Atomic Calibration of LLMs in Long-Form Generations

Caiqi Zhang♥†, **Ruihan Yang**♠†, **Zhisong Zhang**♦*,
Xinting Huang♦, **Sen Yang**♣†, **Dong Yu**♦, **Nigel Collier**♥*
♥University of Cambridge, ♠Fudan University, ♦Tencent AI Lab,
♣The Chinese University of Hong Kong
{cz391, nhc30}@cam.ac.uk, zhisonzhang@tencent.com

Abstract

Large language models (LLMs) often suffer from hallucinations, posing significant challenges for real-world applications. Confidence calibration, which estimates the underlying uncertainty of model predictions, is essential to enhance the LLMs' trustworthiness. Existing research on LLM calibration has primarily focused on short-form tasks, providing a single confidence score at the response level (macro calibration). However, this approach is insufficient for long-form generations, where responses often contain more complex statements and may include both accurate and inaccurate information. Therefore, we introduce **atomic calibration**, a novel approach that evaluates factuality calibration at a fine-grained level by breaking down long responses into atomic claims. We classify confidence elicitation methods into **discriminative** and **generative** types and demonstrate that their combination can enhance calibration. Our extensive experiments on various LLMs and datasets show that atomic calibration is well-suited for long-form generation and can also improve macro calibration results. Additionally, atomic calibration reveals insightful patterns in LLM confidence throughout the generation process.

1 Introduction

While large language models (LLMs), such as Llama (Touvron et al., 2023), Mistral (Jiang et al., 2023), and GPT models (OpenAI, 2022), excel in various tasks, they still struggle with faithfulness and reliability issues. LLMs often suffer from hallucinations, generating factually inaccurate content and misleading responses (Zhang et al., 2023b; Huang et al., 2023), which limits their application in high-risk real-world scenarios (Hu et al., 2023). To address this, *confidence calibration* aims to estimate the underlying uncertainty of model predic-
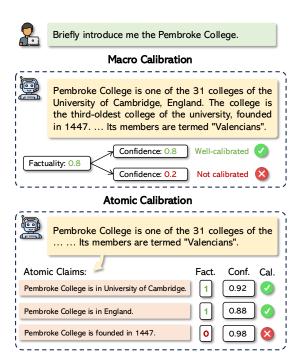

Figure 1: Comparison between traditional macro calibration in response-level and our atomic calibration. The Fact. label is assigned by fact-checking module. We only list three atomic claims for illustration.

tions and reflect the true likelihood of correctness (Guo et al., 2017). A calibrated model is crucial for real-world applications, as it allows us to determine the extent to which we can trust models' predictions (Zhu et al., 2023; Mahaut et al., 2024).

Most existing work on LLM calibration focuses on short-form QA tasks (Jiang et al., 2021; Tian et al., 2023; Zhu et al., 2023; Ulmer et al., 2024). These studies often use datasets like TriviaQA (Joshi et al., 2017) and Natural Questions (Joshi et al., 2017), where answers typically have fewer than 10 words. However, in real-world applications, responses to user queries are often much longer (Zhang et al., 2024), sometimes extending to hundreds or even thousands of words. In such cases, the quality of LLM responses is not simply binary (correct or incorrect), as answers may in-

---

*Corresponding authors.
†Work done during the internship at Tencent AI Lab.

clude both accurate and inaccurate statements. Previous work on long-form calibration has primarily focused on response-level calibration (Zhang et al., 2024; Huang et al., 2024) (which we term *macro calibration*). This approach provides an overall confidence estimation for the entire response (as shown in the upper part of Figure 1). However, it is insufficient for long-form generation, as it cannot adequately capture the model's fine-grained uncertainty across multiple factual statements.

In this work, we introduce the concept of **atomic calibration**, which evaluates calibration at the fine-grained level of atomic claims (as illustrated in the lower part of Figure 1). We focus on the aspect of factuality, since hallucination are a widely recognized issue of LLMs (Zhang et al., 2023b; Huang et al., 2023) and the factuality of statements can be objectively determined. Accurate factual calibration is crucial for mitigating the hallucination problem (Mahaut et al., 2024). To perform atomic calibration, we decompose long responses into self-contained atomic claims and evaluate the calibration for all atomic claims. Various confidence elicitation methods are studied to provide accurate confidence estimation. We provide a detailed investigation into long-form generation calibration, covering 7 LLMs and 3 datasets. The main contributions of our work are as follows:

- We propose the novel concept of **atomic calibration**. Unlike macro calibration, which provides a single response-level confidence score, atomic calibration offers a more fine-grained analysis of model calibration, making it better suited for long-form generation (§2).

- We introduce a categorization of confidence elicitation methods for long-form generation into two types: **discriminative** and **generative**, estimating the model's intrinsic and external confidence, respectively. We examine five confidence elicitation methods and assess their effectiveness for atomic calibration (§3).

- We show that atomic calibration is particularly well-suited for long-form generation, and that macro calibration can be enhanced by incorporating atomic calibration results. We also find the two types of confidence elicitation methods are complementary and their fusion can provide better calibration results. Atomic calibration also enables deeper analysis, revealing patterns in confidence and calibration changes throughout the generation process (§5).

2 Atomic Calibration

For a language model $\mathcal{M}$, let $x \in \mathcal{X}$ represent the response generated by model $\mathcal{M}$ for a query q, and let $y \in \mathcal{Y}_t$ denote the corresponding label, where $\mathcal{Y}_t \subseteq [0, 1]$ indicates a quality score for a specific task $t \in T$. Note that unlike multiple-choice and short-form questions, which typically only focus on the correctness of the answer, the tasks in T involve various aspects, such as factuality, coherence, creativity, etc.

We define a probability prediction function $f : \mathcal{X} \rightarrow \Delta^{|\mathcal{Y}_t|}$, where $\Delta^{|\mathcal{Y}_t|}$ represents the $|\mathcal{Y}_t|$-dimensional simplex. In this context, $f(x)_y$ denotes the probability assigned to the label y given the generated output x. In the remaining of this work, we focus specifically on calibrating for factuality. In this case, $\mathcal{Y}$ represent $\mathcal{Y}_t$ when t corresponds to factuality. $\mathcal{Y}$ takes values in [0, 1], indicating the factuality level of a response. Based on the calibration concepts discussed in (Guo et al., 2017), we define the calibration of each response as follows:

Definition 1 (Macro Calibration on Factuality)
*A language model $\mathcal{M}$ that produces generations $x \sim \mathcal{M}(x \mid q)$ is said to be **response-level (macro) calibrated** if*

$$\mathbb{P}(y \mid f(x)_y = \beta) = \beta, \quad \forall \beta \in \Delta^{|\mathcal{Y}|}.$$

In the context of long-form generation, a single response x may encompass multiple atomic claims. Macro calibration at the response level cannot fully present the fine-grained uncertainty at the atomic level. To address this, we decompose the response x into N atomic claims c_i, represented as $x = \coprod_{i=1}^{N} c_i$. Each atomic claim c_i is assigned a binary label $y_i \in \mathcal{Y}_i$, where $\mathcal{Y}_i = \{0, 1\}$, indicating its truthfulness. The overall factuality score for the response y is computed as $y = \frac{1}{N} \sum_{i=1}^{N} y_i$. Similarly, we define $f(c_i)_{y_i}$ as the probability of the label y_i given the atomic claim c_i. Building on this decomposition, we propose a fine-grained measure of calibration at the atomic level as follows:

Definition 2 (Atomic Calibration on Factuality)
*A language model $\mathcal{M}$, which generates a long-form response x conditioned on the query q, $x \sim \mathcal{M}(x \mid q)$, is considered **atomic-level calibrated** if, for each atomic claim c_i with its corresponding label y_i, the following condition holds:*

$$\mathbb{P}(y_i \mid f(c_i)_{y_i} = \beta_i) = \beta_i, \quad \forall \beta_i \in \Delta^{|\mathcal{Y}_i|}.$$

Remarks: **(1)** Unlike traditional classification problems where $f(x)_y$ is usually represented as a single log probability of the predicted answer, it is much more challenging to measure model confidence in text generation tasks. Different confidence elicitation methods may yield different predictions of the $f(x)_y$; therefore, how to design proper elicitation methods is a key problem. **(2)** Macro calibration is not equivalent to the sum of atomic calibrations, as illustrated by:

$$\mathbb{P}(y \mid f(x)_y = \beta) = \beta$$
$$\not\Rightarrow \frac{1}{N}\sum_{i=1}^{N}\mathbb{P}(y_i \mid f(c_i)_{y_i} = \beta_i) = \beta$$
$$\not\Rightarrow \mathbb{P}(y_i \mid f(c_i)_{y_i} = \beta_i) = \beta, \forall i \in \{1, ..., N\}.$$

3 Confidence Elicitation Methods

To calculate the confidence scores for model outputs in long-form generation, we define two types of confidence elicitation methods: **generative** and **discriminative**. Generative methods assume that the consistency between different generation samples provides a reliable estimation of model uncertainty. The more frequently certain facts are covered by the samples, the higher the model's confidence in those facts. In contrast, discriminative methods assess uncertainties by asking the model itself. This is motivated by the findings that models tend to perform better on discriminative tasks (Saunders et al., 2022), and thus they may already possess the capability to estimate the confidence of their own outputs in a discriminative manner.

We first generate one response x as the answer to the query q. Then, x is broken into atomic claims C. Following (Min et al., 2023; Wei et al., 2024; Zhao et al., 2024), each atomic claim contains a single piece of information and must be self-contained. For generative methods, we sample an additional set of responses K, and compare them against the original response x. For each atomic claim in C, we assign it a confidence score.

3.1 Generative Methods

GEN-BINARY. The basic assumption here is that if a fact is frequently conveyed when sampled multiple times, the model is considered "confident" about that fact. For an atomic claim c_i in C, we utilize a natural language inference model $\mathcal{M}_{\text{NLI}}$ to examine whether c_i is supported or not supported by each of the additional samples. Let K_s be the

set of samples supporting c_i. Then, the confidence in c_i is calculated as:

$$Conf(c_i, K) = \frac{|K_s|}{|K|}$$

GEN-MULTI. GEN-MULTI assumes that the model is more confident in facts that are consistently expressed. Unlike GEN-BINARY, it further divides the "not supported" class K_{ns} into "conflict" (K_c) if the fact is presented differently in the sample, and "not mentioned" (K_{nm}) if the fact is not mentioned in the sample. We then calculate the confidence by only considering supporting and conflicting samples:

$$Conf(c_i, K) = \frac{|K_s|}{|K_s| + |K_c|}$$

3.2 Discriminative Methods

DIS-SINGLE. Following (Kadavath et al., 2022; Tian et al., 2023), we directly ask the model whether one single atomic claim is true or false. The probability the model assigns to token "True" ($P(true)$) in its generation is viewed as the confidence. As each atomic claim is judged individually, one advantage of this method is that there is no cross-claim influences when the model makes confidence judgments.

DIS-CONTEXT. In addition to the method where each claim is judged in a self-contained way, we also consider a setting where additional context is provided. Here, the context denotes the passage where the atomic claim is extracted, or the prompt that generates the response. The context helps the model to more accurately locate the atomic claim, and thus potentially leads to better confidence elicitation. $P(true)$, given the context, is then used as the confidence score, just as in DIS-SINGLE.

DIS-RATING. Instead of using $P(true)$, in DIS-RATING, we directly prompt the model to assign a numerical value representing its confidence in the atomic claim c_i. A score of 0 indicates no confidence, while 10 represents maximum confidence. An alternative approach is to use semantic expressions ranging from "Very Uncertain" to "Very Confident". However, Tian et al. (2023) show that LLMs express uncertainty as effectively, or even more effectively, using numerical values rather than words.

3.3 Confidence Fusion Strategies

Since the generative and discriminative methods provide different perspectives of confidence estimation, it is natural to explore fusion methods that combine different confidence estimates for better calibration. After exploration, we consider four simple but effective options: (1) `MinConf` selects the minimum of the two confidence values, providing a conservative estimate; (2) `HMean` calculates the harmonic mean, providing a balanced fusion that weights lower confidence values more heavily, to help prevent overconfident predictions; (3) `ProdConf` multiplies the confidences, ensuring a high final confidence only when both estimates are strong, thereby effectively penalizing discrepancies between the two estimates; and (4) `WAvg` computes the weighted average, with a hyper-parameter weight, prioritizing the more reliable estimate and allowing flexibility to adjust the influence of each confidence source based on prior knowledge or validation results.

4 Evaluation Metrics

4.1 Atomic Calibration

In atomic calibration, we utilize FACTSCORE (Min et al., 2023) to measure the atomic claims' factuality. For each claim, we have its binary factuality label and a continuous confidence score. Since this corresponds to standard calibration scenarios, we follow the conventions from previous work (Kuhn et al., 2022; Zhu et al., 2023; Tian et al., 2023) and use Expected Calibration Error (ECE) (Naeini et al., 2015), Brier Score (Brier, 1950), and AUROC to assess calibration. Details of the three metrics are in Appendix A.

4.2 Macro Calibration

In macro calibration, factuality scores for responses are represented as continuous values from 0 to 1, therefore, the atomic calibration metrics mentioned above are not directly applicable. Apart from the Spearman Correlation, we propose two novel metrics UCCE and QCCE that extend ECE to continuous factuality scores.

Spearman Correlation. Following (Zhang et al., 2024; Huang et al., 2024), we calculate Spearman Correlation to assess whether samples with higher factuality have correspondingly higher confidence scores. Compared to Pearson Correlation, it focuses on assessing the rank correlation, is robust to outliers and does not require that data is in normal distribution.

Uniform Continuous Calibration Error (UCCE). UCCE evaluates how well a model's continuous predictions align with the true values by dividing the predicted values into bins of equal size. This metric assesses the model's calibration across the range of predictions. The calibration error is computed by comparing the mean predicted and true values within each bin.

Let $\hat{y}_i$ represent the predicted value for the i-th sample, y_i the true value for the i-th sample, B_m the set of samples falling into the m-th bin, M the number of bins, and N the total number of samples. The UCCE is computed as follows:

$$\text{UCCE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \frac{\sum_{i \in B_m} \hat{y}_i}{|B_m|} - \frac{\sum_{i \in B_m} y_i}{|B_m|} \right|.$$

UCCE is particularly effective for scenarios where predicted values are uniformly distributed.

Quantile Continuous Calibration Error (QCCE). QCCE uses a similar approach to UCCE but utilizes quantile binning, ensuring that each bin contains an equal number of samples. Consequently, each bin has a size of $|B_m| = \frac{N}{M}$. QCCE is advantageous for skewed or non-uniform distributions of predicted values, as it avoids sparsity in any bin by ensuring a balanced representation across all bins. The QCCE is calculated as follows:

$$\begin{aligned}
\text{QCCE} &= \frac{1}{M} \sum_{m=1}^{M} \left| \frac{1}{|B_m|} \sum_{i \in B_m} \hat{y}_i - \frac{1}{|B_m|} \sum_{i \in B_m} y_i \right| \\
&= \frac{1}{N} \sum_{m=1}^{M} \left| \sum_{i \in B_m} \hat{y}_i - \sum_{i \in B_m} y_i \right|.
\end{aligned}$$

5 Experiments and Results

5.1 Experiment Setup

Models. We utilize seven LLMs from three model families with varying sizes: Llama3 Instruct (8B and 70B) (Meta, 2024), Mistral Instruct (7B and 8x7B) (Jiang et al., 2023), and Qwen2 Instruct (7B, 52B-A14B, and 72B) (Yang et al., 2024).

Datasets. We use three datasets for long-form QA: *Bios* (Min et al., 2023), which contains 500 individuals from Wikipedia with varying levels of popularity, for which models are tasked to generate biographies; *LongFact* (Wei et al., 2024) extends *Bios*

	Bios			LongFact			WildHallu		
	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	AUROC ↑
Llama3-8B-Instruct									
DIS-CONTEXT	35.5	35.8	74.5	11.9	13.6	74.4	12.5	16.5	**83.5**
DIS-RATING	26.8	29.0	71.1	**3.5**	12.0	66.9	**5.3**	**15.2**	79.8
DIS-SINGLE	32.6	33.9	74.5	14.3	15.2	69.8	19.2	20.9	79.3
GEN-BINARY	**10.0**	**17.8**	**83.1**	8.5	**11.4**	**77.3**	11.1	**15.2**	82.0
GEN-MULTI	37.4	37.3	64.2	12.6	13.1	58.5	21.9	22.1	65.4
Mistral-7B-Instruct									
DIS-CONTEXT	24.8	26.0	77.5	15.7	16.1	75.3	20.6	21.7	79.8
DIS-RATING	44.5	42.5	65.0	10.0	14.2	67.9	19.7	23.9	68.1
DIS-SINGLE	30.2	30.7	75.2	20.4	20.5	66.6	24.0	24.6	75.1
GEN-BINARY	**13.7**	**19.0**	**81.9**	**8.4**	**11.5**	**80.1**	**12.7**	**17.0**	**81.3**
GEN-MULTI	42.2	41.8	65.0	13.4	13.9	61.7	26.6	26.4	64.2
Qwen2-7B-Instruct									
DIS-CONTEXT	26.5	28.3	75.5	13.9	14.8	77.9	17.2	19.4	81.2
DIS-RATING	41.5	39.7	64.2	**3.5**	11.7	62.6	**8.2**	18.1	70.4
DIS-SINGLE	29.3	30.4	75.5	16.1	16.8	74.7	18.7	20.3	80.1
GEN-BINARY	**10.9**	**16.7**	**83.8**	6.3	**9.9**	**81.9**	9.5	**14.0**	**82.5**
GEN-MULTI	41.7	41.1	65.6	11.6	12.1	62.8	21.0	21.0	64.4

Table 1: Atomic Calibration Results. All the numbers are in percentages.

and includes 1,140 questions covering 38 manually-selected topics; *WildHallu* (Zhao et al., 2024) includes 7,917 entities derived from one million user-chatbot interactions in real-world settings.

Atomic Facts Generation and Verification. For all three datasets, we apply a FACTSCORE-based (Min et al., 2023) factuality assessment approach. We first use GPT-4o to decompose the entire response into atomic facts, with each fact containing only a single piece of information. These atomic facts are then verified using GPT-4o, cross-referenced with evidence from Wikipedia and Google Search. The detailed prompts for generating atomic facts are provided in Appendix B.

Confidence Elicitation. We use P(true) (Kadavath et al., 2022), Self-Rating (Tian et al., 2023), and Semantic Entropy (SE) (Kuhn et al., 2022) as the baseline confidence elicitation methods. They are all calculated in response-level. For GEN-BINARY, we apply the Llama3-8B-Instruct for better NLI performance. In WAvg, we assign a weighting 0.7 to generative methods and 0.3 to discriminative ones, based on validation results.

5.2 Results

Overall, the tested LLMs are not well-calibrated at the atomic fact level. Table 1 lists our main atomic calibration results, which indicate that different confidence elicitation methods yield varying calibration scores. Although there is no universally

accepted threshold for low ECE, a well-calibrated model typically achieves an ECE close to 1%, as shown in (Guo et al., 2017) and (Zhu et al., 2023). However, even with the most robust method, GEN-BINARY, the ECE scores remain around 10%, indicating a significant calibration gap. Among the models, Qwen2-7B-Instruct demonstrates slightly better calibration compared to the other two.

Atomic calibration can enhance macro calibration. Table 2 shows the main results of response-level calibration. For the five atomic-level methods, we calculate the average confidence of the facts in a response to obtain the response-level confidence. The results indicate that atomic calibration leads to better overall results compared to the baseline methods, highlighting the helpfulness of more fine-grained calibration analysis.

Confidence fusion can further improve both atomic and macro calibration. Table 4 presents the results of various confidence fusion strategies at the atomic level (with response-level results in Appendix D). The best results are consistently achieved by fusion methods. Among these strategies, the weighted average (WAvg) is the most effective. The benefits of the fusion strategies are further discussed in §6.1. Interestingly, we find that combining methods within the same confidence type (such as DIS-RATING with DIS-CONTEXT) *does not lead to improved calibration* (in Appendix D). A case study showing the effectiveness of confidence fusion is in Figure 4.

	Bios			LongFact			WildHallu		
	SC ↑	UCCE ↓	QCCE ↓	SC ↑	UCCE ↓	QCCE ↓	SC ↑	UCCE ↓	QCCE ↓
Llama3-8B-Instruct									
P(true)	30.2	45.1	46.0	18.9	16.3	16.8	40.5	25.7	26.1
Self-Rating	40.5	38.7	39.3	21.5	14.1	14.3	50.2	18.6	19.0
SE	42.1	37.4	38.0	23.0	13.5	13.8	52.0	17.8	18.2
Dis-Context	55.4	34.0	33.2	29.7	5.6	5.4	65.9	9.5	9.1
Dis-Rating	73.8	25.7	25.7	34.1	**2.9**	**2.8**	**71.7**	**3.6**	**3.8**
Dis-Single	58.0	27.2	27.2	20.9	8.7	8.4	55.9	14.4	14.1
Gen-Binary	**79.8**	**5.6**	**5.6**	**52.7**	3.0	3.0	70.0	7.8	7.2
Gen-Multi	71.4	35.8	36.0	37.5	11.6	11.8	62.6	22.0	24.1
Mistral-7B-Instruct									
P(true)	32.8	44.5	44.9	22.0	16.7	17.0	41.2	24.3	25.2
Self-Rating	42.3	37.1	37.5	26.1	14.5	14.7	52.0	18.1	19.3
SE	44.1	36.5	36.7	28.0	13.8	14.1	53.4	17.4	18.5
Dis-Context	**79.7**	**8.3**	8.2	47.9	4.1	4.2	72.3	**6.1**	**5.7**
Dis-Rating	55.0	41.4	41.5	40.8	4.4	4.2	60.4	16.9	16.7
Dis-Single	70.3	16.0	16.0	32.8	8.8	9.0	65.3	10.4	9.7
Gen-Binary	74.9	8.5	**8.0**	**64.1**	**2.5**	**2.5**	**73.9**	10.3	10.3
Gen-Multi	60.7	38.7	38.7	49.6	11.9	12.6	65.6	26.2	27.5
Qwen2-7B-Instruct									
P(true)	33.5	45.0	44.8	28.3	11.2	11.6	35.4	12.7	12.9
Self-Rating	48.2	24.3	23.9	36.7	6.9	7.0	48.0	9.8	9.5
SE	49.8	22.9	22.6	38.9	6.5	6.4	49.2	8.9	9.2
Dis-Context	66.5	14.8	14.4	40.6	3.9	3.9	66.8	**4.0**	**3.7**
Dis-Rating	63.0	40.7	41.0	29.9	4.4	4.4	54.0	9.1	8.9
Dis-Single	52.8	19.8	19.3	30.9	4.9	5.1	60.4	5.3	5.6
Gen-Binary	**72.4**	**5.4**	**5.7**	**67.6**	**2.0**	**1.8**	**72.2**	6.5	6.4
Gen-Multi	43.1	38.8	38.3	52.1	11.4	11.8	63.2	21.6	23.9

Table 2: Response Calibration Results. All the numbers are in percentages.

Larger model size does not necessarily result in better calibration. Table 3 compares the calibration levels of models with different sizes. Our two key findings are: **(1)** With generative methods, there is little difference in calibration between larger and smaller models; **(2)** With discriminative methods, *larger models generally provide better calibration*. We hypothesize that this is because discriminative methods require models to self-assess the confidence of their own outputs, and larger models typically possess stronger discriminative abilities (Saunders et al., 2022).

6 Discussion

6.1 Confidence Methods Alignment

To further explore the reasons behind the improvements provided by confidence fusion, we show the correlation between different confidence elicitation methods in Figure 2 (using *WildHallu* as the study case and more results are in Appendix E). Our findings are summarized as follows:

Confidence methods within the same type are

better aligned. In Figure 2, warmer colors indicate higher Spearman correlation scores. Confidence elicitaton methods of the same type (top left for generative and bottom right for discriminative) show stronger correlations compared to those across different types. This helps to explain why fusing generative and discriminative methods are effective, since these two types capture different aspects of uncertainty and are complementary to each other.

The alignment is stronger at the response level than at the atomic level. When comparing atomic and macro calibration, we observe that the alignment is stronger for the latter. In atomic calibration, several methods display weak correlations (indicated in blue), while the correlations are generally higher at response level (indicated in red). Similarly, methods from different types show more disagreement than those of the same type. This highlights the need for future research on the discrepancies between generative and discriminative confidence elicitation methods, as well as how to better unify these approaches.

6

	Bios			LongFact			WildHallu		
	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	AUROC ↑
GEN-BINARY									
Llama3-8B-Instruct	10.0	17.8	83.1	8.5	11.4	77.3	11.1	15.2	82.0
Llama3-70B-Instruct	10.0	16.5	82.5	8.3	9.3	73.7	9.5	12.3	78.3
Mistral-7B-Instruct	13.7	19.0	81.9	8.4	11.5	80.1	12.7	17.0	81.3
Mistral-8x7B-Instruct	12.3	18.5	79.8	7.8	9.0	76.3	9.8	13.4	77.8
Qwen2-7B-Instruct	10.9	16.7	83.8	6.3	9.9	81.9	9.5	14.0	82.5
Qwen2-57B-Instruct	10.5	18.1	82.3	7.8	10.0	78.3	9.2	13.6	81.7
Qwen2-72B-Instruct	11.2	16.6	83.4	7.6	8.3	76.6	8.6	11.9	77.7
DIS-RATING									
Llama3-8B-Instruct	26.8	29.0	71.1	3.5	12.0	66.9	5.3	15.2	79.8
Llama3-70B-Instruct	10.6	19.3	73.2	4.2	8.0	74.2	4.3	11.5	81.2
Mistral-7B-Instruct	44.5	42.5	65.0	10.0	14.2	67.9	19.7	23.9	68.1
Mistral-8x7B-Instruct	15.3	22.6	70.8	5.3	8.6	72.6	7.6	14.7	72.9
Qwen2-7B-Instruct	41.5	39.7	64.2	3.5	11.7	62.6	8.2	18.1	70.4
Qwen2-57B-Instruct	23.2	27.0	69.3	2.2	9.8	71.3	5.2	15.2	77.2
Qwen2-72B-Instruct	11.4	21.0	71.6	6.1	7.7	77.1	4.0	11.7	79.2

Table 3: Atomic Calibration Results. All the numbers are in percentages. Our two key findings are: **(1)** for generative methods, there is no significant difference in calibration between larger and smaller models; **(2)** however, with discriminative methods, *larger models tend to provide better calibration.*

	Bios			LongFact			WildHallu		
	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	AUROC ↑
GEN-BINARY	10.0	17.8	83.1	8.5	11.4	77.3	11.1	15.2	82.0
DIS-RATING	26.8	29.0	71.1	3.5	12.0	66.9	5.3	15.2	79.8
DIS-CONTEXT	35.5	35.8	74.5	11.9	13.6	74.4	12.5	16.5	83.5
MinConf	**6.2**	17.1	83.2	10.7	12.2	77.4	9.0	13.9	85.8
HMean	9.8	17.4	84.0	4.1	11.0	79.6	5.6	13.3	**87.0**
ProdConf	7.4	**16.7**	84.1	13.4	12.8	79.6	11.5	14.1	87.0
WAvg	10.9	17.4	**84.4**	**3.3**	**10.3**	**79.9**	**5.1**	**13.0**	87.0

Table 4: Atomic calibration results of confidence fusion strategies for Llama3-8B-Instruct, demonstrating the effectiveness of confidence fusion in combining generative and discriminative confidence estimates.

6.2 Confidence Across Different Positions

As each long-form response contains multiple atomic facts, we analyze how confidence and factuality scores evolve during the generation process. Specifically, we divide all atomic facts C into five equal parts along the generation process. Part 1 represents the beginning of the generation, and part 5 corresponds to the end. We calculate the average confidence score for each part of the responses and present the results in Figure 3.

With discriminative methods, models exhibit decreasing confidence in atomic facts as the generation progresses. We observe similar trends across all discriminative methods. This contrasts with previous findings, which used logits as a measure of confidence and found that models tend to become more confident during long generation sequences (Zhang et al., 2023a). Our results show that dis-criminative methods indicate lower confidence in the model's output toward the latter parts of the generation.

With generative methods, the model shows the lowest average confidence in the middle part of the generation. We hypothesize that this is because the tested models tend to provide general introductions and conclusions at the beginning and the end of the generation. During consistency checking, these statements are frequently cross-referenced, leading to higher confidence. For example, in *Bios*, statements like "[a person] is famous" or "[a person] made a significant impact in his field" are often repeated across samples. On the contrary, in the middle parts where the models address more specific facts about individuals' lives, careers and achievements, they tend to cover different aspects and details.

7

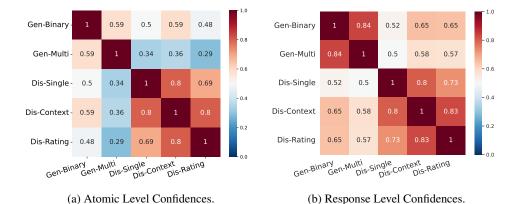(a) Atomic Level Confidences.   (b) Response Level Confidences.

Figure 2: Heatmaps comparing the Spearman correlation between generative and discriminative confidence elicitation methods in Llama3-8B-Instruct on *WildHallu*. Warmer colors are for higher correlations.
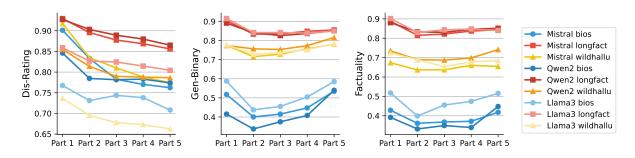

Figure 3: Average confidence scores across different parts of long-form responses. For discriminative methods, confidence decreases as the generation progresses, while generative methods show the lowest confidence in the middle sections.

7 Related Work

Atomic Facts Generation and Verification. Long-form responses often contain both correct and incorrect statements, which impact the overall factuality assessments. Min et al. (2023) propose breaking long responses into atomic facts and calculating the precision of these fact pieces to determine the overall factuality score. Wei et al. (2024) and Zhao et al. (2024) extend this paradigm by expanding the dataset to include more domains beyond biographies. Song et al. (2024) design VERISCORE for diverse long-form generation tasks that feature both verifiable and unverifiable content. Chiang and Lee (2024) introduce D-FACTSCORE, specifically designed for content with ambiguous entities. Overall, the approach of breaking long-form responses into atomic facts for verification is nowadays widely accepted in factuality assessment.

Uncertainty and Calibration in Long-form Generations. Existing research on uncertainty estimation and calibration primarily focuses on multiple-choice or short-form questions (Zhu et al., 2023; Kuhn et al., 2022; Lin et al., 2023; Tian et al.,

2023; Ulmer et al., 2024). However, limited work has explored calibration for long-form generations. Huang et al. (2024) proposed a unified calibration framework for all text generation tasks, comparing distributions of both correctness and the associated confidence of responses. Band et al. (2024) introduced linguistic calibration, where models explicitly express their uncertainty during long-form generation. Zhang et al. (2024) proposed LUQ, an uncertainty estimation method tailored to long-form generation, demonstrating its effectiveness in ensembling different LLMs. None of the aforementioned studies have systematically examined calibration at a fine-grained level. Our work aims to fill this gap with a focus on factuality.

8 Conclusion

We introduce atomic calibration, a fine-grained approach for evaluating LLM calibration at the atomic claim level, addressing the limitations of traditional response-level calibration in long-form generation. Our experiments show that atomic calibration is well-suited for long-form generation and complements macro calibration. We also demon-

Figure 4: An example from *WildHallu* dataset by Mistral-7B-Instruct. We only select five atomic facts for demonstration. The example shows the effectiveness of calculating confidence in atomic level with fusion strategy.

strated that combining different confidence elicitation methods leads to better calibration results and atomic calibration can provide deeper insights into model confidence during the generation process. This approach offers a more reliable way to assess and improve the trustworthiness of LLM outputs in real-world applications.

Limitation

First, our work primarily focuses on the factuality aspect of LLMs. As mentioned in Section 2, the task t can be various aspects of the quality of a long-form response, such as coherence, creativity, writing style, and more. Unlike previous studies that use the overall quality of long-form responses to evaluate calibration (Huang et al., 2024), we concentrate specifically on factuality in this paper. We argue that the hallucination problem is among the most significant challenges faced by LLMs (Zhang et al., 2023b; Huang et al., 2023).

Second, we test the calibration only on open-source LLMs for two main reasons: (1) After assessing the atomic and macro calibration levels of LLMs, our next step is to adjust the model to better reflect its confidence (*i.e.*, for better calibration). Closed-source models are not directly applicable to this calibration process. (2) Our discrimination methods typically require logit access, which is generally unavailable in closed-source models. If logits are accessible, our methods can be directly applied to closed-source models without affecting the atomic calibration process.

Ethics Statement

Our research adheres to strict ethical standards. We ensured compliance with the licenses of all datasets and models used. No human participants were involved in our experiments. After thorough assessment, we do not anticipate any additional ethical concerns or risks related to our work.

9

References

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. 2024. Linguistic calibration of long-form generations. In *Forty-first International Conference on Machine Learning*.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Cheng-Han Chiang and Hung-yi Lee. 2024. Merging facts, crafting fallacies: Evaluating the contradictory nature of aggregated factual claims in long-form generations. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2734–2751, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. *Preprint*, arXiv:1706.04599.

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

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. *Preprint*, arXiv:2311.05232.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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. *Preprint*, arXiv:2207.05221.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *Preprint*, arXiv:2305.19187.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluis Marquez. 2024. Factual confidence of LLMs: on reliability and robustness of current estimators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4554–4570, Bangkok, Thailand. Association for Computational Linguistics.

Meta. 2024. Llama 3 model card.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. 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.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

OpenAI. 2022. Chatgpt blog post. https://openai.com/blog/chatgpt. Accessed: 2024-09-06.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *Preprint*, arXiv:2206.05802.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. *Preprint*, arXiv:2406.19276.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. 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.

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. 2023. Llama 2: Open foundation and finetuned chat models. *Preprint*, arXiv:2307.09288.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. 2024. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15440–15459, Bangkok, Thailand. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *Preprint*, arXiv:2403.18802.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. Luq: Long-text uncertainty quantification for llms. *Preprint*, arXiv:2403.20279.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023a. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. Association for Computational Linguistics.

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

Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, and Yejin Choi. 2024. Wildhallucinations: Evaluating long-form factuality in llms with real-world entity queries. *Preprint*, arXiv:2407.17468.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2023. On the calibration of large language models and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9778–9795, Singapore. Association for Computational Linguistics.

Appendix

A Atomic Calibration Metrics

ECE In computing the Expected Calibration Error (ECE), the predictions are sorted and divided into a fixed number of bins K. The predicted value of each test instance falls into one of the bins. ECE uses empirical estimates as follows:

$$ECE = \sum_{i=1}^{K} P(i) \cdot |o_i - e_i| \,,$$

where o_i is the true fraction of positive instances in bin i, e_i is the mean of the post-calibrated probabilities for the instances in bin i, and $P(i)$ is the empirical probability (fraction) of all instances that fall into bin i. The lower the ECE value is, the better a model is calibrated.

Brier Score The Brier score measures the accuracy of probabilistic predictions. For unidimensional predictions, it is strictly equivalent to the mean squared error as applied to predicted probabilities. Suppose that on each of the n occasions an event can occur in only one of r possible classes or categories and on one such occasion i, the forecast probabilities are $f_{i1}, f_{i2}, \ldots f_{ir}$, that the event will occur in classes $1, 2, \ldots r$, respectively. The r classes are chosen to be mutually exclusive and exhaustive so that

$$\sum_{i=1}^{r} f_{ij} = 1, i = 1, 2, 3, \ldots n.$$

The Brier Score P is then defined by

$$P = \frac{1}{n} \sum_{j=1}^{r} \sum_{i=1}^{n} (f_{ij} - E_{ij})^2 \,,$$

where E_{ij} takes the value 1 or 0 according to whether the event occurred in class j or not.

AUROC Following (Kuhn et al., 2022), AUROC metric is equivalent to the probability that a randomly chosen correct answer has a higher confidence score than a randomly chosen incorrect answer. Higher scores are better for AUROC, and perfect confidence score is 1, while a random confidence measure would be 0.5.

B   Prompts

Prompts to Generate Responses

Bios:
Tell me a bio of <ENTITY>.

- -

LongFact:
Can you detail the concept of Gravitational Lensing and how it is utilized in modern astronomy for the study and understanding of the universe?

Can you explain the process and importance of customer journey mapping in creating effective marketing strategies?

- -

WildHallu:
In a paragraph, could you tell me what you know about <ENTITY>?

Table 5: Prompts to generate responses. We use the default prompt template for *Bios* and *WildHallu*. For *LongFact*, it has corresponding Prompt for each question, so we only list two examples here.

GEN-BINARY

```
Sample: {sample}
Sentence: {sentence}
Is the sentence supported by the sample above?
You should answer the question purely based on the given sample and not your
own knowledge.
Do not output the explanations.
Your answer should be within "yes" or "no".
Answer:
```

Table 6: Prompt for GEN-BINARY.

GEN-MULTI

```
Sample: {sample}
Sentence: {sentence}
Is the sentence supported, refuted, or not mentioned by the sample above?
You should answer the question purely based on the given sample and not your
own knowledge.
Do not output the explanations.
Your answer should be within "supported", "refuted", or "not mentioned".
Answer:
```

Table 7: Prompt for GEN-MULTI.

DIS-SINGLE

Please evaluate the truthfulness of the following statement based on your internal knowledge.
Respond only with "True" or "False" with no other information.
Statement: {atomic_fact_piece}
True or False:

Table 8: Prompt for DIS-SINGLE.

DIS-CONTEXT

You are given a topic and a statement related to that topic.
Evaluate the truthfulness of the statement based solely on your internal knowledge.
A statement is true only if it is factually accurate according to what you know.
Do not use the topic to evaluate the statement.
Respond with "True" or "False" only, without any additional information.
Topic: {context}
Statement: {atomic_fact_piece}
True or False:

Table 9: Prompt for DIS-CONTEXT.

DIS-RATING

Please rate your confidence in the truth of the following statement on a scale from 0 to 10.
A confidence of 0 means you are not confident at all whether the statement is true, while 10 means you are absolutely confident the statement is true.
The topic provides context for the statement, but focus on the statement itself when rating your confidence.
Provide only the numerical value for your confidence level, bounded by dollar signs ($), without any additional information.
Topic: {context}
Statement: {atomic_fact_piece}
Confidence:

Table 10: Prompt for DIS-RATING.

C Experiment Details

We use vLLM (Kwon et al., 2023) for our LLM inference tasks, with the following parameters: temperature = 1, top-p = 0.95, and a maximum output of 512 tokens. For discriminative confidence elicitation methods, we set the temperature to 0 and only consider the top 10 logits. For generative methods, we use $N = 20$ samples. The experiments are conducted on A100-SXM-40GB GPUs. Running the discriminative methods takes 30 minutes for 500 samples, while the generative methods take 1.3 hours for the same number of samples. We use GPT-4o as the auxiliary model for generating atomic claims and fact-checking the LLM.

D Confidence Fusion Results

	Bios			LongFact			WildHallu		
	SC ↑	UCCE ↓	QCCE ↓	SC ↑	UCCE ↓	QCCE ↓	SC ↑	UCCE ↓	QCCE ↓
GEN-BINARY	79.8	5.6	5.6	52.7	3.0	3.0	70.0	7.2	7.8
DIS-RATING	73.8	25.7	25.7	34.1	2.8	2.9	71.7	3.8	3.6
DIS-CONTEXT	55.4	33.2	34.0	29.7	5.4	5.6	65.9	9.1	9.5
MinConf	80.0	**5.5**	**5.4**	45.2	5.2	5.3	75.8	4.0	4.1
HMean	81.7	11.5	11.5	54.5	2.0	2.0	**78.7**	**3.2**	**3.0**
ProdConf	**82.2**	7.7	7.1	54.9	14.2	14.3	78.7	12.2	12.1
WAvg	81.8	10.9	11.0	**55.6**	**1.5**	**0.9**	77.2	5.0	5.0

Table 11: Macro calibration results of confidence fusion strategies for Llama3-8B-Instruct.

	Bios			LongFact			WildHallu		
	SC ↑	UCCE ↓	QCCE ↓	SC ↑	UCCE ↓	QCCE ↓	SC ↑	UCCE ↓	QCCE ↓
GEN-BINARY	74.9	8.0	8.5	64.1	2.5	2.5	73.9	10.3	10.3
DIS-RATING	55.0	41.5	41.4	40.8	4.2	4.4	60.4	16.7	16.9
DIS-CONTEXT	79.7	8.2	8.3	47.9	4.2	4.1	72.3	5.7	6.1
MinConf	82.4	**2.1**	**3.2**	61.1	3.2	3.3	77.3	**3.1**	**3.7**
HMean	83.5	5.8	6.2	62.5	1.9	2.1	77.9	6.5	6.6
ProdConf	**83.8**	13.1	13.0	62.5	10.6	10.4	78.0	7.2	7.1
WAvg	82.0	8.0	8.0	**65.1**	**1.8**	**2.0**	**78.4**	8.9	8.9

Table 12: Macro calibration results of confidence fusion strategies for Mistral-7B-Instruct.

	Bios			LongFact			WildHallu		
	SC ↑	UCCE ↓	QCCE ↓	SC ↑	UCCE ↓	QCCE ↓	SC ↑	UCCE ↓	QCCE ↓
GEN-BINARY	72.4	5.7	5.4	**67.6**	1.8	2.0	72.2	6.4	6.5
DIS-RATING	63.0	41.0	40.7	29.9	4.4	4.4	54.0	8.9	9.1
DIS-CONTEXT	66.5	14.4	14.8	40.6	3.9	3.9	66.8	3.7	4.0
MinConf	76.0	**4.1**	**3.9**	61.1	3.1	3.4	72.6	3.8	3.9
HMean	**76.2**	7.8	8.0	62.4	1.4	**1.4**	74.7	**2.3**	**2.2**
ProdConf	76.0	12.9	12.6	62.5	10.4	10.4	74.9	12.7	12.7
WAvg	75.7	7.5	8.4	66.4	**1.2**	1.4	**75.6**	4.5	4.3

Table 13: Macro calibration results of confidence fusion strategies for Qwen2-7B-Instruct.

	Bios			LongFact			WildHallu		
	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	AUROC ↑
GEN-BINARY	13.7	19.0	81.9	8.4	11.5	80.1	12.7	17.0	81.3
DIS-RATING	44.5	42.5	65.0	10.0	14.2	67.9	19.7	23.9	68.1
DIS-CONTEXT	24.8	26.0	77.5	15.7	16.1	75.3	20.6	21.7	79.8
MinConf	14.1	18.3	82.0	8.6	12.7	80.7	**7.6**	16.0	83.2
HMean	14.3	18.3	82.1	7.6	12.2	81.0	11.5	16.6	83.4
ProdConf	14.2	18.3	82.3	9.5	13.0	81.0	7.9	**15.9**	83.5
WAvg	**10.6**	**16.6**	**84.7**	**5.5**	**10.7**	**82.1**	12.3	16.4	**84.4**

Table 14: Atomic calibration results of confidence fusion strategies for Mistral-7B-Instruct.

	Bios			LongFact			WildHallu		
	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	AUROC ↑
GEN-BINARY	10.9	16.7	83.8	6.3	9.9	81.9	9.5	14.0	82.5
DIS-RATING	41.5	39.7	64.2	3.5	11.7	62.6	8.2	18.1	70.4
DIS-CONTEXT	26.5	28.3	75.5	13.9	14.8	77.9	17.2	19.4	81.2
MinConf	11.3	16.9	82.4	6.3	10.3	80.5	5.0	13.8	82.6
HMean	11.1	16.9	82.7	2.7	9.6	81.7	**4.9**	13.6	83.8
ProdConf	12.4	17.0	83.1	8.3	10.6	81.7	7.2	13.7	83.9
WAvg	**10.7**	**15.9**	**84.8**	**2.6**	**9.2**	**82.8**	6.8	**13.5**	**84.3**

Table 15: Atomic calibration results of confidence fusion strategies for Qwen2-7B-Instruct.
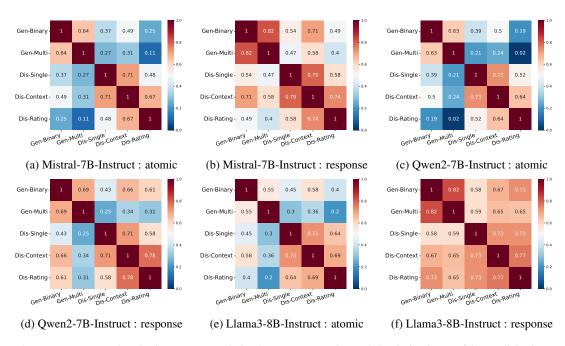
E Confidence Alignment

Figure 5: Heatmaps comparing the Spearman correlation between generative and discriminative confidence elicitation methods for *Bios*. Results shown for Mistral-7B-Instruct, Qwen2-7B-Instruct, and Llama3-8B-Instruct.
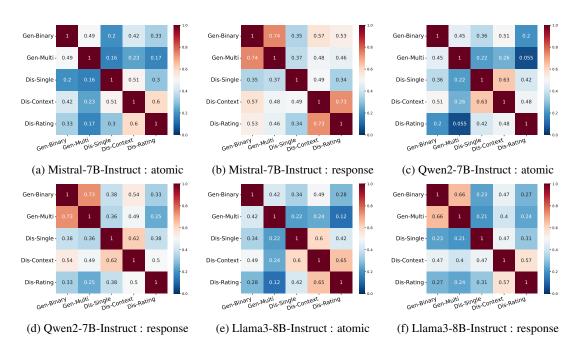

Figure 6: Heatmaps comparing the Spearman correlation between generative and discriminative confidence elicitation methods for *LongFact*. Results shown for Mistral-7B-Instruct, Qwen2-7B-Instruct, and Llama3-8B-Instruct.

(a) Mistral-7B-Instruct : atomic (b) Mistral-7B-Instruct : response (c) Qwen2-7B-Instruct : atomic

(d) Qwen2-7B-Instruct : response (e) Llama3-8B-Instruct : atomic (f) Llama3-8B-Instruct : response
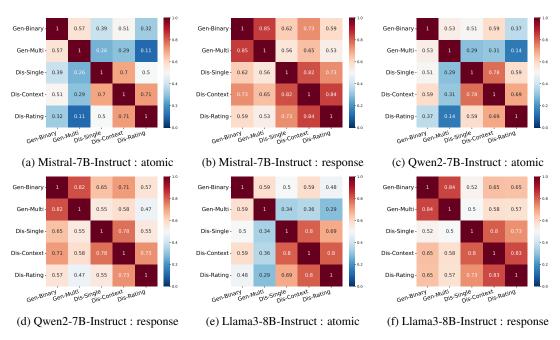
Figure 7: Heatmaps comparing the Spearman correlation between generative and discriminative confidence elicitation methods for *WildHallu*. Results shown for Mistral-7B-Instruct, Qwen2-7B-Instruct, and Llama3-8B-Instruct.