Fact-Level Confidence Calibration: Empowering Confidence-Guided LLM Self-Correction

Anonymous ACL submission

Abstract

Confidence calibration in LLMs, i.e., aligning their self-assessed confidence with the actual accuracy of their responses, enabling them to self-evaluate the correctness of their outputs. However, current calibration methods for LLMs typically estimate two scalars to represent overall response confidence and correctness, which is inadequate for long-form generation where the response includes multiple atomic facts and may be partially confident and correct. These methods also overlook the relevance of each fact to the query. To address these challenges, we propose a Fact-Level Calibration framework that operates at a finer granularity, calibrating confidence to relevance-weighted correctness at the fact level. Furthermore, comprehensive analysis under the framework inspired the development of Confidence-guided Fact-level self-correction (ConFact), which uses high-confidence facts within a response as additional knowledge to improve low-confidence ones. Extensive experiments across four datasets and six models demonstrate that ConFact effectively mitigates hallucinations without requiring external knowledge sources such as retrieval systems[1].

1 Introduction

Large Language Models (LLMs) have recently achieved notable breakthroughs in various tasks (Brown et al., 2020), demonstrating their ability to comprehend and generate language that bears a striking resemblance to human communication (OpenAI, 2023). Nonetheless, a major obstacle to their reliability is the prevalence of hallucinations (Lin et al., 2021; Zhang et al., 2023; Li et al., 2023a; Golovneva et al., 2022; Bang et al., 2023), a phenomenon where the models generate incorrect and unreliable outputs. This issue not only undermines user trust but also restricts the application of LLMs in domains where reliability is crucial, such as in the legal, financial, and educational fields.

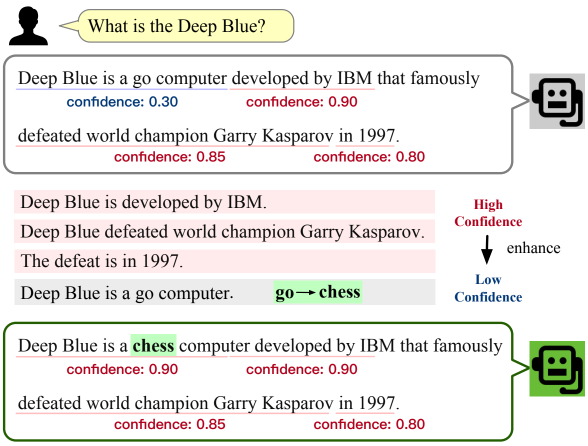

Figure 1: Motivation of our fact-level confidence calibration and confidence-guided self-correction.

Echoing the ancient adage that *"To know what you know and what you do not know, that is true wisdom"*, confidence calibration in LLMs emerges as an effective approach to mitigate the issue of hallucinations (Li et al., 2024; Liu et al., 2023; Huang et al., 2024). By confidence calibrating, models can better align their self-assessed confidence with the actual accuracy of their responses, empowering them to self-evaluate the correctness of their outputs. This mechanism offers an effective way to identify hallucinations by using the model's confidence as a basis for users to either trust or question the model's response.

However, current confidence calibration methods for LLMs (Guo et al., 2017a; Nguyen and O'Connor, 2015) typically estimate two scalars to represent the overall confidence and correctness for the entire response. This approach is unreasonable for long-form generation, where responses may contain multiple atomic facts (illustrated in Fig.1). In such cases, considering the varying of facts in one response, the confidence and correctness

[1]Code is available at https://anonymous.4open.science/r/fact-cal-correct

should also be diverse, capable of reflecting higher certainty in some facts and greater uncertainty in others. Furthermore, within long-form responses, certain facts exist that may indeed be correct but lack relevance to the query. Previous calibration methodologies predominantly focus on assessing correctness while neglecting to incorporate considerations of relevance.

To address these challenges, we propose a novel framework for confidence calibration operating at a finer fact-level granularity. Within this framework, the confidence assessment of each fact incorporates two key aspects: correctness and relevance. Correctness indicates the factual accuracy of the fact, while relevance measures the extent to which the fact is related to the query. Calibration of a response is defined as the degree of alignment between confidence and correctness weighted by relevance across all facts. This framework endows the model with the capability to exhibit partial confidence and correctness in individual facts. Extensive analysis based on the aforementioned framework yields three interesting findings: (1) fact-level calibration imposes a stricter standard than response-level calibration. (2) fact-Level can mitigate over-confidence issues. (3) the variance in confidence distribution among different facts within the same response is considerable.

The aforementioned three observations inspire the development of **Con**fidence-Guided **Fact**-Level Self-Correction (`ConFact`) to enhance the generation and mitigate hallucinations (illustrated in Fig.1). For a response, `ConFact` first leverages the aforementioned framework to segment the response into multiple facts and evaluate their confidence vector. It then uses the high-confidence facts and their associated confidence score as additional knowledge to augment low-confidence facts, with the aim of all facts within the response achieving high confidence. `ConFact` can self-enhance to mitigate hallucinations without the need for external knowledge sources such as retrieval systems. Experiments with `ConFact` across four datasets and six models reveal that it can significantly reduce the occurrence of hallucinations, thereby increasing the models' reliability and enabling their practical application in real-world scenarios.

Our main contributions include:

- **Fact-Level Calibration Framework**: The proposed fact-level calibration framework operates at a finer level of granularity to align the confidence with the correctness weighted by relevance across all facts. This framework endows the model with the capability to exhibit partial confidence and correctness in individual facts.

- **Insightful Observations**: We uncover insightful observations regarding the model's scale and its calibration capability.

- **Self-Correction Method**: We propose `ConFact` method based on the fact-level calibration framework to enhance the generation and reduce hallucinations without relying on external knowledge sources.

2 Preliminary and Problem Formulation

2.1 Preliminary

Consider a dataset defined as $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_i$ denotes the i-th query, with a total count of N queries. Let the model's responses to queries be represented as $A = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$, where each $(\mathbf{x}_i, \mathbf{y}_i)$ forms a query-answer pair. The confidence $conf_i$ signifies the model's degree of certainty in its answer y_i to the query $\mathbf{x}_i$. The correctness $corr_i$ measures the objective truthfulness of the response y_i to the query $\mathbf{x}_i$. The aim of confidence calibration is to ensure that, for every confidence interval, the average confidence of the query-answer pairs within that interval aligns with their average correctness.

2.2 Problem Formulation

Considering the long-form generation nature of LLMs, our proposed confidence calibration is defined at a fact-level granularity, where both the correctness and relevance of each fact will be considered. We define the problem as follows: different from the traditional definition of confidence calibration, we assume the response $\mathbf{y}_i$ contains M_i facts represented as $\{f_i^j\}_{j=1}^{M_i}$. Each fact f_i^j will be evaluated with a relevance value rel_i^j and a correctness value $corr_i^j$. Meanwhile, this fact is also associated with a confidence score $conf_i^j$ representing the LLM's level of uncertainty regarding that fact. The goal of fact-level calibration is to align the confidence with the relevance-weighted correctness in terms of the response $\mathbf{y}_i$ across M_i facts.

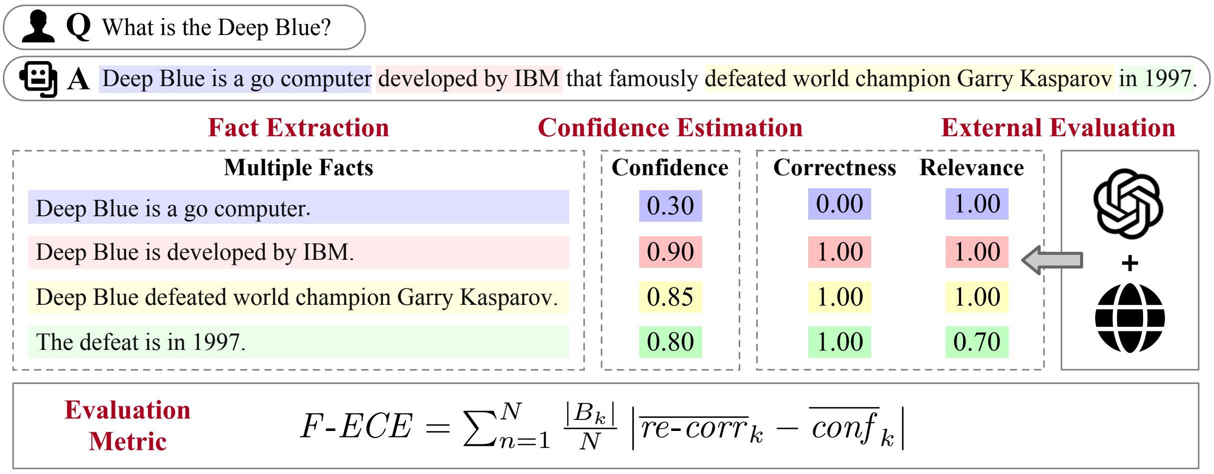

Figure 2: An illustration of our fact-level confidence calibration framework for fine-grained LLM calibration.

3 Fact-Level Confidence Calibration

In this section, we begin by presenting our motivation and offering a detailed introduction to the architecture of Fact-Level Confidence Calibration framework. Subsequently, we delve into three intriguing observations within our framework. Finally, we summarize how these observations inspire our approach to self-correction.

3.1 Motivation

Compared to the confidence calibration for short-form generation or traditional classification problems, a significant challenge in calibrating long-form text generation is that a response may contain multiple facts, making it unreasonable to assign a single correctness measure and a single confidence score to the entire response. The reason is that the answer might be partially correct, and the model might also be partially confident in only a subset of the facts of a response. Meanwhile, some facts in response are irrelevant to the query, so the calibration based solely on correctness is insufficient.

Based on the above motivation, our proposed calibration framework aims to calibrate the confidence to relevance-weighted correctness on the fact level, which leads to the following two advantages: (1) **Finer Granularity**: we assign a confidence vector rather than a scalar to a response, where each item represents confidence for a single fact. This fine-grained framework allows for more nuanced and precise calibration. (2) **Relevance Awareness**: we assess both the correctness and the relevance of each fact, which ensures that the confidence score attributed to each fact can reflect its significance and appropriateness within the given context.

3.2 Architecture

To calibrate the confidence with the relevance-weighted correctness on the fact level, our framework includes four components as illustrated in fig. 2: fact extraction, correctness and relevance evaluation, confidence estimation, and evaluation based on fact-level calibration metric.

Fact Extraction Given a query-answer pair $(\mathbf{x}_i, \mathbf{y}_i)$ from a model to be calibrated, we first dissect the response to identify the contained facts. This process can be performed by a powerful external language model (e.g., GPT-4 (Brown et al., 2020)), resulting in a set of facts $\{f_i^j\}_{j=1}^{M_i}$ for the response $\mathbf{y}_i$.

Correctness and Relevance Evaluation After extracting facts, this component aims to assess the correctness and relevance of each fact to the query. The correctness of each fact is evaluated for its factuality using GPT models in conjunction with retrieval methods based on search engines and the ground truth answers in datasets to obtain $\{corr_i^j\}_{j=1}^{M_i}$. The relevance $\{rel_i^j\}_{j=1}^{M_i}$ of each fact is also obtained based on GPT models, representing its pertinence to the query within the context of the response.

Confidence Estimation The confidence estimation measures the confidence of the targeted LLM for each fact, considering both its correctness and relevance. To obtain a confidence vector $\{conf_i^j\}_{j=1}^{M_i}$ for each response, a verbalization-based method (Tian et al., 2023) is employed, where the model is prompted to provide a confidence score for each fact within response. Confidence of a fact $fact_i^j$ can be represented as:

$$conf_i^j = \mathcal{C}(LLM(\cdot), p_c(f_i^j, \mathbf{x}_i, \mathbf{y}_i)), \quad (1)$$

where p_c is the prompt, which includes: (1) A clear task description. (2) The criteria to give confidence scores. (3) Several instances containing the input query, the complete response, one extracted fact, the associated confidence score with an explanation. (4) The task containing the input query, the complete response, and the target fact. The model is expected to output its confidence for the target fact in a verbalization manner, accompanied by an explanation. For a detailed prompt template, please refer to Appendix C.

Evaluation based on Fact-level Calibration Metric We define F-ECE (Fact-Level Expected Calibration Error) as the evaluation metric that quantifies the discrepancy between confidence and the relevance-weighted correctness across all responses and their respective facts. For each fact within a response, we compute the relevance-weighted correctness as the product of the fact's correctness score and its relevance score. Response-level relevance-weighted correctness and confidence are then determined by averaging these relevance-weighted correctness scores and the confidence scores across all facts within each response, as shown in eq. (2).

$$\begin{aligned} re\text{-}corr_i &= \frac{1}{M}\sum_{j=1}^{M} corr_i^j \times rel_i^j, \\ conf_i &= \frac{1}{M}\sum_{j=1}^{M} conf_i^j, \end{aligned} \quad (2)$$

where $recorr$ denotes the relevance-weighted correctness and $conf$ denotes the confidence.

F-ECE is finally calculated by the average relevance-weighted correctness and confidence of responses in bin k, where B is the number of bins for grouping confidence scores, and B_k is the set of responses in the k-th bin. Let $\overline{re\text{-}corr}_k = \frac{1}{|B_k|}\sum_{i \in B_k} corr_i$ and $\overline{conf}_k = \frac{1}{|B_k|}\sum_{i \in B_k} conf_i$,

$$F\text{-}ECE = \sum_{n=1}^{N} \frac{|B_k|}{N} \left| \overline{re\text{-}corr}_k - \overline{conf}_k \right| \quad (3)$$

3.3 Key Observations

This section discusses three important phenomena observed under our fact-level calibration.These findings not only demonstrate the superiority of our framework over traditional response-level calibration, but also inspire the development of a confidence-guided fact-level self-correction method based on these insights.

Observation 1: Fact-level calibration imposes a stricter standard than response-level calibration. As illustrated in fig. 3, by comparing the histogram between the left side and right side, it is evident that our fact-level framework can accentuate the differences in calibration performance across different scale models with various capabilities. Specifically, the models (e.g., Llama-2-7b) that appear well-calibrated under traditional response-level perform worse in fact-level calibration. This capability stems from fact-level calibration, which takes into account the fine-grained correctness at the fact level and considers the relevance of each fact to the query, highlighting the importance of utilizing a more granular calibration assessment to uncover hidden deficiencies in model performance.

Observation 2: Fact-Level Can Mitigate Over-Confidence Issue The distribution of confidence across datasets is illustrated in fig. 4. The response-level calibration assigns a single confidence value to the entire response, shown in gray. In contrast, our fact-level method assigns a confidence value to each fact within the response, resulting in a confidence vector for one response. We calculate the mean, minimum, and maximum values of each confidence vector, and depict the statistical distribu-

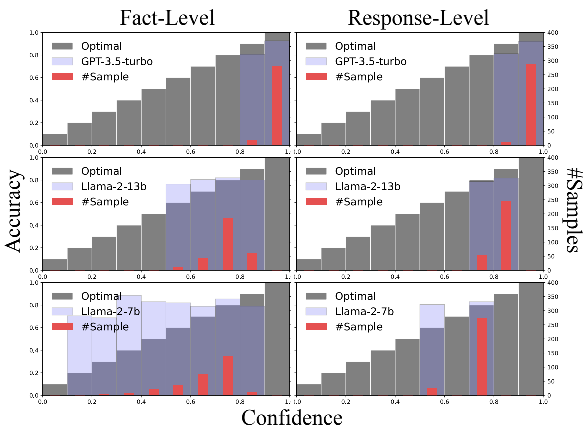

Figure 3: Comparison of calibration measures between fact-level and response-level based on models with three different scales: Llama-7B, Llama-13B, and GPT-3.5.

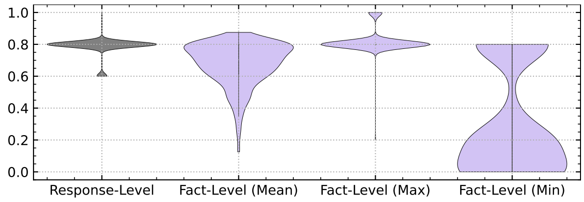

Figure 4: Comparison of across-responses confidence distribution between fact-level and response-level.

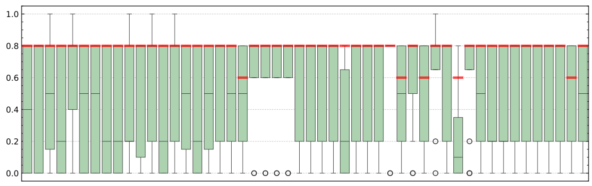

Figure 5: Confidence distribution within responses at the fact level, the red bar is the response-level score.

tions of these values across all responses via violin plot (Hintze and Nelson, 1998). Two intriguing phenomena can be observed: (1) The confidence distribution of the response-level is narrow and centered around a high confidence value. Our distribution of mean confidence values, on the other hand, is wider and shows a lower response level. (2) The distribution of the response-level is highly similar to the distribution of the maximum confidence values in our fact-level method.

These two phenomena suggest that response-level confidence is dominated by the fact in the response with implicitly highest confidence, which can lead to over-confidence. Our framework, by breaking down the facts and evaluating confidence individually, can explicitly emphasize less confident aspects within the response, thereby mitigating the over-confidence issue.

Observation 3: High Variance exists in Fact-level Confidence within a Response fig. 5 illustrates the distribution of confidence levels for facts within specific responses, depicted by the green box plots. The red dots represent the response-level confidence for the entire response. Two phenomena can be observed: (1) Fact-level confidence varies significantly within individual responses, while response-level confidence is relatively concentrated at a higher level. (2) Outlier facts tend to exhibit lower confidence levels. The numerous white dots in the box plots indicate the presence of these outliers, which typically correspond to facts with significantly lower confidence scores, generally falling below the overall distribution. This suggests that certain facts within a response are generated with considerably less confidence by the model.

4 `ConFact`: Confidence-Guided Fact-Level Self-Correction

In this section, we introduce the motivation and architecture of **Con**fidence-guided **Fact**-level self-correction, dubbed **`ConFact`**. `ConFact` utilizes facts with high confidence as references to revise facts with low confidence, thereby enhancing the generation process and mitigating hallucinations. `ConFact` operates in real-time during the generation process, avoiding the need for fine-tuning or training, thereby lowering costs and increasing flexibility. Moreover, it does not rely on external knowledge, significantly enhancing its universality.

4.1 Motivation

The development of Confidence-Guided LLM Self-Correction is inspired by the aforementioned three observations. The rationale behind these observations supporting Self-Correction lies in: (1) Our observations 1 and 2 show that even under strict conditions, the fact-level framework can reduce over-confidence and improve the model's calibration, aligning confidence more closely with accuracy. This improved calibration is essential for effective confidence-guided self-correction. (2) Our observation 3 shows that high-confidence and low-confidence facts often coexist within the same response. Even when confidence levels are generally consistent, outliers tend to be lower confidence facts. This allows high-confidence facts to provide the necessary knowledge to correct the low-confidence ones.

4.2 Architecture

The overall architecture of `ConFact` is illustrated in fig. 6. As can be seen, `ConFact` includes three steps: fact extraction and confidence estimation, factor extraction and fact correction, and fact confidence re-estimation.

Step 1: Fact Extraction and Confidence Estimation Given a response $\mathbf{y}_i$, `ConFact` first conducts fact extraction and confidence estimation for each extracted fact, following the same process as described in section 3.2. After obtaining the facts $\{f_i^j\}_{j=1}^{M_i}$ for $\mathbf{y}_i$ and their corresponding confidence scores $\{conf_i^j\}_{j=1}^{M_i}$, we then split the facts into two groups: high-confidence and low-confidence, based on a confidence threshold τ.

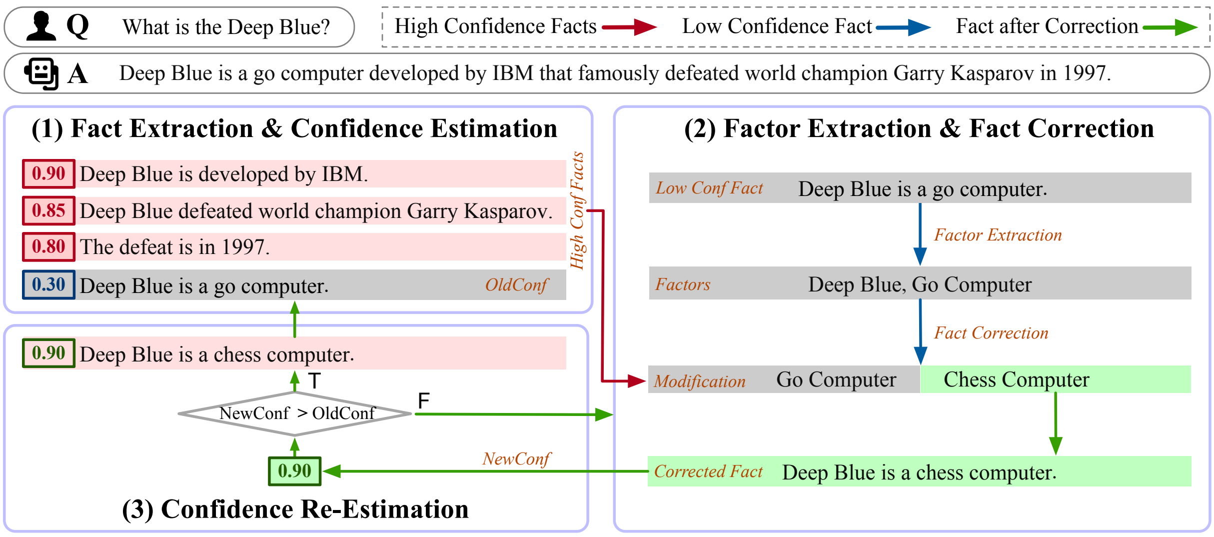

Figure 6: An illustration of our confidence-guided fact-level self-correction framework.

The high-confidence group in eq. (4) is used as a form of internal knowledge base, whose knowledge is leveraged to reinforce and augment facts within the low-confidence group in eq. (5),

$$f_h = \{f_i^j \mid conf_i^j \geq \tau\} \tag{4}$$

$$f_l = \{f_i^j \mid conf_i^j < \tau\}, \tag{5}$$

where the threshold is defined as the mean confidence score across facts $\tau = \frac{1}{M_i}\sum_{j=1}^{M_i} conf_i^j$.

Step 2: Factor Extraction and Fact Correction To ensure that only the erroneous parts of the low-confidence facts are modified without changing the overall meaning, we restrict the modifiable parts. Specifically, we first parse the key factors through factor extraction. Let $\{fa_i^{j,k}\}_{k=1}^{K_i^j}$ represent the K_i^j factors extracted from the target fact $f_i^j \in f_l$,

$$\{fa_i^{j,k}\}_{k=1}^{K_i^j} = \mathcal{F}(LLM(\cdot), p_f(f_i^j)), \tag{6}$$

where p_f is the prompt, which includes: (1) A clear task description. (2) Several instances. (3) The task containing the input sentence. The model is expected to output its extracted factors.

After extracting factors, we then perform fact correction, targeting only the extracted factors for modification. This process can be represented as,

$$\hat{f}_i^j = \mathcal{R}(LLM(\cdot), p_r(f_i^j, \{fa_i^{j,k}\}_{k=1}^{K_i^j}, f_h) \tag{7}$$

where p_r is the prompt, which includes: (1) A clear task description. (2) Several instances. (3) The task containing the input target fact, the extracted factors and the high-confidence reference facts. The model is expected to output the modified target fact, noting that the model allows for returning "NoError" to make no modifications to the input.

Step 3: Fact Confidence Re-Estimation Finally, the modified facts undergo the confidence estimation process again to obtain new confidence scores:

$$\hat{conf}_i^j = \mathcal{C}(LLM(\cdot), p_c(\hat{f}_i^j, \mathbf{x}_i, \mathbf{y}_i)), \tag{8}$$

where $\hat{conf}_i^j$ represents the confidence score of the modified fact $\hat{f}_i^j$. Finally, if $\hat{conf}_i^j > conf_i^j$, the modification is deemed successful and is accepted. Otherwise, ConFact will repeat the process of factor extraction, fact correction, and confidence re-estimation. This iterative process continues until either a satisfactory confidence score is achieved or a predetermined maximum number of iterations N is reached, where the model return "NoError" and make no modifications to the input.

5 Experiment

5.1 Experiment Setup

This section outlines the experimental setups, including the datasets, models, and evaluation.

Datasets We employ two datasets: (1) Long-Fact (Wei et al., 2024): A dataset consisting of prompts designed to assess a model's factuality in long-form responses created by GPT-4. (2) ASQA (Stelmakh et al., 2022): A dataset designed for long-form question answering that uniquely centers on ambiguous factoid questions.

Models We use five models from different families and scales to validate our method, including: (1) Llama (Touvron et al., 2023): include Llama-7b-chat and Llama-13b-chat. (2) Vicuna (Chiang et al., 2023): include Vicuna-7b and Vicuna-13b. (3) GPT (Brown et al., 2020): GPT-3.5-turbo.

Correctness and Relevance Evaluation For correctness and relevance evaluation, we use the Search-Augmented Factuality Evaluator (SAFE), which is a pipeline proposed by (Wei et al., 2024) that employs LLMs as agents to automatically evaluate the factuality of long-form responses. It utilizes a multi-step reasoning process that includes sending search queries to Google Search (Hillis et al., 2012) to verify the information provided. For the fact correction evaluation, we use GPT-4 for zero-shot pair-wise evaluation (see prompts in Appendix C).

Evaluation Metrics For calibration evaluation, we use Expected Calibration Error (ECE) (Guo et al., 2017a; Naeini et al., 2015) at the response-level, and our F-ECE at the fact-level as introduced in section 3.2. For self-correction evaluation, the evaluation metrics are twofold. Firstly, we use Accuracy, Precision, and Recall (Powers, 2020) to evaluate error detection. Then, we use improvement ratio, same ratio, and regression ratio to evaluate self-correction.

5.2 Results for Fact-Level Calibration

This section provides detailed implementation and comprehensive experiment results of our fact-level calibration framework. As introduced in section 3.3, we have three key observations.

Table 1: Comparison of response-level and our fact-level calibration performance of five base models in terms of (F-)ECE under ASQA and LongFact datasets.

Base Model	Method	ASQA	LongFact
Llama-2-7b	Fact	0.261	0.211
	Response	0.251	0.141
Llama-2-13b	Fact	0.240	0.156
	Response	0.261	0.131
Vicuna-7b	Fact	0.337	0.151
	Response	0.352	0.137
Vicuna-13b	Fact	0.254	0.113
	Response	0.269	0.109
GPT-3.5-turbo	Fact	0.179	0.086
	Response	0.185	0.094

Calibration Comparison for Observation 1 For Observation 1, we compare our fact-level and response-level calibration in accordance with the protocol in (Guo et al., 2017a). We illustrate reliability histograms and compute the summary statistics of ECE and our F-ECE to evaluate calibration. The procedures are implemented as follows: For fact-level, we evaluate confidence, correctness, and relevance as described in section 3.2. For response-level, we use a verbalization-based method following the procedure in (Huang et al., 2024), where the model is prompted to provide a single confidence score for the whole response. For a detailed prompt template, please refer to Appendix C. For the reliability histogram, we divided the model's predictions into ten bins based on the confidence score and calculated the average accuracy for each bin. From the perspective of the histogram, an optimally calibrated model should have its bar graph in a diagonal shape to achieve the smallest gap area. The results are depicted in fig. 3 and table 1.

Across-Responses Confidence Distribution for Observation 2 For Observation 2, we examine how our fact-level calibration can mitigate the over-confidence issue by analyzing the distribution of confidence scores. The procedures are implemented as follows: For the response-level, we use a verbalization-based method to obtain a score for each response and visualize its distribution across the entire dataset using violin plots. For the fact-level, since the confidence for a single response is represented as a vector rather than a scalar, we compute three different statistical measures: the mean, maximum, and minimum of the vector. We then visualize these measures as three separate violin plots. The results are depicted in fig. 7.

Within-Responses Confidence Distribution for Observation 3 For Observation 3, we investigate the variance in fact-level confidence within individual responses. The procedures are implemented as follows: For each response, we obtain its confidence vector and visualize its distribution using box plots. Due to space limitations, we have visualized 10 responses for each model in each dataset in fig. 8, whereas this number is 50 in fig. 5. The red bar is the confidence score of the whole response at response-level.

5.3 Results for Self-Correction

Error Detection table 2 presents the error detection results of our proposed method based on five

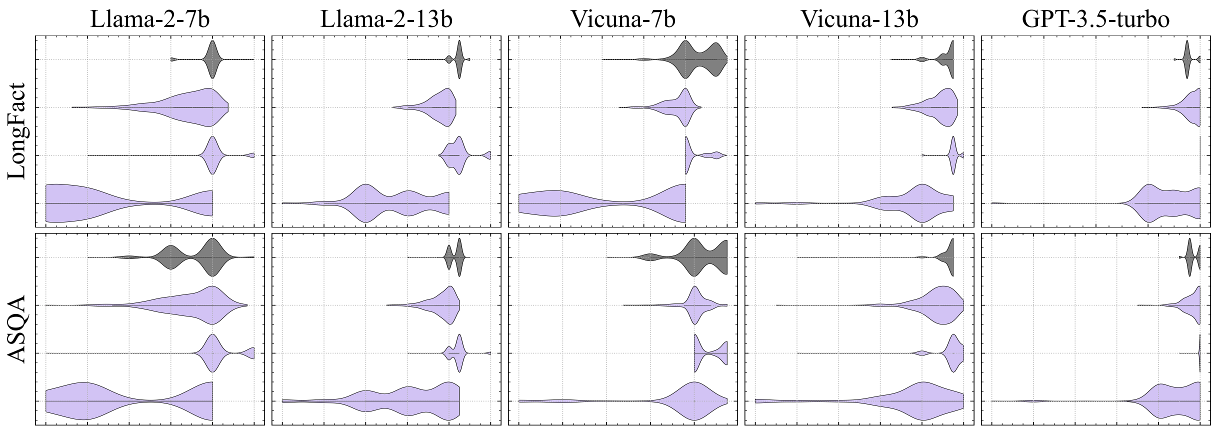

Figure 7: Comparison of confidence distribution across different responses between fact-level and response-level. The purple are our fact-level distribution under different statistical metrics, the gray is the response-level distribution.

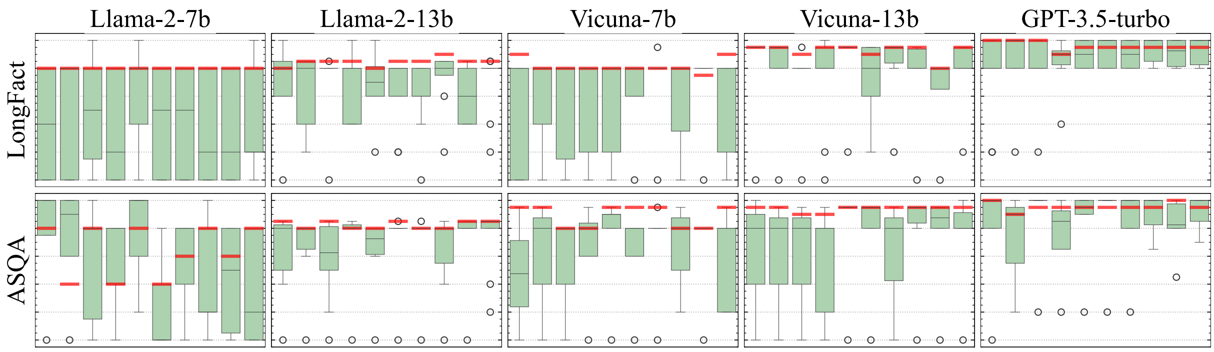

Figure 8: Confidence distribution within responses at the fact level, the red bar is the response-level score.

different base models. It can be seen that, in terms of Accuracy and Precision, larger models perform better than smaller models, i.e., GPT > 13b > 7b. However, all models somewhat fall short in Recall, indicating that many erroneous facts are not being detected. This suggests that all models exhibit a certain degree of overconfidence, often considering incorrect answers to be correct.

Error Correction table 3 presents the error correction results of our proposed method based on five different base models. It can be seen that, among the three outcomes "improve," "same," and "regress," our method achieves the highest proportion of "improve" for all models except LLaMA. This indicates that our method effectively enables the models to self-correct and achieve better generation results.

6 Conclusion

This paper introduces a novel fact-level calibration framework to address hallucination issues in long-form responses generated by LLMs. Traditional single-estimate confidence methods are inadequate for complex outputs with multiple facts. By evaluating each fact's correctness and relevance individually, both externally and internally, our framework enables fine-grained confidence assessments. It sets a higher standard than response-level approaches, mitigates over-confidence, and reveals significant confidence variance among facts within responses. Leveraging high-confidence facts for in-context learning effectively mitigates hallucination, as validated across multiple datasets and models.

Table 2: Acc., Precision and Recall of error-detection.

Base Model	Accuracy (%)	Precision (%)	Recall (%)
GPT-3.5-turbo	83.29	87.89	13.71
Vicuna-7b	60.06	99.90	0.15
Vicuna-13b	74.81	77.68	8.46
Llama-2-7b	64.26	67.86	13.35
Llama-2-13b	70.45	77.45	30.62

Table 3: GPT-4 evaluation of the self-correction.

Base Model	Improved (%)	Same (%)	regressed (%)	#revised
GPT-3.5-turbo	46.30	24.07	29.63	108
Vicuna-7b	50.00	50.00	0.00	2
Vicuna-13b	49.40	28.92	21.69	83
Llama-2-7b	6.76	12.56	80.68	207
Llama-2-13b	53.35	19.59	27.07	418

7 Limitations and Broader Impacts

In this work, we propose a fact-level calibration framework and, based on this framework, introduce a confidence-guided fact-level self-correction method. However, for this self-correction method to be effective, the model itself must possess a certain level of calibration ability. In our paper, we discuss how our calibration framework can alleviate over-confidence. In future work, we will further explore ways to enhance calibration ability within the calibration framework, paving the way for more effective confidence-guided self-correction.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Silei Cheng, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable. In International Conference on Learning Representations (ICLR 23).

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Soham Dan and Dan Roth. 2021. On the effects of transformer size on in-and out-of-domain calibration. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2096–2101.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. arXiv preprint arXiv:2003.07892.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. arXiv preprint arXiv:2404.04475.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. Preprint, arXiv:2305.14387.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2022. Roscoe: A suite of metrics for scoring step-by-step reasoning. In The Eleventh International Conference on Learning Representations.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017a. On calibration of modern neural networks. In International conference on machine learning, pages 1321–1330. PMLR.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017b. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of Proceedings of Machine Learning Research, pages 1321–1330. PMLR.

Ken Hillis, Michael Petit, and Kylie Jarrett. 2012. Google and the Culture of Search. Routledge.

Jerry L Hintze and Ray D Nelson. 1998. Violin plots: a box plot-density trace synergism. The American Statistician, 52(2):181–184.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications. arXiv preprint arXiv:2306.04459.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024. Calibrating long-form generations from large language models. arXiv preprint arXiv:2402.06544.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.

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.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. In Proceedings of the

2023 Conference on Empirical Methods in Natural Language Processing, pages 6449–6464.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024. Think twice before assure: Confidence estimation for large language models through reflection on multiple answers. arXiv preprint arXiv:2403.09972.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958.

Xin Liu, Muhammad Khalifa, and Lu Wang. 2023. Lightweight language model calibration for open-ended question answering with varied answer lengths. In The Twelfth International Conference on Learning Representations.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence, 2015:2901–2907.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, pages 1587–1598. The Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. CoRR, abs/2303.08774.

David MW Powers. 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. arXiv preprint arXiv:2204.06092.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. arXiv preprint arXiv:2305.14975.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. CoRR, abs/2307.09288.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, et al. 2024. Long-form factuality in large language models. arXiv preprint arXiv:2403.18802.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.

A Datasets

LongFact (Wei et al., 2024): A dataset consisting of prompts designed to assess a model's factuality in long-form responses created by GPT-4. This dataset includes a diverse range of topics and ensures that the prompts require detailed and nuanced answers, making it a robust benchmark for evaluating the factual accuracy of language models in generating extended text. The dataset is particularly valuable for testing the capabilities of models in maintaining factual consistency over longer passages, which is crucial for applications such as content creation, summarization, and complex question answering.

ASQA (Stelmakh et al., 2022): A dataset designed for long-form question answering that uniquely centers on ambiguous factoid questions. ASQA provides a challenging testbed for models as it includes questions that can have multiple valid answers depending on the interpretation of the ambiguity. This dataset emphasizes the need for models to not only retrieve accurate information but also to handle the inherent uncertainty and provide comprehensive responses. ASQA is instrumental in pushing the boundaries of model performance in scenarios where clarity and precision are essential, such as in education and interactive AI systems.

B Models

Llama-7b-chat & Llama-13b-chat (Touvron et al., 2023): These models are part of the LLaMA family, known for their strong performance in various natural language processing tasks. The "chat" versions are particularly fine-tuned for conversational contexts, making them suitable for generating coherent and contextually appropriate responses in dialogue settings. LLaMA models are designed to balance performance and computational efficiency, making it a popular choice for research and application in interactive AI systems.

Vicuna-7b and Vicuna-13b (Chiang et al., 2023) Vicuna is an open-source chatbot trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT. Preliminary evaluation using GPT-4 as a judge shows Vicuna-13B achieves more than 90% quality of ChatGPT and Bard while outperforming other models like LLaMA and Stanford Alpaca (Li et al., 2023b; Dubois et al., 2024, 2023) in more than 90% of cases.

GPT-3.5-turbo (Brown et al., 2020) This model is part of OpenAI's well-known GPT series. GPT-3.5-turbo is an enhanced version of GPT-3, offering improved performance and efficiency. It is designed to handle a wide range of language tasks, from text generation to comprehension and translation. The "turbo" variant is optimized for faster inference and lower latency, making it ideal for real-time applications where quick response times are crucial. GPT-3.5-turbo is widely used in both research and industry due to its versatility and high-quality output.

C Prompts

C.1 Prompt for Fact-Level Confidence Estimation

The specific prompts used for fact-level confidence estimation are detailed below.

```
Instructions:
1. The following STATEMENT has been extracted
from the broader context of the given RESPONSE to
the given QUESTION.
2. Indicate how confident you are in the accuracy
of the STATEMENT when answering the QUESTION,
based on your knowledge.
3. The confidence evaluation should be a value
between 0 and 1 (with two decimal places retained),
based on the following scoring criterion:
{Criterion} 4. Your task is to do this for the
STATEMENT, RESPONSE and QUESTION under "Your
Task".
Some examples have been provided for you to learn
how to do this task.


{Some Examples}

Your Task:
QUESTION:
{Question}

RESPONSE:
{Response}

STATEMENT:
{Statement}
```

Table 4: Prompt for fact-level confidence estimation {Criterion}, {Question}, {Response} and {Statement} are placeholders.

C.2 Prompt for Response-Level Confidence Estimation

The specific prompts used for response-level confidence estimation are detailed below.

```
Instructions:
1. The following RESPONSE is the answer to the given QUESTION.
2. Indicate how confident you are in the accuracy of the RESPONSE when answering the QUESTION, based on your knowledge.
3. The confidence evaluation should be a value between 0 and 1 (with two decimal places retained), based on the following scoring criterion:
{Criterion}
4. Your task is to do this for the RESPONSE and QUESTION under "Your Task".
Some examples have been provided for you to learn how to do this task.


{Some Examples}

Your Task:
QUESTION:
{Question}

RESPONSE:
{Response}
```

Table 5: Prompt for response-level confidence estimation {Criterion}, {Question}, {Response} and {Statement} are placeholders.

C.3 Prompt for Factor Extraction

The specific prompts used for factor extraction are detailed below.

```
Instructions:
You are to read a sentence and identify the key factors within it.
The task involves pinpointing the essential elements or aspects that significantly influence or characterize the situation, event, or subject described.
Return the identified key factors using the format <[factor1, factor2, ...]>
Some examples have been provided for you to learn how to do this task.


{Some Examples}

Your Task:
SENTENCE:
{Sentence}
```

Table 6: Prompt for factor extraction {Sentence} is placeholders.

C.4 Prompt for Fact Correction

The specific prompts used for fact correction are detailed below.

C.5 GPT-4 Judgments for Self-Correction

For the self-correction, we utilize GPT-4 for zero-shot pair-wise evaluation. We use `gpt-4-0314` for all our experiments. The specific prompts used for GPT-4 evaluation are detailed below.

```
Instructions:
You have been provided with a sentence and some reference knowledge.
The sentence has been analyzed, and its factors have been identified.
However, it is acknowledged that there may be errors or inaccuracies in the identified factors.
Your task is to first review the identified factors and check for any errors or inaccuracies.
If there are no errors, simply return "NoError" to indicate that no corrections are needed.
If errors are present, proceed to make the necessary corrections.
Ensure that the corrections are limited to the existing factors without adding new content.
Use the format <old factor -> new factor> for each correction.


{Some Examples}

Your Task:
SENTENCE:
{Sentence}

FACTORS:
{Factor}

REFERENCE:
{Reference}
```

Table 7: Prompt for fact correction {Sentence}, {Factor} and {Reference} are placeholders.

D Related Work

The concept of confidence calibration was first introduced to nerual networks by (Guo et al., 2017a) to prevent logits from making incorrect classifications with high probability. This concept has since been extended to NLP models (Desai and Durrett, 2020; Dan and Roth, 2021; Hu et al., 2023). Common methods for estimating confidence scores include logit-based methods, consistency-based methods, and verbalization-based methods. Logit-based methods (Guo et al., 2017b; Cheng et al., 2023; Kadavath et al., 2022) assess model confidence by examining the logits predicted by the model. Consistency-based methods (Wang et al., 2023; Kuhn et al., 2023) rely on the principle that language models tend to produce similar outputs consistently when they are confident. Recently, research has indicated that verbalization-based methods (Tian et al., 2023) might offer superior confidence estimation.

```
You will be provided with a QUESTION, its RESPONSE, and all facts extracted from the RESPONSE under the heading "ALL FACTS". You will also be provided with a specific fact under the heading "TARGET FACT 1", which is included in ALL FACTS. Additionally, you will be given a modified version of this target fact under the heading "TARGET FACT 2".

Based on your knowledge, evaluate whether the modification of the target fact is an improvement, the same, or a regression.

An improvement implies:
1. More accurate information.
2. Greater relevance to the question.
3. Minimal overlap with other facts in ALL FACTS.

A regression implies: 1. Introduction of erroneous or inaccurate information.
2. Lower relevance to the question.
3. Repetition or introduction of information that is already provided with other facts in ALL FACTS.


QUESTION:
{Question}

RESPONSE:
{Response}

ALL FACTS:
{All Facts}

TARGET FACT 1:
{Original Fact}

TARGET FACT 2:
{New Fact}

First, provide a one-sentence comparison of the two facts and explain whether you think the modification is an improvement, the same, or a regression. Second, on a new line, the state only "IMPROVED", "SAME", or "REGRESSED" to indicate the effectiveness of the modification. Your response should use the following format:
COMPARISON: <one-sentence comparison and explanation>
REVISION: <"IMPROVED", "SAME", or "REGRESSED">
```

Table 8: Prompt for GPT-4 evaluation for the self-correction {Question}, {Response}, {All Facts}, {All Facts}, {Original Facts} and {New Fact} are placeholders.