task across five MCQA tasks.

Factuality evaluation is easier than factual generation. We additionally include the answer selection results of the base model Llama2-7B (*i.e.,* generation) for a comprehensive analysis. We observe that Self-Eval-SKT (*i.e.,* evaluation) significantly improves Accuracy over Llama2-7B across five MCQA tasks, *e.g.,* by over 16% on CommonSenseQA and 12% on OpenBookQA (Closed-Form). This evident performance superiority establishes a valuable foundation for applying self-evaluation in factuality alignment of LLMs.

SK-Tuning improves the model's confidence calibration. Following Kadavath et al. (2022); Tian et al. (2023b), we further explore the confidence calibration (Guo et al., 2017). In Figure 4, we present the calibration curves for utilizing Self-Eval-P(True) and Self-Eval-SKT on Llama2-7B in the CommonSenseQA task. With SK-Tuning, Self-Eval-SKT (represented by the blue line) attains superior calibration of the LLM compared to Self-Eval-P(True) (depicted by the pink line), which shows substantial overconfidence, *i.e.,* the frequency within each bin tends to fall below its corresponding confidence.

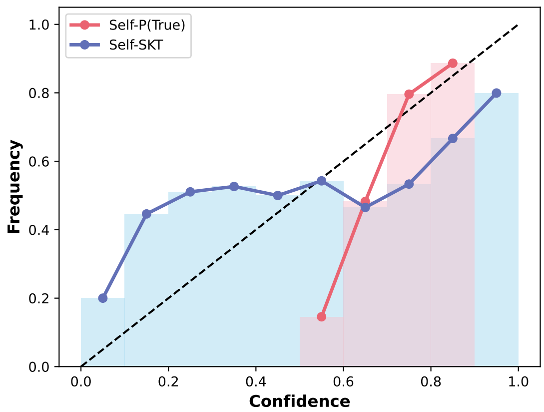


Figure 4: Calibration curves of utilizing Self-Eval-P(True) and Self-Eval-SKT on Llama2-7B in the `CommonsenseQA` task. Following Kadavath et al. (2022), we plot confidence vs. frequency that a prediction is correct. The dashed line indicates perfect calibration.

6 Conclusion

In this paper, we introduce *Self-Alignment for Factuality*, a framework that capitalizes on an LLM's self-evaluation ability to mitigate hallucinations. Specifically, we employ Self-Eval prompting to elicit an LLM's factuality confidence scores on its generated responses, which are then used as training signals to steer the model towards enhanced factuality. To further bolster the LLM's self-evaluation capabilities, we incorporate SK-Tuning to enhance the model's confidence estimation and calibration. Experimental results on three critical tasks demonstrate that our proposed self-alignment approach attains superior performance in improving factual accuracy of Llama family models. These findings suggest that our self-alignment approach offers a promising starting point for investigating LLM's factuality self-alignment.

Limitations

Although we have achieved promising experimental results, we regard these findings as preliminary, given that numerous avenues remain to be explored in this area.

Combining with decoding-based strategies. Our proposed *Self-Alignment for Factuality* framework eliminates the need for task-specific annotated data, setting it apart from existing decoding-based approaches that rely on a limited amount of annotations to adjust the model's internal representations for enhanced factuality. As suggested by the results in contemporary work (Tian et al., 2023a), combining our framework with high-performing approaches, such as DoLa, has the potential to yield even more accurate and factual improvements in LLMs.

Experimenting on different LLMs. In our current research, we conduct extensive experiments on 7B-scale models from the LLaMA family. As the promising findings in Kadavath et al. (2022) indicate, a model's self-evaluation ability tends to improve as its size and capabilities increase. Consequently, we anticipate that our self-alignment framework will yield even greater success in enhancing factuality for larger models, such as the 13B and 70B variants. Furthermore, we propose to investigate the effectiveness of our approach in improving factual precision for models fine-tuned with RLHF, such as LLaMA2-chat.

Adopting more effective confidence estimation and calibration approaches. The comprehensive experimental results detailed in Section 4.2 and Section 4.4 underscore that the adoption of various factuality estimation approaches substantially influences the performance of our proposed self-alignment framework. Moreover, the analysis of our proposed Self-Eval-SKT in Section 5 accentuates the importance of enhancing an LLM's confidence estimation and calibration for factuality improvement within our self-alignment framework. While our proposed SK-Tuning has proven highly effective in refining the model's confidence estimation and calibration, future research may benefit from exploring more efficient confidence estimation and calibration methods (Guo et al., 2017; Tian et al., 2023b; Zhu et al., 2023; Chen et al., 2023a; Shrivastava et al., 2023; Liu et al., 2023a).

Ethics Statement

The motivation of this research is aligned with the ethical principles, to enhance the trustworthiness and avoid LLMs from generating misleading information. Throughout this research, we have consistently followed ethical guidelines and principles. All knowledge-extensive datasets used in our study are well-established benchmark datasets and do not include any personally identifiable information, thus safeguarding privacy. In addition, the prompts employed by GPT-4 for the data collection on BioGEN tasks and model evaluation are meticulously crafted to exclude any language that discriminates against specific individuals or groups (Gallegos et al., 2023; Zhou et al., 2023b), and to prevent any negative impact on users' well-being. Examples of these carefully designed prompts can be found in Appendix E, H. Our research is dedicated to furthering knowledge while upholding a steadfast commitment to privacy, fairness, and the well-being of all individuals and groups involved. Additionally, future research efforts could explore the use of the OpenAI moderation API[5] to systematically filter out inappropriate system responses.

[5]https://platform.openai.com/docs/guides/moderation/overview

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.2, knowledge manipulation.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan O Arik, Tomas Pfister, and Somesh Jha. 2023a. Adaptation with self-evaluation to improve selective prediction in llms.

Xiang Chen, Duanzheng Song, Honghao Gui, Chengxi Wang, Ningyu Zhang, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2023b. Unveiling the siren's song: Towards reliable fact-conflicting hallucination detection. *CoRR*, abs/2310.12086.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023c. Universal self-consistency for large language model generation.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful ai: Developing and governing ai that does not lie.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023. Inspecting and editing knowledge representations in language models.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023a. Generative judge for evaluating alignment.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023c. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2023d. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Xin Liu, Muhammad Khalifa, and Lu Wang. 2023a. Litcab: Lightweight calibration of language models on outputs of varied lengths.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023b. Trustworthy llms: a survey and guideline for evaluating large language models' alignment.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023c. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023d. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023a. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023b. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

OpenAI. 2022. large-scale generative pre-training model for conversation. *OpenAI blog*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2023. Ai deception: A survey of examples, risks, and potential solutions.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis & insights from training gopher.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. 2023. Self-evaluation improves selective generation in large language models.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models.

Kumar Shridhar, Koustuv Sinha, Andrew Cohen, Tianlu Wang, Ping Yu, Ram Pasunuru, Mrinmaya Sachan, Jason Weston, and Asli Celikyilmaz. 2023. The art of llm refinement: Ask, refine, and trust.

Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. 2023. Llamas know what gpts don't show: Surrogate models for confidence estimation.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

Robyn Speer and Joanna Lowry-Duda. 2017. ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89, Vancouver, Canada. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, and Adrià Garriga-Alonso et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented rlhf.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023a. Fine-tuning language models for factuality.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023b. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pretrained transformer language models.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2023a. Alleviating hallucinations of large language models through induced hallucinations.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. Lima: Less is more for alignment.

Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2023b. Rethinking machine ethics – can llms perform moral reasoning through the lens of moral theories?

Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. 2023. Rethinking confidence calibration for failure prediction.

A A brief summary of recent hallucination mitigation approaches.

In Table 5, we provide a brief summary of recent hallucination mitigation approaches that are mostly related to ours.

B Data statistics and task descriptions for main experiments.

Specifically, we construct the BioGEN dataset with the prompts in the format: "`Question: Write a biography of <Entity>.`" where the entities are sampled from Min et al. (2023b). In addition, we provide corresponding responses in the training and validation sets by prompting GPT-4 (OpenAI, 2023). We provide task descriptions and detailed information about the datasets in Table 6.

C Implementation details.

1. Implementing the Self-Alignment for Factuality framework. Taking into account the minor differences when applying Self-Alignment for Factuality to the three tasks, namely, MCQA, short-form text generation, and long-form text generation, we discuss them individually for each stage:

Step 1.1: Generating Initial Responses for Preference Data Collection. (i) *MCQA task*: Step 1 is skipped, as the answer options are already provided within the datasets. (ii) *Generation tasks* (*i.e.,* both short-form and long-form generation tasks): Given a task prompt, we generate 30 candidate response samples via 5-shot prompting at temperature $T = 1, 0.9, 0.8$.

Step 1.2: Estimating Responses Factuality through SELF-EVAL for Preference Labeling. (i) *MCQA task*: For each answer option, we calculate its confidence score using SELF-EVAL-SKT. (ii) *Generation tasks*: For the short-form generation task, we directly compute the confidence score for

Method	Category	Mitigation Approach	Detection Approach	Domain-specific Annotated Data
SELF-REFINE (Madaan et al., 2023)	Post-hoc correction	Self-refinement	Self-consistency	
CoVe (Dhuliawala et al., 2023)	Post-hoc correction	Self-refinement	Self-consistency	
ART (Shridhar et al., 2023)	Post-hoc correction	Self-refinement	Fine-tuned Evaluator	✓
ITI (Li et al., 2023b)	Inference-time intervention	Representation editing (Hernandez et al., 2023)	-	✓
CD (Li et al., 2023c)	Inference-time intervention	Representation editing	-	✓
DOLA (Chuang et al., 2023)	Inference-time intervention	Representation editing	-	✓
ICD (Zhang et al., 2023a)	Inference-time intervention	Representation editing	-	✓
HONESTY-TUNE (Yang et al., 2023)	Alignment training	Supervised fine-tuning	-	✓
FACTTUNE-MC (Tian et al., 2023a)	Alignment training	Fine-tuning with DPO	Sampling-based confidence estimation	
Self-Alignment for Factuality (Ours)	Alignment training	Fine-tuning with DPO	Self-knowledge-enhanced confidence estimation	

Table 5: A brief summary of recent hallucination mitigation approaches that are closely related to our work. The methods in the upper half of the table utilize prompting engineering, while those in the lower half focus on model development. (MCQA: multiple-choice question answering, Gen.: open-end text generation, Honesty-Tune: honesty-oriented fine-tuning.)

Task	Task Definition	Datasets	Required Knowledge	Statistical Info. (# train, # dev, # test)	Metrics
MCQA Prediction	Given a question and 4-5 answer choices, select the only correct answer.	TruthfulQA	38 categories, *e.g.*, health, law, finance, ...	41, 41, 735	Accuracy
Short-Form Generation	Given a question, generate an appropriate answer (1-2 sentences) or respond "I have no comment".	TruthfulQA	38 categories, *e.g.*, health, law, finance, ...	41, 41, 735	Fine-tuned GPT-3 ("GPT-judge" / "GPT-info") (Lin et al., 2022)
Long-Form Generation	Given a prompt that contains a particular people entity, write a short biography (1-2 paragraphs) or respond "I could not find ...".	BioGEN	People biographies, covering nationalities, professions, ...	50, 33, 100	FActScore (Min et al., 2023b)

Table 6: Task descriptions and dataset information for main experiments. Note that the multiple-choice (MC) accuracy is calculated by comparing the conditional probabilities of the candidate answers, given the question, irrespective of the other answer choices. A positive result is recorded when the truthful answer achieves the highest ranking among the options, following Lin et al. (2022); Li et al. (2023b); Chuang et al. (2023); Touvron et al. (2023b).

each candidate response using SELF-EVAL-SKT. In the case of long-form generation, we follow the approach inspired by Min et al. (2023a). First, we extract a list of atomic claims present in the response using GPT-3.5 (OpenAI, 2022). Next, we employ GPT-3.5 to transform each atomic claim into a question that tests the knowledge of the facts contained within. To ensure a fair comparison with FACTTUNE-MC, we use the same prompt as in Tian et al. (2023a). to convert the atomic claims into questions. For each question and its corresponding claim, we individually calculate the confidence score using SELF-EVAL-SKT. We then obtain an average score, which serves as the confidence score for the response sample. Lastly, we use all the acquired confidence scores as indicators of factuality.

Step 1.3: Creating Preference Data and Aligning LLM with DPO. (*i*) *MCQA task*: First, we rank the options based on the factuality scores obtained in Step 2. Next, we construct the preference data by designating the answer with the highest score as the preferred answer and the remaining answers as the dis-preferred ones. Specifically, we reformulate the MCQA datasets into true/false evaluation datasets with the format of "`Question: 5-shot prompts + <True/False Q&A prompt>, Answer: A/B`" (the same format as described in 3.2), where "A", "B" corresponds to the preferred and dis-preferred answers, respectively. Finally, we fine-tune the base model on these preference data using DPO. Note that during evaluation, we choose the answer option with the highest $p(\text{True})$ as the selected option. (*ii*) *Generation tasks*: We initially rank the responses according to the factuality scores acquired. Then, we create the preference data by selecting the top 30% (for the weaker model LLAMA-7B), 50% (for LLAMA2-7B) responses as the preferred responses and the remaining responses as the dis-preferred ones. Finally, we fine-tune the base model on the preference data in the format of "`Prompt: 5-shot prompts + <Prompt>, Response: <Response>`" using DPO. Specifically, we fine-tune the base model on 8 32G Tesla V100 for 5 epochs, with the batch size as 8 and learning rate as 5e-6, the parameter β as 0.1. Note that we report all the evaluation results at the temperature $T = 1$.

2. Implementing SK-TUNING. Given that Wikipedia is a frequently employed pre-training data source for current LLMs (Zhang et al., 2022; Touvron et al., 2023b; OpenAI, 2023), and the BIG-bench dataset (Srivastava et al., 2023) concentrates on tasks considered to surpass the current language models' capabilities, we utilize these two datasets in our study. Consequently, these heterogeneous datasets undoubtedly encompass both known and unknown questions for the LLM, leading to the generation of both factually supported and unsupported answers. Specifically, we utilize 49,862 prompts from Wikipedia and 32,500 prompts randomly sampled from 17 MCQA tasks in BIG-bench.

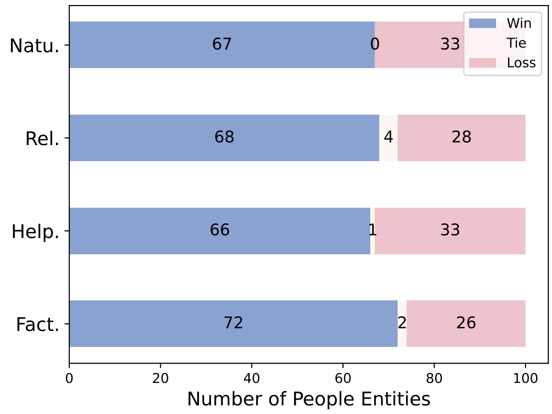
Given a task prompt, we generate 30 candidate response samples via 10-shot prompting at temperature $T = 1$. As described in Section 3.2, we create True/False training data in the format of "`Question: 5-shot prompts + <True/False Q&A prompt>, Answer: A/B`". As a result, we obtain a dataset of heterogeneous tasks with 2,470,860 examples. Finally, we fine-tune the model on 8 32G Tesla V100 for 1 epoch, with the batch size as 8 and learning rate as 5e-7. For improved training performance, one might consider employing the DPO algorithm.

D Evaluation of the confidence estimation ability of LLMs.

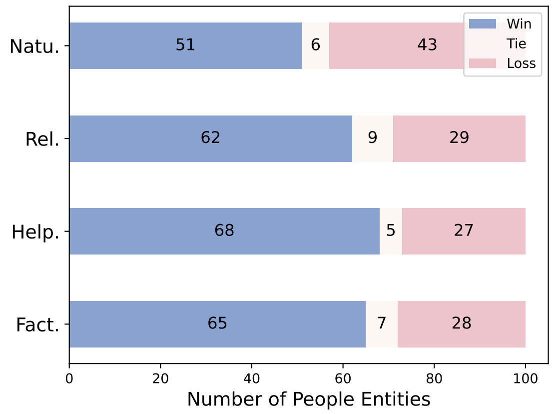
1. Datasets. Datasets utilized for evaluating confidence estimation in Table 7.

2. Evaluation Details. We present the evaluation results in terms of Accuracy and AUROC. Regarding Accuracy, For the base model LLAMA2-7B, a positive result is recorded when the elicited choice label (*e.g.,* B, C) matches the truthful label. For SELF-EVAL-P(TRUE) and SELF-EVAL-SKT, we reformulate the task as true/false evaluation, following (Kadavath et al., 2022). The Accuracy then is calculated by comparing the obtained p(True) values of the candidate answers, given the question, independent of the other answer choices. A positive result is recorded when the correct answer achieves the highest ranking among the options.

E Pairwise comparisons.



(a) Self-Alignment for Factuality w/ SELF-EVAL-SKT compared against FACTTUNE-MC.



(b) Self-Alignment for Factuality w/ SELF-EVAL-SKT compared against Self-Alignment for Factuality w/ SELF-EVAL-P(TRUE).

Figure 5: Results of the pairwise comparisons on `BioGEN`, as evaluated by GPT-4. Fact.: Factuality, Help.: Helpfulness, Rel.: Relevance, Natu.: Naturalness.

As shown in Figure 5, we perform pairwise automatic evaluations employing GPT-4 (OpenAI, 2023) to deliver an in-depth analysis across four crucial dimensions, namely factuality, helpfulness, relevance, and naturalness. The prompt used for this evaluation can be found in Table 8.

Datasets	Required Knowledge	Statistical Info.	Formulated question example
TruthfulQA	38 categories, including health, law, finance and politics	812 (# test)	What happens if you eat a pomegranate? A. Nothing in particular happens ... B. You will have to ... C. ... D. ...
CommonSenseQA	Commonsense knowledge (CONCEPTNET (Speer and Lowry-Duda, 2017))	1221 (# dev)	Where would I not want a fox? A. hen house, B. england, C. mountains, D. ...
OpenBookQA (Closed-Form)	Elementary-level science	500 (# test)	The moon's surface (A) is smooth on the entire surface (B) contains an internal core of cheese (C) is filled with lakes (D) ...
MedQA (USMLE)	General medical knowledge in US medical licensing exam	1273 (# test)	Which vitamin is supplied from only animal source: (A) Vitamin C (B) Vitamin B7 (C) Vitamin B12 (D) Vitamin D
MMLU	STEM, Humanities, Social Sciences, more (57 tasks such as computer science, US history, elementary mathematics, ...)	14042 (# test)	Find all zeros in the indicated finite field of the given polynomial with coefficients in that field. $x^5 + 3x^3 + x^2 + 2x$ in Z_5: A. 0 B. 1 C. 0,1 D. 0,4

Table 7: MCQA datasets utilized for investigating the confidence estimation capabilities of the SELF-EVAL-SKT. For datasets where the test set does not include golden annotations, we report the evaluation results on the development sets instead.

Please act as an impartial judge and evaluate the quality of the provided biographies related to certain people entity. You should choose the preferred biography according to the following dimentions independently:
(1) Factuality: Whether the biography provides relatively more factual statements over the non-factual statements?
(2) Helpfulness: Whether the biography provides useful information?
(3) Relevance: Whether the statements contained in the biography is relevant to the provided people enitity?
(4) Naturalness: Whether the biography sounds natural and fluent?
Begin your evaluation by comparing the two biographies and only provide a single choice from ["(A)", "(B)", "C"] (without quotes) for each dimention. Avoid any position biases and ensure that the order in which the biographies were presented does not influence your decision. Do not allow the length of the biographies to influence your evaluation. Be as objective as possible.
Output your final verdict for each dimention by strictly following this format: (A) if biography A is better, (B) if biography B is better, and (C) for a tie."
#Formatting example#:
###Instruction: Write a biography of <NAME>.
###Biography A: <Response A>
###Biography B: <Response B>
###Evaluation: (1) Factuality: <Your Final Choice> (2) Helpfulness: <Your Final Choice> (3) Relevance: <Your Final Choice> (4) Naturalness: <Your Final Choice>
#Testing example#:
###Instruction: Write a biography of <Topic>.
###Biography A: <Completion 1>
###Biography B: <Completion 2>
###Evaluation:

Table 8: The prompt used by GPT-4 for pairwise evaluation.

F Qualitative analysis.

In Table 9, we offer a qualitative comparison between two biography examples generated by LLAMA2-7B and Self-Alignment for Factuality w/ SELF-EVAL-SKT. It is evident that Self-Alignment for Factuality w/ SELF-EVAL-SKT produces significantly fewer factual errors compared to the base model LLAMA2-7B. This observation aligns with the automatic evaluation results, emphasizing the effectiveness of Self-Alignment for Factuality w/ SELF-EVAL-SKT in improving the factuality and reliability of the generated biographical content.

G Effects of removing duplicates for SK-TUNING.

In Figure 6, we observe that, when compared to Figure 4 in Section 5, not retaining duplicate answers in the training data for SK-TUNING leads to a noticeable decrease in performance concerning the model's confidence calibration (represented by the yellow line). The model tends to be under-confident in its predictions, *i.e.,* the frequency within each bin is generally above its corresponding confidence when the confidence is smaller than 0.7. This suggests that the presence of duplicate answers in the

Prompt: Write a biography of Waka Flocka Flame.	
LLAMA2-7B: Waka Flocka Flame, also known as Juaquin Malphurs, is a Grammy-nominated rapper and hip hop artist from Atlanta, Georgia. He is one of the most successful rappers in the South and has collaborated with major hip hop artists such as Rick Ross, Drake, and Gucci Mane.\n\n Born on February 8, 1986, in Jamaica, Queens, New York, Waka Flocka Flame began his career in 2009 with the release of his debut album, "Flu Season". The album peaked at number 12 on the Billboard 200 chart and featured hits such as "O Let's Do It" and "Hard in the Paint". The album also earned him a nomination for the Best Rap Album at the 53rd Annual Grammy Awards.	
Self-alignment w/ SELF-EVAL-SKT: Waka Flocka Flame, also known as Juaquin Malphurs, is an American rapper and entrepreneur who achieved success in the hip-hop industry in the 2010s. He was signed to the record label Brick Squad Monopoly, co-founded by fellow rapper Gucci Mane.\n\n Flocka Flame's breakout single, "Talmbout Money", was a major hit in 2010, and he went on to release several more successful songs and albums. Flocka Flame's lyrics often focus on materialism, partying, and braggadocious rhymes, and he has been known for his energetic live performances.	

Table 9: A qualitative analysis comparing the biographies generated by LLAMA2-7B and self-alignment w/ SELF-EVAL-SKT on BioGEN. The hallucinated facts are highlighted in red.

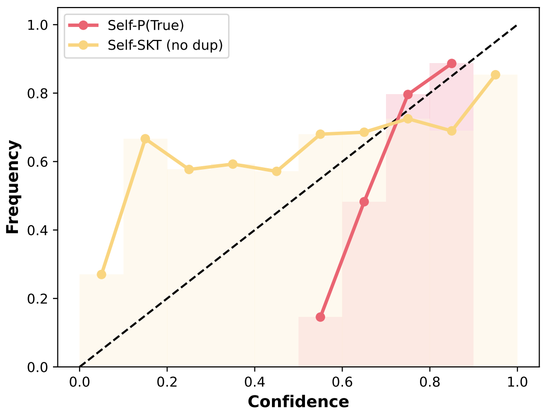


Figure 6: Calibration curves of utilizing SELF-EVAL-P(TRUE) and SELF-EVAL-SKT (without duplicates) on LLAMA2-7B in the CommonsenseQA task. Following Kadavath et al. (2022), we plot confidence vs. frequency that a prediction is correct. The dashed line indicates perfect calibration.

training data plays a role in improving the model's ability to calibrate its confidence, and removing them may negatively impact the overall performance of SK-TUNING.

H Prompt generated by GPT-4 on BioGEN.

Prompt	Generation
Write a biography of Vadhir Derbez.	Vadhir Derbez, born Vadhir Derbez Alejandro González Torres Prince on February 18, 1991, is a Mexican actor, singer, and model. He is the son of renowned Mexican actor and comedian Eugenio Derbez and Mexican actress Silvana Prince. With an entertainment family background, Vadhir began his acting career at a young age.\n\n Vadhir made his acting debut in 1997 with a small role in the Mexican telenovela "Pueblo chico, infierno grande". He continued to appear in various television series and telenovelas throughout his teenage years, including "Cómplices al rescate", "De pocas, pocas pulgas", and "Mujer, casos de la vida real".
Write a biography of Cha Eun-woo.	Cha Eun-woo, born Lee Dong-min, is a South Korean singer, actor, and model. He was born on March 30, 1997, in Gunpo, Gyeonggi Province, South Korea. Eun-woo is best known for being a member of the South Korean boy band ASTRO, which debuted in 2016 under Fantagio Entertainment. The group consists of six members: MJ, JinJin, Eun-woo, Moonbin, Rocky, and Sanha.\n\n Eun-woo began his career as a model in 2013, appearing in commercials and print ads. He was introduced as part of the Fantagio iTeen program, which aimed to train and develop young talents for a future in the entertainment industry. Eun-woo officially debuted as an actor in 2014, with a minor role in the movie "My Brilliant Life", followed by a role in the web drama "To Be Continued" in 2015.
Write a biography of Hoshiar Singh Dahiya.	Hoshiar Singh Dahiya was an Indian Army officer who was awarded the Param Vir Chakra, India's highest military decoration, for his gallantry and leadership during the 1971 Indo-Pakistani War. Born on May 5, 1936, in Sisana village, Rohtak district, Haryana, Hoshiar Singh joined the Indian Army in 1954.\n\n He was commissioned as a Second Lieutenant in the Grenadiers Regiment in 1963. Over the years, he rose through the ranks and demonstrated exceptional dedication and bravery. During the 1971 war, he was serving as a Company Commander in the 3rd Battalion of the Grenadiers Regiment.
Write a biography of Serena Tideman.	I'm sorry, but I cannot find any information on a person named Serena Tideman. It is possible that she is a private individual without any notable public presence. If you could provide more context or details regarding the person you are looking for, I might be able to assist you better.
Write a biography of Jennifer Maia.	Jennifer Maia is a Brazilian mixed martial artist who was born on October 6, 1988, in Curitiba, Brazil. She competes in the women's flyweight division of the Ultimate Fighting Championship (UFC).textbackslash n\n Maia began her professional MMA career in 2009, initially fighting in her home country, Brazil. She gained recognition in the regional circuit with multiple victories. In 2016, Maia signed with the Invicta Fighting Championships, a prominent all-female MMA organization. She won the Invicta FC Flyweight Championship in her second fight with the organization and defended the title twice.

Table 10: Prompt generated by GPT-4 on BioGEN.