Atomic Calibration of LLMs in Long-Form Generations

Anonymous ACL submission

Abstract

Large language models (LLMs) often suffer from hallucinations, posing significant challenges for real-world applications. Confidence calibration, as an effective indicator of hallucination, is thus essential to enhance the trustworthiness of LLMs. Prior work mainly focuses on short-form tasks using a single response-level score (macro calibration), which is insufficient for long-form outputs that may contain both accurate and inaccurate claims. In this work, we systematically study **atomic calibration**, which evaluates factuality calibration at a fine-grained level by decomposing long responses into atomic claims. We further categorize existing confidence elicitation methods into **discriminative** and **generative** types, and propose two new confidence fusion strategies to improve calibration. Our experiments demonstrate that LLMs exhibit poorer calibration at the atomic level during long-form generation. More importantly, atomic calibration uncovers insightful patterns regarding the alignment of confidence methods and the changes of confidence throughout generation. This sheds light on future research directions for confidence estimation in long-form generation.

1 Introduction

While large language models (LLMs) (Touvron et al., 2023; Jiang et al., 2023; OpenAI, 2022) excel in various tasks, they still struggle with trustworthiness 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 predictions 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

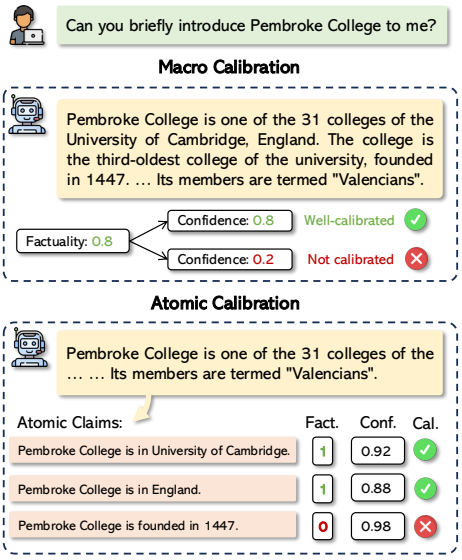


Figure 1: Comparison between traditional macro calibration in response-level and our atomic calibration. The `Fact.` label is assigned by fact-checking module. We only list three atomic claims for illustration.

(Zhu et al., 2023; Mahaut et al., 2024). Improved calibration enables more reliable confidence estimation, warning users when **not** to trust the model and *thus mitigating the impact of hallucinations*.

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), using datasets like TriviaQA and Natural Questions (Joshi et al., 2017), where answers are typically under 10 words. In contrast, real-world queries often elicit much longer responses (Zhang et al., 2024; Yang et al., 2024b), spanning hundreds or thousands of words. In such cases, response quality is not simply binary, as answers may mix accurate and inaccurate statements.

Recent work has begun to address calibration in long-form generation (Zhang et al., 2024; Huang et al., 2024; Liu et al., 2023; Fadeeva et al., 2024; Jiang et al., 2024). Several approaches estimate a single confidence score for the entire response

(*macro calibration*; upper, Figure 1), while others assess confidence at the level of atomic claims (Liu et al., 2023; Fadeeva et al., 2024; Jiang et al., 2024) (*atomic calibration*; lower, Figure 1). However, previous work leaves several key research questions unanswered: Why is it important to evaluate calibration at the atomic-claim level? What factors influence calibration results at this level? What patterns can be observed by analyzing calibration at the atomic-claim level?

In this work, we systematically examine **atomic calibration**: A long response is decomposed into atomic claims, each containing a single factual statement, and confidence scores are assigned using various elicitation methods. To analyze different confidence elicitation methods, we categorize them into **discriminative** (intrinsic confidence estimation) and **generative** (external confidence assessment). Our experiments on three long-form QA datasets with seven LLMs reveal that: **(1)** Models that appear *well-calibrated* at the response level *perform poorly at the atomic level* (Figure 2, Table 1); **(2)** Leveraging atomic calibration enhances macro calibration (Table 2). These two reasons highlight the need for research on atomic-level calibration to develop better-calibrated models.

We further investigate the characteristics of discriminative and generative confidence. Our analysis yields two main findings: **(1)** Discriminative and generative methods are *complementary*; combining them improves calibration, while combinations within the same category offer limited gains. **(2)** *Generative methods* maintain consistent calibration across different model sizes, whereas *discriminative methods* benefit from increased model sizes. Motivated by finding (1), we propose two novel fusion strategies based on *confidence agreement* to integrate generative and discriminative confidence. Our strategies outperform existing fusion methods.

Our atomic-level analysis (Section 6) offers deeper insights into **confidence method alignment** and **confidence changes during generation**. Confidence methods within the same category align better, explaining why cross-category fusion is more effective. Interestingly, with discriminative methods, model confidence in atomic facts tends to *decrease* as generation progresses. In contrast, generative methods show the *lowest average confidence in the middle* of the generation process. These results highlight the necessity of fine-grained calibration evaluation for long-form generation, given us insights on model trustworthiness and usability.

2 Related Work

Atomic Claims 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. Decomposing long-form responses into atomic claims and fact-checking them individually has become a widely adopted pipeline.

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). There is an increasing interest on 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. Another line of work (Liu et al., 2023; Fadeeva et al., 2024; Jiang et al., 2024; Yuan et al., 2024) decomposes sentences into atomic claims and assigns confidence scores to each claim. However, a unified definition of atomic-level calibration remains lacking. Clarifying this concept and identifying key influencing factors are essential steps toward improving calibration in long-form generation.

3 Atomic Calibration

For a language model $\mathcal{M}$, let $x \sim M(x \mid q)$ denote the response generated by $\mathcal{M}$ for a query q, $x \in \mathcal{X}$. Let $y \in \mathcal{Y}_t$ be the corresponding label, representing a quality score ranging from 0 to 1 for a specific task $t \in T$. Unlike multiple-choice or short-form

questions, which mainly assess correctness, tasks in T cover diverse dimensions such as factuality, coherence, and creativity.

We define a probability prediction function $f : \mathcal{X} \rightarrow \Delta^{|\mathcal{Y}_t|}$, where $\Delta^{|\mathcal{Y}_t|}$ denotes the $|\mathcal{Y}_t|$-dimensional probability simplex. Here, $f(x)_y$ represents the probability assigned to label y for a generated output x. In this work, we focus on calibrating factuality, as hallucinations are a well-known issue in LLMs (Zhang et al., 2023b; Huang et al., 2023), and the factuality of atomic claims can be assessed objectively. In this setting, $\mathcal{Y}$ denotes $\mathcal{Y}_t$ for the factuality task t, where $\mathcal{Y} \subseteq [0, 1]$ reflects the factuality level of a response. Following 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*

$$\mathbb{P}(y \mid f(x)_y = \beta) = \beta, \quad \forall \beta \in [0, 1].$$

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:*

$$\mathbb{P}\left(y_i \mid f(c_i)_{y_i} = \beta_i\right) = \beta_i, \quad \forall \beta_i \in [0, 1].$$

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:

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

4 Confidence Elicitation Methods

In this section, we define two types of confidence elicitation methods: **generative** and **discriminative**. We then introduce two novel confidence fusion strategies that considers confidence agreement when combining confidence scores. For the response x to a query q, x is broken into atomic claims C. Following previous work (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.

4.1 Generative Methods

Generative methods assume that the consistency between different generation samples provides a reliable estimation of model uncertainty (Zhang et al., 2024; Jiang et al., 2024). Generally, an additional natural language inference (NLI) model is used to calculate the consistency. In particular, we have the following two variations:

GEN-BINARY. The basic assumption 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 NLI 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" (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:

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

4.2 Discriminative Methods

Discriminative methods assess uncertainties by asking the model itself (Tian et al., 2023; Xiong et al., 2023). 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.

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) demonstrate LLMs achieve comparable or even better results using numerical values.

4.3 Confidence Fusion Strategies

Combining confidence scores has proven effective for calibration (Huang et al., 2024; Rivera et al., 2024), but existing methods typically only use a single fixed weight, α, to combine the scores, ignoring the **confidence disagreement**. For instance, Rivera et al. (2024) computes the weighted average (**`WAvg`**) for confidence scores A and B: $C = A \cdot \alpha + B \cdot (1 - \alpha)$. This approach *does not account for the agreement between the two scores*. For example, when $\alpha = 0.5$, confidences of 0 and 1 are treated the same as confidences of 0.4 and 0.6, although the former may indicate higher uncertainty due to a larger disagreement. To address this, we propose two simple but effective methods that consider confidence disagreement $d = B - A$.

`AdjustedAlpha` adjusts the weight α based on the confidence difference:

$$\alpha' = \alpha + \gamma_a \cdot d,$$

where γ_a is a small constant (*e.g.*, 0.1), followed by $C' = A \cdot \alpha' + B \cdot (1 - \alpha')$.

`DampedFusion` applies a damping factor based on the agreement:

$$\gamma(d) = 1 - k \cdot |d|,$$

where k is a small constant (*e.g.*, 0.02) that controls the damping sensitivity, followed by $C' = C \cdot \gamma(d)$. For baselines, we also include: **`MinConf`**, which selects the minimum confidence; **`HMean`**, which calculates the harmonic mean; and **`ProdConf`**, which multiplies the confidences.

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., 2024a).

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* 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. These atomic facts are then verified using GPT-4o, cross-referenced with

	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.

evidence from Wikipedia and Google Search. The detailed prompts for generating atomic facts are provided in Appendix I.

Confidence Elicitation. We use P(true) (Kadavath et al., 2022), Self-Rating (Tian et al., 2023), Semantic Entropy (SE) (Kuhn et al., 2022), and Sum of Eigenvalues (EigV) (Lin et al., 2023) as the baseline confidence elicitation methods. They are all calculated in response-level. For GEN-BINARY, we apply the Llama-3-8B-Instruct for better NLI performance. For `WAvg`, `AdjustedAlpha`, and `DampedConf`, we use a separate validation set for hyper-parameter tuning.

Metrics. We use Expected Calibration Error (ECE) (Naeini et al., 2015) and Brier Score (BS) (Brier, 1950) as the primary metrics. These metrics are applicable to both atomic and macro calibration (see details in Appendix A), enabling a direct comparison between them. Additionally, we include AUROC to evaluate atomic calibration and Spearman Correlation for a more instance-specific assessment in macro calibration.

5.2 Results

Overall, the tested LLMs are not well-calibrated at the atomic fact level. Table 1 lists our main atomic calibration results. 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.

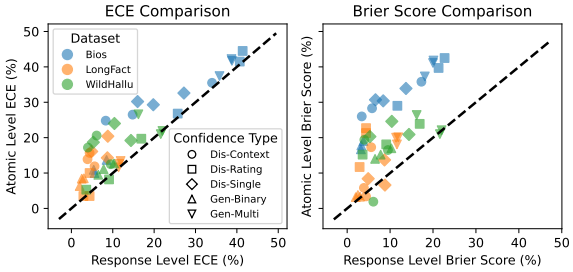


Figure 2: Comparison of atomic level and response-level calibration for ECE and Brier Score. Atomic-level performance is generally **worse** than response-level performance, with data points consistently lying **above** the identity line.

Models that appear well-calibrated at the response level still perform poorly at the atomic level Figure 2 compares atomic and response-level scores for ECE and Brier Score across different datasets and confidence types. The data points consistently lie above the identity line, indicating that *atomic-level errors are higher than response-level errors*. This suggests that atomic calibration is crucial for fine-grained evaluation.

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.

	Bios			LongFact			WildHallu		
	ECE ↓	BS ↓	SC ↑	ECE ↓	BS ↓	SC ↑	ECE ↓	BS ↓	SC ↑
Llama3-8B-Instruct									
P(true)	45.1	25.9	30.2	16.3	4.8	18.9	25.7	13.5	40.5
Self-Rating	38.7	23.4	40.5	14.1	4.2	21.5	18.6	12.9	50.2
SE	37.4	21.8	42.1	13.5	3.4	23.0	17.8	11.7	52.0
EigV	36.8	21.2	43.0	13.0	3.2	23.8	17.2	11.3	53.0
DIS-CONTEXT	34.0	17.3	55.4	5.6	1.9	29.7	9.5	4.8	65.9
DIS-RATING	25.7	11.7	73.8	**2.9**	1.6	34.1	**3.6**	**3.5**	**71.7**
DIS-SINGLE	27.2	13.7	58.0	8.7	2.6	20.9	14.4	7.3	55.9
GEN-BINARY	**5.6**	**3.3**	**79.8**	3.0	**1.1**	**52.7**	7.8	4.6	70.0
GEN-MULTI	35.8	18.1	71.4	11.6	2.7	37.5	22.0	10.5	62.6
Mistral-7B-Instruct									
P(true)	44.5	27.1	32.8	16.7	7.4	22.0	24.3	19.8	41.2
Self-Rating	37.1	26.4	42.3	14.5	6.5	26.1	18.1	14.5	52.0
SE	36.5	23.9	44.1	13.8	3.7	28.0	17.4	14.3	53.4
EigV	35.9	23.3	45.3	13.3	3.5	36.5	16.9	13.9	54.4
DIS-CONTEXT	**8.3**	**3.4**	**79.7**	4.1	1.4	47.9	**6.1**	**4.3**	72.3
DIS-RATING	41.4	22.7	55.0	4.4	1.7	40.8	16.9	9.8	60.4
DIS-SINGLE	16.0	6.6	70.3	8.8	3.1	32.8	10.4	6.5	65.3
GEN-BINARY	8.5	3.8	74.9	**2.5**	**1.0**	**64.1**	10.3	5.0	**73.9**
GEN-MULTI	38.7	20.1	60.7	11.9	2.8	49.6	26.2	13.4	65.6
Qwen2-7B-Instruct									
P(true)	45.0	27.9	33.5	11.2	5.6	28.3	12.7	15.6	35.4
Self-Rating	24.3	25.1	48.2	6.9	4.9	36.7	9.8	14.7	48.0
SE	22.9	23.1	49.8	6.5	3.7	38.9	8.9	13.9	49.2
EigV	22.4	22.5	50.7	6.2	3.5	39.8	8.5	13.5	50.2
DIS-CONTEXT	14.8	5.8	66.5	3.9	1.7	40.6	**4.0**	3.5	66.8
DIS-RATING	40.7	21.3	63.0	4.4	1.9	29.9	9.1	6.3	54.0
DIS-SINGLE	19.8	8.4	52.8	4.9	2.4	30.9	5.3	4.8	60.4
GEN-BINARY	**5.4**	**3.2**	**72.4**	**2.0**	**0.9**	**67.6**	6.5	**3.2**	**72.2**
GEN-MULTI	38.8	20.0	43.1	11.4	2.6	52.1	21.6	9.5	63.2

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

The confidence fusion method considering confidence agreement outperforms other methods. Table 4 presents the results of various confidence fusion strategies at the atomic level (more results in Appendix F). The best performance is consistently achieved by `AdjustedAlpha` and `DampedFusion`. Notably, we observe that combining methods of the same confidence type (e.g., DIS-RATING with DIS-CONTEXT) *does not lead to improved calibration*. A case study demonstrating the effectiveness of confidence fusion is shown in Figure 11.

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 3 (using *WildHallu* as the study case and more results are in Appendix G). Our findings are summarized as follows:

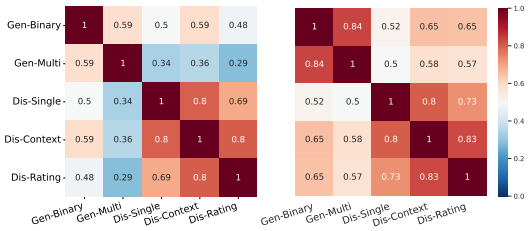


Figure 3: Heatmaps of Spearman Correlation between different confidences in Llama3-8B-Instruct on *WildHallu*. Warmer colors indicate higher correlations. Atomic level: left; response level: right.

Confidence methods within the same type are better aligned. In Figure 3, 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 cross-category fusion strategies are effective, since these two types *capture different aspects of*

	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 with different model sizes. All the numbers are in percentages.

	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
`AdjustedAlpha`	**4.1**	15.8	**85.2**	3.4	10.2	**80.4**	**4.3**	12.6	**88.3**
`DampedFusion`	5.0	**15.6**	84.7	3.5	**9.8**	80.0	4.8	**12.4**	87.9

Table 4: Atomic calibration results of different confidence fusion strategies for Llama3-8B-Instruct. The fusion results are based on GEN-BINARY and DIS-RATING.

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.

6.2 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 4.

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 discriminative 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

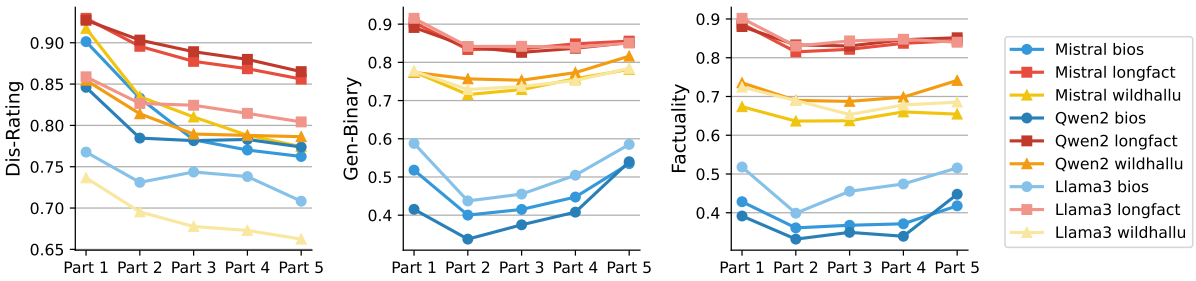


Figure 4: Average confidence scores across different parts of long-form responses. For discriminative methods, confidence decreases as the generation progresses, while generative methods show the lowest confidence in the middle sections.

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.

6.3 The Utilities of Atomic Calibration

While the primary goal of atomic calibration is to provide fine-grained calibration evaluation for models, we also explore its utilities in several downstream tasks, including: (1) **Selective Question Answering** (Kamath et al., 2020; Cole et al., 2023; Yang et al., 2023), which involves setting a confidence threshold to selectively reject low-confidence answers, ensuring that only high-confidence responses are retained; (2) **LLM-Ensemble** (Zhang et al., 2024), which leverages multiple models to generate responses to the same question, selecting the answer with the highest confidence, thereby combining the strengths of each model; and (3) **Atomic Claims Reunion** (Thirukovalluru et al., 2024; Jiang et al., 2024), which involves sampling multiple responses, breaking then into atomic claims, evaluating their confidence, and reassembling only high-confidence claims to produce a more reliable final answer. Among these applications, we observe consistent improvements in factuality with atomic-level examination. Detailed experimental settings and results can be found in Appendix C.

It is important to note that, unlike previous work on Selective Question Answering (Huang et al., 2024) and LLM-Ensemble (Zhang et al., 2024) for long-form generation, which mainly rely on atomic-level confidence estimation to *enhance the overall quality of responses* (with responses either being entirely accepted or rejected), *Atomic Claims Reunion* **does not** require an overall response-level score. Instead, it relies entirely on the confidence of atomic claims to select and combine the most accurate claims. This means the final answer may contain claims from different sampled answers. More importantly, we observe that models with better atomic-level calibration (e.g., Qwen2 in Table 7, Appendix C) exhibit greater improvements after the reunion process, emphasizing the importance of examining and refining atomic calibration.

7 Conclusion

Our main contributions are three-fold: (1) We systematically study **atomic calibration**, which evaluates confidence calibration at the level of individual atomic claims. Our experiments reveal that models that appear *well-calibrated* at the response level *perform poorly at the atomic level*. (2) To analyze confidence elicitation methods, we categorize them into discriminative and generative methods. We also propose two novel fusion strategies to combine the confidence scores based on confidence agreement. (3) Our atomic-level analysis provides further insights into confidence methods alignment and confidence changes during generation. We find with discriminative methods, models show *decreasing confidence* in atomic facts as generation progresses. In contrast, generative methods show the *lowest average confidence in the middle* of the generation. Last but not least, we demonstrate the utilities of atomic calibration and propose for future research on more fine-grained confidence in long-form generation.

Limitation

First, our work primarily focuses on the factuality aspect of LLMs. As mentioned in Section 3, 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.

Third, in this work, we mainly focus on exploring different confidence elicitation methods and therefore do **not** apply post-hoc calibration techniques such as histogram binning or temperature scaling. Applying these methods makes it difficult to disentangle improvements due purely to *elicitation* from those due to *recalibration*. To isolate the contribution of our elicitation designs and to avoid conflating them with downstream post-processing effects, we report raw atomic and macro confidence scores, leaving a systematic study of post-hoc techniques to future work.

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.

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.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9367–9385, 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.

Mingjian Jiang, Yangjun Ruan, Prasanna Sattigeri, Salim Roukos, and Tatsunori Hashimoto. 2024. Graph-based uncertainty metrics for long-form language model outputs. *Preprint*, arXiv:2410.20783.

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.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

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.

Xin Liu, Muhammad Khalifa, and Lu Wang. 2023. Litcab: Lightweight language model calibration over short- and long-form responses. *Preprint*, arXiv:2310.19208.

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.

Mauricio Rivera, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Combining confidence elicitation and sample-based methods for uncertainty quantification in misinformation mitigation. *Preprint*, arXiv:2401.08694.

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.

Raghuveer Thirukovalluru, Yukun Huang, and Bhuwan Dhingra. 2024. Atomic self-consistency for better long form generations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12681–12694, Miami, Florida, USA. Association for Computational Linguistics.

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 fine-tuned 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.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *Preprint*, arXiv:2306.13063.

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. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Qi Yang, Shreya Ravikumar, Fynn Schmitt-Ulms, Satvik Lolla, Ege Demir, Iaroslav Elistratov, Alex Lavaee, Sadhana Lolla, Elaheh Ahmadi, Daniela Rus, Alexander Amini, and Alejandro Perez. 2023. Uncertainty-aware language modeling for selective question answering. *Preprint*, arXiv:2311.15451.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Sen Yang, Nigel Collier, Dong Yu, and Deqing Yang. 2024b. Logu: Long-form generation with uncertainty expressions. *arXiv preprint arXiv:2410.14309*.

Yige Yuan, Bingbing Xu, Hexiang Tan, Fei Sun, Teng Xiao, Wei Li, Huawei Shen, and Xueqi Cheng. 2024. Fact-level confidence calibration and self-correction. *Preprint*, arXiv:2411.13343.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics.

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, the better a model is calibrated.

When labels are continuous values between 0 and 1, the ECE formulation can be generalized. Instead of binning instances based on binary outcomes, the continuous predictions are grouped into bins according to their predicted probability values. Specifically, the observed calibration error o_i in each bin is the average of the continuous label values for the instances in that bin, and e_i is the mean predicted probability for those instances. This ensures that the calibration error accounts for all possible real-valued outcomes within the range [0,1], providing a more nuanced measure of calibration when the labels are continuous.

Brier Score The Brier score measures the accuracy of probabilistic predictions. In binary classification, it compares the predicted probability of the positive class with the actual binary outcome (0 or 1). The Brier score is defined as:

$$P = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \,,$$

where $\hat{y}_i$ is the predicted probability for instance i and $y_i \in \{0, 1\}$ is the actual binary outcome.

For continuous labels in the range $[0, 1]$, the Brier score can still be used, where y_i is now a continuous value between 0 and 1. In this case, the Brier score becomes equivalent to the mean squared error (MSE) between predicted probabilities and the true values, and minimizing the Brier score for continuous labels is analogous to minimizing MSE. Both metrics aim to reduce the squared differences between predicted and true values, with lower scores indicating better calibration and accuracy.

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.

Spearman Correlation Following Zhang 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.

B Statistics in Atomic and Macro Calibration

To assess the confidence of a model, we generate responses using various questions (e.g., N questions). For each response, a single confidence score is too coarse-grained. Instead, we evaluate the confidence of each atomic claim, with an average of M atomic claims per response. These individual confidences are then aggregated into a response-level confidence score.

Atomic calibration is computed over MN data points, where M is the average number of atomic claims per response, and N is the number of responses. In contrast, response-level calibration is based on N data points. This distinction highlights the trustworthiness of the model's confidence at both the atomic and response levels, providing a more granular view of its performance.

From the above discussion, it follows that to ensure sufficient data points for atomic calibration, MN must be large. In our datasets, N typically exceeds 1,000, ensuring that MN remains robust even when some responses have only a few calims.

The detailed generation statistics are further illustrated in Figures 5, 6, and 7. Figure 5 presents the average answer length, while Figure 6 shows the average number of atomic claims per answer. Finally, Figure 7 highlights the percentage of answers containing fewer than 10 atomic facts.

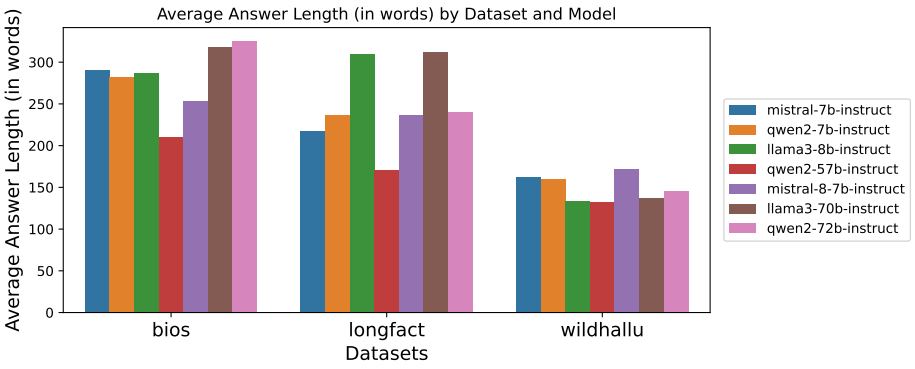


Figure 5: Average answer length (in words) for different models on `Bios`, `longfact`, and `wildhallu`.

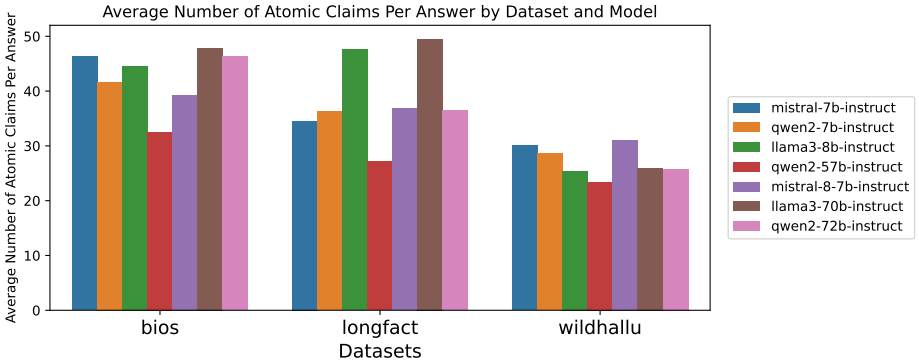


Figure 6: Average number of atomic claims per answer for different models on `Bios`, `longfact`, and `wildhallu`.

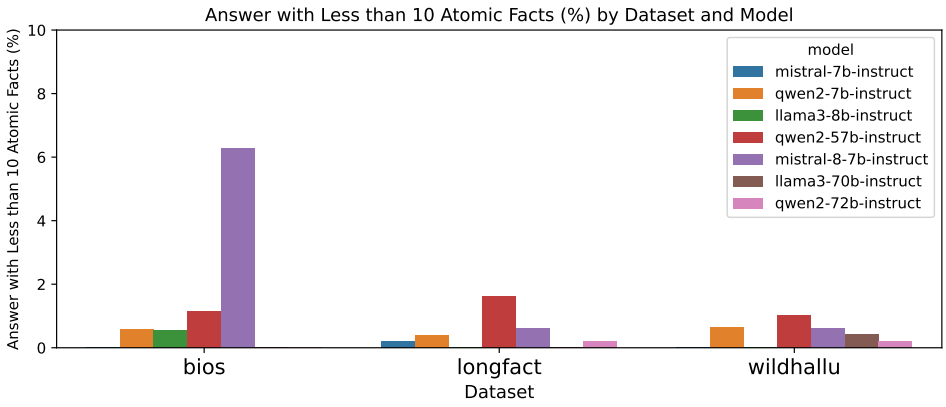


Figure 7: Answer with less than 10 atomic facts (%) by dataset and model. Notably, `Mistral-8×7B-Instruct` has 6% short answers. Human evaluation reveals that these responses are primarily instances where the model refuses to answer.

C Applications

Selective Question Answering: Selective question answering involves setting a confidence threshold, which can be derived from a validation set, to selectively reject questions with low confidence. This approach aims to improve the overall factuality of the responses by eliminating potentially unreliable answers.

Using the *Bios* dataset, we evaluate the performance of three models: Llama3-8B-Instruct, Mistral-7B-Instruct, and Qwen2-7B-Instruct with DIS-GEN and Semantic Entropy (SE). We observe an improvement in overall factuality as we gradually rejected more questions (from 0% to 10%). The table below illustrates this trend, highlighting the utility of DIS-GEN in identifying accurate responses and improving selective question answering. Our comparison between DIS-GEN and Semantic Entropy (SE) indicates that DIS-GEN brings more significant improvements in factuality, suggesting that better calibration methods can substantially enhance the results of selective question answering.

LLM Ensemble: In the *LLM Ensemble* method, we use three models to generate answers to the same question and select the response from the model with the highest confidence. This approach aims to enhance factuality by leveraging the strengths of each model. The **Answer Distribution (AD)** shows the proportion of the final response contributed by each model, highlighting the benefit of ensemble methods. The table below presents the results of applying this method on the Bios and WildHallu datasets, comparing two different selection strategies: DIS-GEN and SE.

The results demonstrate that the ensemble method with DIS-GEN significantly improves factuality. For instance, in the Bios dataset, the factuality score increases from 0.475 to 0.556, and in the WildHallu dataset, it increases from 0.655 to 0.752. In contrast, using SE results in no improvements, with factuality scores even lower than the best individual model (0.484 vs 0.502 and 0.671 vs 0.701) for the Bios and WildHallu datasets, respectively. These findings suggest that **ensembling does not always guarantee better results**, and the selection strategy, such as DIS-Gen, plays a crucial role in improving factuality.

Atomic Reunion: In *Atomic Reunion*, for each question, we begin by sampling the model's output five times (this is also what we need to calculate GEN-DIS. These outputs are then broken down into atomic claims, which are individual, verifiable statements. Each claim is evaluated for confidence, and only those with a high confidence level are retained. Subsequently, we prompt a LLM, such as GPT-4o, to reassemble the selected atomic claims into a cohesive and factually accurate response.

This method seeks to enhance factuality by utilizing smaller, more manageable pieces of information, allowing the LLM to generate a more reliable answer by combining only high-confidence atomic claims. The table below presents the factuality scores of the new answers generated through the Atomic Reunion approach. As observed, this approach leads to a significant improvement in factuality compared to the baseline models.

D Reliability of Atomic Facts Generation and Verification

The processes of atomic fact generation and verification have been extensively studied and validated in prior work (Min et al., 2023; Wei et al., 2024; Zhao et al., 2024). For instance, FActScore (Min et al., 2023) reports an error rate of 2%. In this work, we leverage their pipeline while employing stronger models from GPT-3.5 and GPT-4o, to further enhance performance.

We, the authors, conducted additional tests comparing GPT's atomic decompositions with ground-truth manual segmentations. We manually selected 30 samples for this evaluation. The results of our assessment are as follows:

- **Consistency:** Over 10 trials, GPT demonstrated a high inter-run consistency of 95%, indicating stable and repeatable outcomes.
- **Error Rate:** The error rate, which includes missing or overly segmented claims and misclassification of factuality, was measured at 6.4%. This error rate is manageable within the context of our calibration framework, suggesting that the model's atomic fact generation is reliable.

E 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

Refuse Rate	Llama3-8B-Instruct	Mistral-7B-Instruct	Qwen2-7B-Instruct
SE			
0%	0.475	0.403	0.502
5%	0.479	0.407	0.506
7.5%	0.483	0.411	0.511
10%	0.485	0.416	0.518
DIS-Gen			
0%	0.475	0.403	0.502
5%	0.496	0.419	0.517
7.5%	0.511	0.438	0.533
10%	**0.528**	**0.465**	**0.557**

Table 5: Factuality Scores with Varying Refuse Rates for Selective Question Answering

Model	Bios		WildHallu	
	Factuality Scores	Answer Distribution	Factuality Scores	Answer Distribution
DIS-Gen				
Llama3-8B-Instruct	0.475	31%	0.655	29%
Mistral-7B-Instruct	0.403	23%	0.631	27%
Qwen2-7B-Instruct	0.502	46%	0.701	44%
Ensemble	**0.556**	/	**0.752**	/
SE				
Llama3-8B-Instruct	0.475	18%	0.655	15%
Mistral-7B-Instruct	0.403	37%	0.631	44%
Qwen2-7B-Instruct	**0.502**	45%	**0.701**	41%
Ensemble	0.484	/	0.671	/

Table 6: LLM Ensemble Results on Bios and WildHallu Datasets

Model	Before Atomic Reunion	After Atomic Reunion
DIS-Gen		
Llama3-8B-Instruct	0.475	0.501
Mistral-7B-Instruct	0.403	0.441
Qwen2-7B-Instruct	0.502	**0.575**

Table 7: Factuality Scores Before and After Atomic Reunion using DIS-Gen

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.

F Confidence Fusion Results

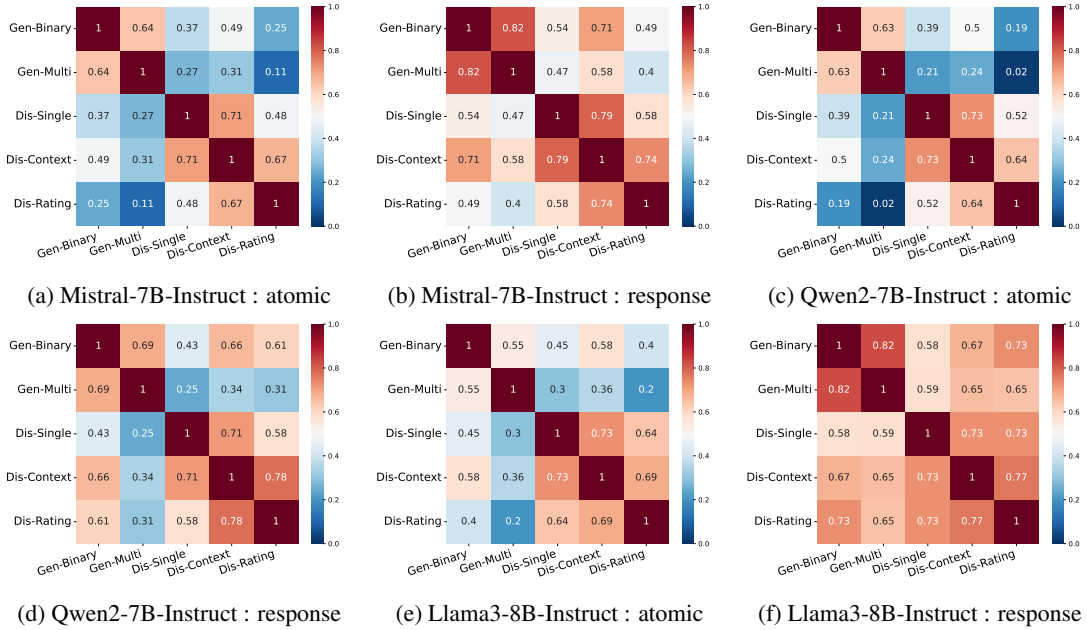
	Bios			LongFact			WildHallu		
	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	AUROC ↑
GEN-BINARY	13.7	19.0	81.9	8.4	11.5	80.1	12.7	17.0	81.3
DIS-RATING	44.5	42.5	65.0	10.0	14.2	67.9	19.7	23.9	68.1
DIS-CONTEXT	24.8	26.0	77.5	15.7	16.1	75.3	20.6	21.7	79.8
MinConf	14.1	18.3	82.0	8.6	12.7	80.7	7.6	16.0	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**
AdjustedAlpha	**9.8**	16.7	**85.0**	5.8	**10.5**	81.8	**6.5**	15.2	84.0
DampedFusion	10.2	**16.5**	84.6	5.9	10.9	81.9	7.1	15.4	83.8

Table 8: Atomic calibration results of confidence fusion strategies for Mistral-7B-Instruct. The fusion results are based on GEN-BINARY and DIS-CONTEXT.

	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
AdjustedAlpha	**8.9**	16.0	84.6	2.9	**9.1**	82.6	**4.5**	**13.2**	84.1
DampedFusion	10.2	**15.8**	84.5	**2.6**	9.3	**82.9**	5.2	13.3	**84.4**

Table 9: Atomic calibration results of confidence fusion strategies for Qwen2-7B-Instruct. The fusion results are based on GEN-BINARY and DIS-CONTEXT.

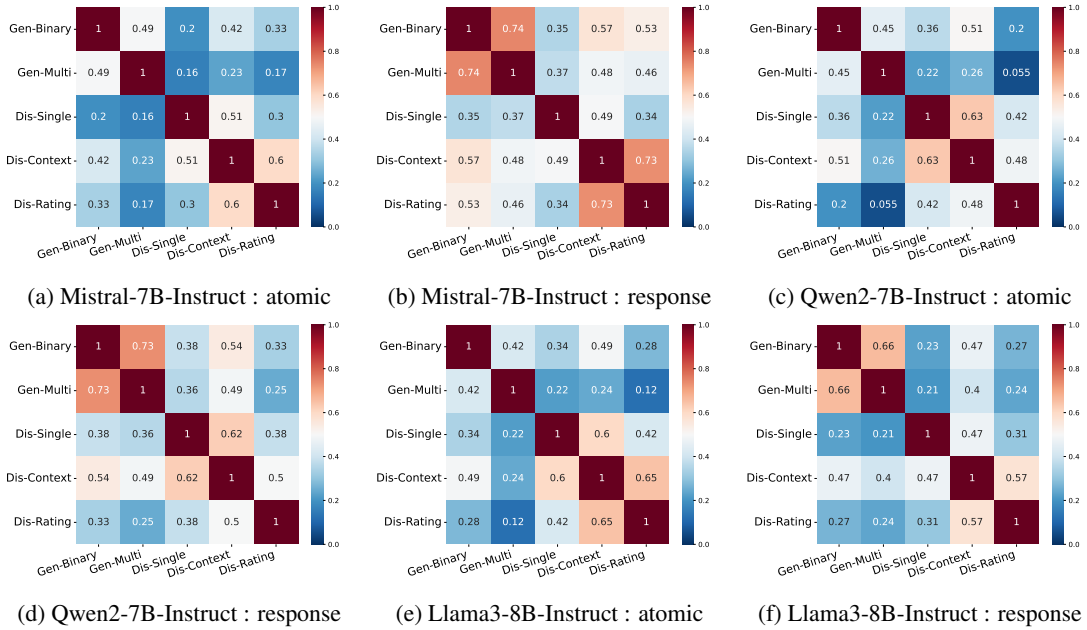
G Confidence Alignment



(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

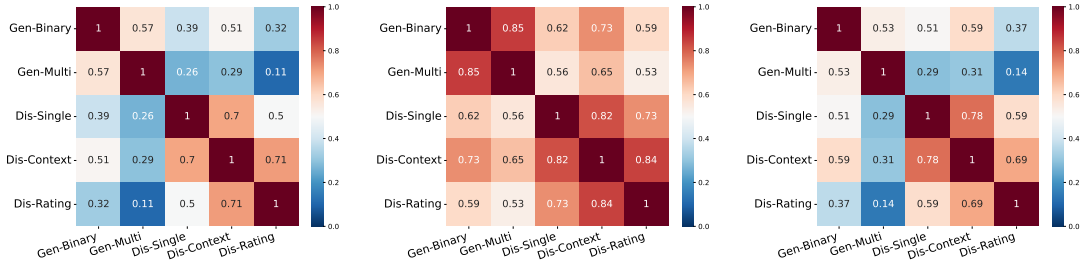
Figure 8: Heatmaps comparing the Spearman correlation between generative and discriminative confidence elicitation methods for *Bios*. Results shown for Mistral-7B-Instruct, Qwen2-7B-Instruct, and Llama3-8B-Instruct.



(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 9: 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 10: 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.

H Case Study

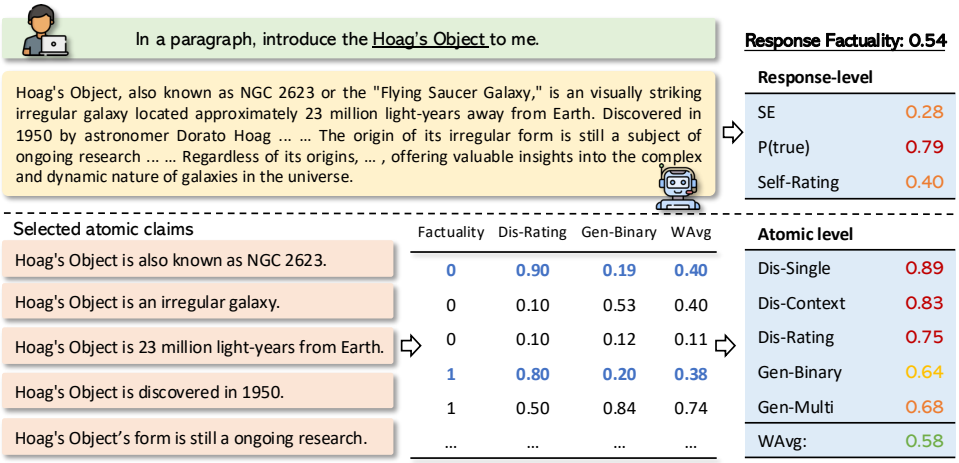


Figure 11: An example from *WildHallu* dataset by Mistral-7B-Instruct. We only select five atomic facts for demonstration. The example shows the effectiveness of calculating confidence in atomic level with fusion strategy.

I 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 10: 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 11: 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 12: 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 13: 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 14: 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 15: Prompt for DIS-RATING.