Balancing Truthfulness and Informativeness with Uncertainty-Aware Instruction Fine-Tuning

Anonymous ACL submission

Abstract

Instruction fine-tuning (IFT) can increase the informativeness of large language models (LLMs), but may reduce their truthfulness. This trade-off arises because IFT steers LLMs to generate responses containing long-tail outputs that include knowledge that may not be well-covered during pre-training. As a result, models become more informative but less accurate when generalizing to unseen tasks. In this paper, we empirically demonstrate how unfamiliar knowledge in IFT datasets can undermine the truthfulness of LLMs, and propose two new IFT paradigms, $UNIT_{cut}$ and $UNIT_{ref}$, to address this issue. $UNIT_{cut}$ detects and removes unfamiliar knowledge from IFT data to enhance truthfulness, while $UNIT_{ref}$ adds a reflection section that flags uncertain claims. Experiments demonstrate that $UNIT_{cut}$ substantially improves truthfulness and $UNIT_{ref}$ preserves informativeness while addressing hallucinations by signaling uncertainty.[1]

1 Introduction

General-purpose alignment pursues responses that "provide a clear, complete, and detailed answer with additional information valuable for users." (Zheng et al., 2023), where informativeness plays a critical role. To encourage detailed, user-valuable responses, prior works have invested significant effort in collecting high-quality Instruction Fine-Tuning (IFT) data and fine-tuning LLMs on them (Zhao et al., 2024a; Liu et al., 2024; Zhou et al., 2023). However, using such high-quality IFT data to steer LLMs to be informative might harm their truthfulness, as they might be taught to extrapolate beyond their parametric knowledge to provide extensive details. For example, in LIMA (Zhou et al., 2023), LLMs are tuned to cite "(Klämbt, 2009)" to support the claim that "brain glial cells migrate."—a reference likely too niche to be familiarised as part

---

[1]We will open-source all data, code, and training recipes.
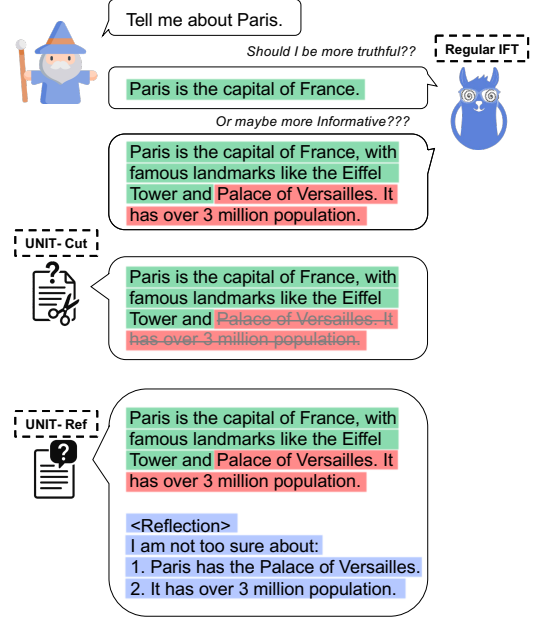

Figure 1: Right: in regular IFT, tuning LLMs for better informativeness may encourage LLMs to produce uncertain claims that are less likely to be correct than certain claims . Left: In uncertain-aware IFT, LLMs are tuned to either leave out uncertain claims ($UNIT_{cut}$) or reflect on them ($UNIT_{ref}$) while maintaining informativeness.

of the model's parametric knowledge during pre-training. This knowledge gap between pre-training and fine-tuning may encourage LLMs to generate informative, confident, but inaccurate answers when generalizing to unseen tasks, inducing confident hallucinations (Gekhman et al., 2024; Kang et al., 2024).

To mitigate hallucinations, previous work focuses on constructing specialized, honesty-oriented training data to improve LLM truthfulness, achieving promising results (Zhang et al., 2024a; Yang et al., 2024b; Band et al., 2024). However, these methods largely overlook how to safely fine-tune LLMs on existing high-quality IFT data to enhance informativeness without reducing truthfulness. This gap is particularly relevant for practition-

ers who depend on IFT to adapt LLMs to specific downstream tasks or domains (Niklaus et al., 2025; Wu et al., 2024).

In this work, we uncover how using high-quality instruction fine-tuning data affects the truthfulness of LLMs, and how to safely use such data without compromising truthfulness. Particularly, we investigate this by answering two research questions in a logical order:

RQ1. Does unfamiliar knowledge in human-annotated IFT datasets affect truthfulness? IFT achieves generalizable informativeness by having detailed long-form generation with diverse, factually-accurate, information-dense instruction-response pairs (e.g., LIMA). However, it remains unclear whether fine-tuning on this dense information that may be unfamiliar to LLMs' parametric knowledge would cost a trade-off in the truthfulness of LLMs (Zhao et al., 2024a).

We first propose (UNcertainty-aware Instruction Tuning by Cutting) $UNIT_{cut}$, where we pinpoint the LLMs' uncertainty about the claims in an instruction fine-tuning dataset and reconstruct a "more familiar" variant of the IFT dataset for instruction fine-tuning. Our findings reveal that incorporating more unfamiliar knowledge in IFT reduces model truthfulness. Furthermore, by removing uncertain knowledge within the IFT dataset using $UNIT_{cut}$, the truthfulness of the responses increases, while the informativeness might be negatively affected.

RQ2. How can we leverage original high-quality IFT data without compromising trustworthiness? While $UNIT_{cut}$ improves truthfulness by brutally removing unfamiliar knowledge from IFT data, this may hurt informativeness by sacrificing valuable details or structures. This raises a key question: can we directly apply the informative, elaborately crafted IFT datasets from prior work while mitigating hallucinations? To address this, we propose the second algorithm $UNIT_{ref}$ (UNcertainty-aware Instruction Tuning by Reflecting), an IFT paradigm that fine-tunes models to report their uncertainty after responses. Specifically, instead of reconstructing a less uncertain variant dataset like $UNIT_{cut}$, $UNIT_{ref}$ appends a "reflection" after each response that lists the uncertain claims to teach the model to reflect on its uncertainty. This approach preserves the original IFT datasets' response information richness while promoting honesty: the model learns to sig-

nal uncertainty, enabling users to perform targeted post-verification of specific claims. A comparison between regular IFT and $UNIT_{ref}$ can be found in Fig. 1.

In summary, our contributions are: (1) We empirically demonstrate that fine-tuning on unfamiliar knowledge in high-quality IFT datasets always reduces the truthfulness of LLMs' responses, and find removing unfamiliar knowledge ($UNIT_{cut}$) an effective algorithm to reduce IFT-caused hallucination (§ 3). (2) We introduce $UNIT_{ref}$, an IFT paradigm that takes advantage of the original high-quality IFT datasets to improve informativeness while preserving trustworthiness by flagging uncertain claims (§ 4).

2 Related Work

LLM Honesty Alignment. Various previous works investigate how to align LLMs to appropriately express their knowns and unknowns (Yang et al., 2024a,b; Xu et al., 2024a; Cheng et al., 2024; Zhang et al., 2024a; Band et al., 2024). Typically this involves (1) prompting the model with information-seeking queries; (2) grading its responses against the ground truth; and (3) fine-tuning the model to express higher confidence for correct answers and lower confidence for incorrect ones. Confidence indicators include refusals (Zhang et al., 2024a), numerical confidence scores (Yang et al., 2024b), or linguistic uncertainty markers (Band et al., 2024; Yang et al., 2024a). While these works focus on constructing their own honesty-oriented training data, they largely overlook how to safely leverage high-quality IFT data (Zhou et al., 2023) to improve other aspects of LLM performance—such as informativeness—without compromising truthfulness. This challenge is particularly relevant for practitioners who rely on elaborate IFT data to adapt LLMs to specific tasks or domains (Niklaus et al., 2025; Wu et al., 2024; Fatemi et al., 2024; Zhang et al., 2024b). Addressing this gap, our work first demonstrates that unfamiliar knowledge in high-quality IFT data can induce hallucination, and we propose $UNIT_{cut}$ and $UNIT_{ref}$ as a safer paradigm for leveraging such data. Our work takes the first step to improve LLM honesty while considering other dimensions, taking informativeness as a case study.

IFT's Impact on Informativeness and Truthfulness. Prior work in IFT alignment emphasizes informativeness as a core element for helpfulness.
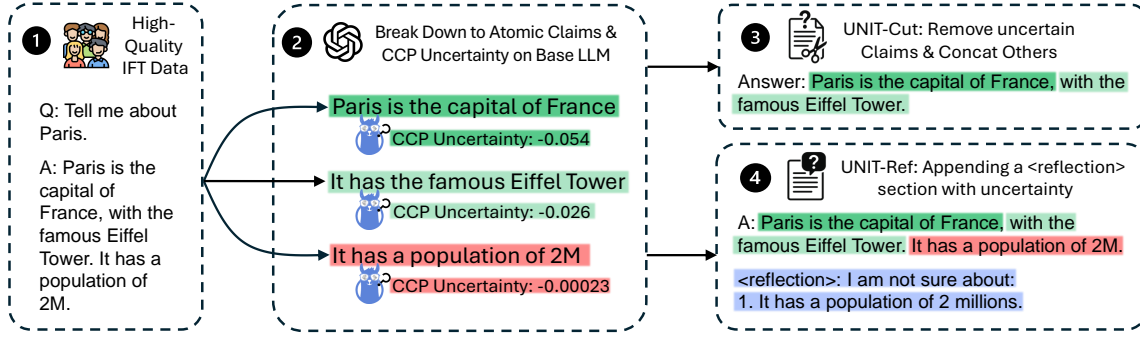
2

Figure 2: Illustration of the procedure of UNIT$_{cut}$ and UNIT$_{ref}$. Given a high-quality IFT dataset (e.g., manually annotated like LIMA and LFRQA), we first measure claim-level uncertainty for each response (❷). Then we construct new IFT data by removing the uncertain claims (❸ UNIT$_{cut}$) or reflecting them after the responses (❹ UNIT$_{ref}$).

For example, Zhou et al. (2023) collects "complete and detailed" responses; Liu et al. (2024) enhances response helpfulness by "depth and details"; and Zhao et al. (2024b) finds that longer responses can benefit alignment. Therefore, our work focuses on how to benefit from informative IFT data without compromising truthfulness. Gekhman et al. (2024) find that IFT rarely increases hallucination with early stop (e.g., under 5 epochs). Zhao et al. (2024a) report that IFT does not degrade performance on factual-knowledge benchmarks. However, we find that even with small epoch numbers, incorporating unfamiliar knowledge can still harm truthfulness.

Uncertainty Measurement for LLMs. Uncertainty measurement is a critical technique for detecting hallucinations, since higher uncertainty often corresponds to lower generation quality (Xiong et al., 2024; Vashurin et al., 2025). Uncertainty measures can be divided into two types: sequence-level measures, which evaluate uncertainty across entire generated sequences (Kuhn et al., 2023; Duan et al., 2024), and claim-level measures, which assess uncertainty for individual factual claims (Fadeeva et al., 2024). In this work, we adopt Claim-Conditioned Probability (CCP) (Fadeeva et al., 2024), identified by Vashurin et al. (2025) as the best-performing claim-level measure, to quantify LLMs' familiarity with the claims in high-quality IFT data.

3 RQ1: Does Unfamiliar Knowledge in IFT Affect Truthfulness?

To investigate RQ1, we design controlled experiments comparing IFT outcomes before and after removing unfamiliar knowledge. We first introduce UNIT$_{cut}$ (§ 3.1), an algorithm that fine-tunes LLMs on IFT data with unfamiliar content removed. We then introduce the training datasets (§ 3.2) and experimental settings (§ 3.3). Finally, we present the results and key takeaways (§ 3.4).

3.1 Methodology - UNIT$_{cut}$

In this section, we introduce UNIT$_{cut}$ (illustrated in Fig. 2 ❷ and ❸), an IFT paradigm with unfamiliar knowledge removed. The training data construction of UNIT$_{cut}$ consists of three steps: First, it measures CCP-based uncertainty (Fadeeva et al., 2024) for atomic claims in IFT data responses. Second, it categorizes claims into familiar or unfamiliar based on a given uncertainty threshold. Third, it concatenates the familiar claims into a new response using a language model rewriter, and fine-tunes the target LLM. The procedure is detailed as follows.

Firstly: Finding Unfamiliar Atomic Claims with CCP. Given an instruction dataset D and a target LLM M, our first step is to measure the uncertainty of all atomic claims in D using the CCP algorithm. The dataset D contains N instruction-response pairs $D = \{(I_i, R_i)\}_{i=1}^N$. Let $x_{i,j}$ denotes the j-th token in response R_i. From each response R_i, CCP extracts a set of atomic factual claims $\mathcal{C}_i = \{C_{i,1}, C_{i,2}, \ldots, C_{i,m_i}\}$, where each $C_{i,j} \subset R_i$ representing a coherent factual statement. For each token $x_{i,j}$ in a claim $C_{i,j}$, the target model M samples the top-K alternatives

$$\mathcal{A}(x_{i,j}) = \{x_{i,j}^1, x_{i,j}^2, \ldots, x_{i,j}^K\},$$

with probabilities $P(x_{i,j}^k \mid x_{i,<j})$, where $x_{i,<j} = \{x_{i,1}, \ldots, x_{i,j-1}\}$. A natural language inference (NLI) model then assesses the semantic relationship between each alternative $x_{i,j}^k$ and the original token $x_{i,j}$ by comparing the contexts $x_{i,<j} \circ x_{i,j}^k$ and $x_{i,1:j} = x_{i,<j} \circ x_{i,j}$, assigning one of three

labels: entailment (e), contradiction (c), or neutral (n). Define:

$$\text{Me}(x_{i,j}) = \left\{ x_{i,j}^k \;\middle|\; \text{NLI}(x_{i,<j} \circ x_{i,j}^k,\, x_{i,1:j}) = e \right\}$$

$$\text{CT}(x_{i,j}) = \left\{ x_{i,j}^k \;\middle|\; \text{NLI}(x_{i,<j} \circ x_{i,j}^k,\, x_{i,1:j}) \in \{e, c\} \right\}$$

where Me (Meaning) intuitively denotes alternative tokens with the same meaning as $x_{i,j}$, while CT (ClaimType) denotes alternative tokens with the same claim type but may contradict $x_{i,j}$. The token-level uncertainty is computed as

$$\text{CCP}(x_{i,j}) = \frac{\sum_{x_{i,j}^k \in Me(x_{i,j})} P(x_{i,j}^k \mid x_{i,<j})}{\sum_{x_{i,j}^l \in CT(x_{i,j})} P(x_{i,j}^l \mid x_{i,<j})},$$

and the overall uncertainty for a claim is aggregated by

$$\text{CCP}_{\text{claim}}(C_{i,j}) = 1 - \prod_{x \in C_{i,j}} \text{CCP}(x).$$

Using CCP, for each response R_i we obtain a set of atomic claims with their corresponding uncertainty values, i.e., $\{(C_{i,j}, \text{CCP}_{\text{claim}}(C_{i,j}))\}_{j=1}^{m_i}$, where a higher CCP value means the model is more uncertain about each claim.

Secondly: Labeling Uncertain Atomic Claims in Responses. In the previous step, we compute CCP-based uncertainty scores $\text{CCP}_{\text{claim}}(C_{i,j})$ for all atomic claims $\mathcal{C}_i = \{C_{i,j}\}_{j=1}^{m_i}$ extracted from response R_i. The next step is to categorize them as familiar or unfamiliar based on the CCP scores.

To do this, we compute the 75th quantile[2] of CCP scores across the entire dataset D, and use this value as the threshold τ for distinguishing unfamiliar (i.e., uncertain) knowledge to the target LLM:

$$\tau = Q_{0.75}\left(\{\text{CCP}_{\text{claim}}(C) \mid C \in \mathcal{C}\}\right).$$

Then, for each claim $C \in \mathcal{C}$, we assign its uncertainty label as follows:

$$\ell(C) = \begin{cases} \text{uncertain}, & \text{if } \text{CCP}_{\text{claim}}(C) > \tau, \\ \text{certain}, & \text{otherwise}. \end{cases}$$

Finally: Removing Unfamiliar Knowledge. For each instruction–response pair (I_i, R_i) we collect the claims previously labeled as *certain* into a list, keeping their order in the original response:

$$\mathcal{L}_i = \big[C_{i,j}\big]_{j:\,\ell(C_{i,j})=\text{certain}}$$

We pass the list to an auxiliary language-model rewriter f_{LLM}. Conditioned on both the instruction I_i and the list $\mathcal{L}_i$, the model returns a fluent answer that contains only vetted information:

$$R_i^{\text{cut}} = f_{\text{LLM}}(I_i, \mathcal{L}_i)$$

Substituting every (I_i, R_i) with (I_i, R_i^{cut}) yields the IFT data with unfamiliar knowledge (above an uncertainty threshold) removed:

$$D^{\text{cut}} = \left\{(I_i, R_i^{\text{cut}})\right\}_{i=1}^{N}$$

If all claims in a response R_i are labeled as uncertain, R_i^{cut} should apologize without providing any information. If all claims in response R_i are labeled as certain, $R_i^{\text{cut}} = R_i$. Finally, we fine-tune the target model M on D^{cut}. This completes UNIT$_{\text{cut}}$, ensuring that no claim with high CCP uncertainty influences the model's subsequent instruction-following behaviour.

Notably, the original CCP is proposed to measure uncertainty for on-policy data (i.e., LLM generations), while we establish a new use case of it to measure LLMs' familiarity to off-policy data (i.e., elaborate IFT data), and prove it effective.

CCP (and other uncertainty measures) was originally designed for information-seeking queries that decompose into independent factual claims. Vashurin et al. (2025) validates CCP's reliability only on information-seeking tasks, but also shows its limitations on non-information-seeking tasks like math and translation. Extending it to other tasks (e.g., reasoning, creative writing, or editing) can be conceptually inaccurate: e.g., in a 2-step reasoning task of first solving a and b given $a = 2 + 2, b = a + 1$; if a model answers "$a = 3$" (wrong) and then "$b = 4$" (correct given the earlier mistake), CCP may flag high uncertainty on the first step but low on the second, conditioned on the generated "$a = 3$". We give more examples where CCP may fail in App. A. Therefore, in this work, we focus our use of CCP on information-seeking IFT data, which is directly relevant to our objective of understanding the impact of unfamiliar knowledge on model truthfulness. We detail our prompts for claim extraction, instruction classification, and LLM rewriting in App. B.

---

[2] We choose 75th quantile for demonstration simplicity. Theoretically, the threshold can be set to other values to control the conservativeness of the algorithm.

3.2 Training Data

We conduct experiments on two datasets: LIMA (Zhou et al., 2023) and LFRQA (Han et al., 2024). LIMA is selected because of its status as a high-quality, human-annotated instruction-following fine-tuning (IFT) dataset. LFRQA also contains long-form human-written instruction-response pairs, but has a stronger focus on information-seeking tasks, where we can apply CCP to measure claim-level uncertainty. Moreover, the queries of LFRQA span across 7 domains (e.g., biomedical, finance, recreation, technology, etc.), which often require domain-specific niche knowledge to answer. Therefore, LFRQA may effectively emulate real-world domain-specific IFT and help us to investigate potential challenges there.

Since our exploration focuses on IFT data and data-centric approaches to alleviate IFT-induced hallucinations, we investigate various combinations of datasets to draw statistically significant observations. Therefore, we vary LFRQA from 10% to 100% of its examples and apply UNIT_{cut}. We then augment each LFRQA subset with LIMA to assess performance when including general-domain helpfulness data.

3.3 Evaluation and Training Details

Truthfulness. We use FactScore (Min et al., 2023) and WildFactScore (Zhao et al., 2024c) to fact-check atomic claims in LLMs' long-form outputs. FactScore prompts LLMs to generate 500 biographies (*Bio*), while WildFactScore prompts to introduce 7K entities absent from Wikipedia (*WildHalu*[3]). FactScore decomposes each text into atomic claims and verifies them using a retrieval-augmented LLM agent. The final truthfulness score is the percentage of atomic claims verified as true.

Informativeness. We investigate how changes in informativeness dynamically affect truthfulness by evaluating both dimensions on benchmark prompts used in FactScore and WildFactScore. To facilitate informativeness measurement on FactScore and WildFactScore's benchmark prompts, we adapt the MT-Bench (Zhou et al., 2023) pairwise LLM-judge format, which, although originally designed to assess general helpfulness, is well-suited for our needs. Because Bio and WildHalu tasks are inherently information-seeking, making assessing helpfulness a reasonable proxy for informativeness.

---

[3]We randomly sample 500 entities from WildHalu for budget control.

Model / Data	LFRQA%	Method	Truth.↑		Info.↑	
			Bio.	Wild.	Bio.	Wild.
Llama3.1-8B / LFRQA	10%	Vanilla	56.54	82.22	22.10	25.90
		UNIT_cut	+3.17↑	+4.91↑	-6.10↓	-9.00↓
	40%	Vanilla	53.41	79.00	15.60	18.75
		UNIT_cut	+12.12↑	+8.68↑	-6.10↓	-8.45↓
	70%	Vanilla	49.15	75.32	16.80	19.70
		UNIT_cut	+20.88↑	+11.16↑	-6.20↓	-8.50↓
	100%	Vanilla	50.15	73.77	15.60	18.80
		UNIT_cut	+20.39↑	+15.10↑	-5.60↓	-8.35↓
Llama3.1-8B / LFRQA +LIMA	10%	Vanilla	50.97	79.43	34.70	33.75
		UNIT_cut	+10.30↑	+1.94↑	-11.70↓	-8.85↓
	40%	Vanilla	47.51	73.89	29.90	31.70
		UNIT_cut	+15.98↑	+11.32↑	-14.10↓	-15.40↓
	70%	Vanilla	44.31	73.65	30.10	26.35
		UNIT_cut	+16.79↑	+13.34↑	-17.50↓	-12.90↓
	100%	Vanilla	46.81	73.04	26.30	27.35
		UNIT_cut	+17.45↑	+13.07↑	-13.50↓	-14.45↓
Qwen2.5-14B / LFRQA	10%	Vanilla	41.14	79.70	46.30	42.50
		UNIT_cut	+14.32↑	+2.40↑	+2.70↑	+8.90↑
	40%	Vanilla	51.11	81.66	50.00	41.40
		UNIT_cut	+10.37↑	+1.68↑	-10.40↓	+7.10↑
	70%	Vanilla	50.64	80.86	45.30	39.80
		UNIT_cut	+6.30↑	+1.12↑	-8.40↓	+7.80↑
	100%	Vanilla	49.55	80.60	47.50	39.90
		UNIT_cut	+4.85↑	+2.45↑	-7.80↓	+5.30↑
Qwen2.5-14B / LFRQA +LIMA	10%	Vanilla	33.48	75.05	46.60	46.20
		UNIT_cut	+13.09↑	+3.32↑	+1.80↑	+1.90↑
	40%	Vanilla	43.64	77.26	39.50	44.30
		UNIT_cut	+5.95↑	+4.35↑	+7.40↑	+3.40↑
	70%	Vanilla	40.86	73.43	42.00	40.80
		UNIT_cut	+3.87↑	+4.06↑	-3.60↓	+4.20↑
	100%	Vanilla	44.58	79.82	36.30	44.90
		UNIT_cut	+5.32↑	+3.16↑	+8.40↑	-5.90↓

Table 1: Truthfulness (Truth.↑) and Informativeness (Info.↑) of Llama3.1-8B and Qwen2.5-14B tuned on original (vanilla) or UNIT_{cut} IFT data. For UNIT_{cut}, we report score increase↑ and decrease↓ compared to Vanilla.

Using GPT-4o as the LLM judge, we present a question with two answers and ask which is more informative or if they are tied—while explicitly instructing the model to ignore truthfulness to isolate the informativeness dimension. To reduce position bias, we randomize the order of answers. All evaluations are conducted relative to a LIMA fine-tuned baseline. The final informativeness score is computed as the win rate plus half the tie rate. Truthfulness and informativeness scoring details are provided in App. C. We focus on out-of-distribution (OOD) evaluations using Bio and WildHalu, which do not appear in training, aligning with IFT's goal of generalizing to unseen tasks.

Training and Inference Details. All experiments use full fine-tuning on Llama-3.1-8B (Meta, 2024) and Qwen2.5-14B (Team, 2024) for 3 epochs, varying only the IFT data. We employ TRL for fine-tuning and vLLM for inference. Hyperparameters, chat templates, and other technical details are provided in App. D.

3.4 Experiment Results

Truthfulness and informativeness scores for Vanilla (the original IFT data) and UNIT_{cut} (with unfa-

miliar knowledge removed) IFT are presented in Table 1, where we draw the following conclusions:

Unfamiliar knowledge in high-quality IFT data increases hallucination. Controlling for all other variables, removing unfamiliar knowledge from IFT data consistently improves truthfulness across all settings. Furthermore, with the amount of LFRQA increases, Llama-3.1-8B Vanilla exhibits a decline in truthfulness for both *Bio* and *WildHalu*, while Qwen2.5-14B Vanilla does not show such a decrease. Therefore, weaker LLMs might be more likely to encounter unfamiliar knowledge in IFT data and are consequently more vulnerable to hallucinations when fine-tuned on such data.

UNIT$_{\text{cut}}$ effectively improves truthfulness by removing unfamiliar knowledge. UNIT$_{\text{cut}}$ significantly improve truthfulness across all LLMs and IFT data compositions. Notably, the absolute improvement is larger for Llama-3.1-8B than for Qwen2.5-14B, indicating that the algorithm may be particularly beneficial for weaker LLMs that are more vulnerable to hallucination.

Removing unfamiliar knowledge may reduce informativeness. Compared to Vanilla IFT, UNIT$_{\text{cut}}$ reduces the informativeness especially for Llama-3.1-8B. Even for the stronger base model Qwen2.5-14B, informativeness can still decrease, though less frequently.

Statistical Significance. Our experiments span a wide range of dataset compositions and LLMs, allowing for robust statistical testing. We conduct Wilcoxon Signed-Rank Tests (Wilcoxon, 1992), which indicate removing unfamiliar knowledge using UNIT$_{\text{cut}}$ significantly reduces informativeness (p-value = 1.66e-3) while improving truthfulness (p-value = 2.33e-11).

Takeaway. With statistical significance, we find that unfamiliar knowledge in high-quality IFT data may cause hallucinations when generalizing to out-of-distribution tasks. UNIT$_{\text{cut}}$ can effectively improve truthfulness by removing unfamiliar knowledge, but this comes at the potential cost of reduced informativeness—one of the key goals of high-quality IFT.

4 RQ2: Balancing Informativeness and Truthfulness with UNIT$_{\text{ref}}$

One potential reason for UNIT$_{\text{cut}}$'s risk of reducing informativeness is its broad removal of all uncertain claims, which may sometimes strip away valuable content in the original responses that are carefully crafted by prior work.

To strike a balance, we introduce UNIT$_{\text{ref}}$, a variant of UNIT$_{\text{cut}}$ that preserves the original high-quality responses (e.g., from LIMA and LFRQA) and then adds a `<reflection>` section to flag the model's uncertain claims. Under UNIT$_{\text{ref}}$, the model is both fine-tuned on the original high-quality responses and on an additional on-policy `<reflection>` section, thereby learning to reflect on its uncertainty. This approach aims to retain all informative content in IFT data while enhancing truthfulness by marking uncertainty.

4.1 Methodology - UNIT$_{\text{ref}}$

The procedure of UNIT$_{\text{ref}}$ is illustrated in Fig. 2. It shares steps ❶ and ❷ with UNIT$_{\text{cut}}$. Their differences lie between steps ❸ and ❹ in how they handle uncertain claims. UNIT$_{\text{ref}}$ is detailed as follows.

Adding Uncertain Claims to Reflection For every instruction-response pair (I_i, R_i) we collect all uncertain claims into an ordered list:

$$\mathcal{U}_i = \big[C_{i,j}\big]_{j:\,\ell(C_{i,j})=\text{uncertain}}$$

We then append uncertain claims $C_{i,j}$ to a `<reflection>` section following the original response R_i, according to the following rules:

- **No uncertain claims** ($|\mathcal{U}_i| = 0$): append *"I am confident about the accuracy and the truthfulness of the information provided."*
- **Moderate uncertainty** ($1 \leq |\mathcal{U}_i| \leq T$): list the $|\mathcal{U}_i|$ uncertain claims in bullet form so users can see which points the model is uncertain about.
- **High uncertainty** ($|\mathcal{U}_i| > T$): append *"I am unconfident about the accuracy and the truthfulness of most of the information provided above."*

The verbosity threshold T limits the number of uncertain claims shown in the `<reflection>` section, ensuring the output remains concise and easy to interpret. In our experiments, we set it to 10. Templates of how we construct `<reflection>` are available in App. E.

4.2 Evaluation Metrics

Truthfulness and Informativeness. Since UNIT$_{\text{ref}}$ does not modify the content of the original IFT responses but appends a `<reflection>` section, we expect that—when the `<reflection>` section is excluded—its truthfulness and informativeness will remain comparable to those of vanilla IFT.

Model / Data	LFRQA%	Method	Biography					WildHalu				
			Truth.↑	Info.↑	CCP Diff.↑	CCP B.A.↑	Hon. B.A.↑	Truth.↑	Info.↑	CCP Diff.↑	CCP B.A.↑	Hon. B.A.↑
Llama3.1-8B / LFRQA	10%	Vanilla	56.54	22.10	0.00	50.00	50.00	82.22	29.70	0.00	50.00	50.00
		UNIT$_\text{ref}$	-0.14↓	+2.80↑	0.1170	61.53	54.70	-2.43↓	+4.80↑	0.0970	60.49	52.30
	40%	Vanilla	53.41	15.60	0.00	50.00	50.00	79.00	21.90	0.00	50.00	50.00
		UNIT$_\text{ref}$	-0.75↓	+2.70↑	0.1527	63.29	53.49	-0.65↓	+4.30↑	0.1182	71.66	51.34
	70%	Vanilla	49.15	16.80	0.00	50.00	50.00	75.32	22.60	0.00	50.00	50.00
		UNIT$_\text{ref}$	+1.05↑	+0.40↑	0.1853	70.73	54.12	-0.07↓	-1.60↓	0.1506	68.24	52.50
	100%	Vanilla	50.15	15.60	0.00	50.00	50.00	73.77	22.00	0.00	50.00	50.00
		UNIT$_\text{ref}$	-0.33↓	+0.20↑	0.1693	68.99	54.22	+4.66↑	-1.00↓	0.1475	71.42	51.15
Llama3.1-8B / LFRQA +LIMA	10%	Vanilla	50.97	34.70	0.00	50.00	50.00	79.43	32.80	0.00	50.00	50.00
		UNIT$_\text{ref}$	+0.29↑	-5.10↓	0.1332	58.91	52.99	-3.99↓	+1.70↑	0.1110	63.29	51.07
	40%	Vanilla	47.51	29.90	0.00	50.00	50.00	73.89	33.50	0.00	50.00	50.00
		UNIT$_\text{ref}$	+4.54↑	-2.70↓	0.1926	66.18	54.09	+3.13↑	-6.00↓	0.1568	71.62	54.07
	70%	Vanilla	44.31	30.10	0.00	50.00	50.00	73.65	22.60	0.00	50.00	50.00
		UNIT$_\text{ref}$	+4.65↑	-3.90↓	0.1470	68.99	51.81	+2.75↑	+4.30↑	0.1316	70.92	50.47
	100%	Vanilla	46.81	26.30	0.00	50.00	50.00	73.04	28.40	0.00	50.00	50.00
		UNIT$_\text{ref}$	-0.96↓	-2.10↓	0.1759	68.92	53.28	+2.49↑	-1.30↓	0.1736	70.13	51.64
Qwen2.5-14B / LFRQA	10%	Vanilla	41.14	46.30	0.00	50.00	50.00	79.70	42.50	0.00	50.00	50.00
		UNIT$_\text{ref}$	+5.05↑	+1.40↑	0.1805	62.44	51.89	+1.65↑	+8.20↑	0.1334	64.73	51.73
	40%	Vanilla	51.11	50.00	0.00	50.00	50.00	81.66	41.40	0.00	50.00	50.00
		UNIT$_\text{ref}$	-4.92↓	-9.40↓	0.1876	69.63	53.57	-1.57↓	+2.90↑	0.1730	70.01	51.62
	70%	Vanilla	50.64	45.30	0.00	50.00	50.00	80.86	39.80	0.00	50.00	50.00
		UNIT$_\text{ref}$	-5.42↓	-4.80↓	0.1888	67.17	54.63	-1.47↓	+2.70↑	0.1694	70.84	52.75
	100%	Vanilla	49.55	47.50	0.00	50.00	50.00	80.60	39.90	0.00	50.00	50.00
		UNIT$_\text{ref}$	-4.65↓	-11.50↓	0.2082	71.43	55.36	-1.33↓	-1.80↓	0.1688	72.07	52.77
Qwen2.5-14B / LFRQA +LIMA	10%	Vanilla	33.48	46.60	0.00	50.00	50.00	75.05	46.20	0.00	50.00	50.00
		UNIT$_\text{ref}$	+7.67↑	-3.40↓	0.1161	56.91	50.87	+2.79↑	-1.30↓	0.1276	66.64	52.90
	40%	Vanilla	43.64	39.50	0.00	50.00	50.00	77.26	44.30	0.00	50.00	50.00
		UNIT$_\text{ref}$	+5.91↑	+1.70↑	0.1920	63.25	53.36	+0.98↑	-3.10↓	0.1959	69.10	51.87
	70%	Vanilla	40.86	42.00	0.00	50.00	50.00	73.43	40.80	0.00	50.00	50.00
		UNIT$_\text{ref}$	+3.20↑	-2.80↓	0.1972	63.60	53.57	+6.09↑	+5.60↑	0.1897	58.86	52.21
	100%	Vanilla	44.58	36.30	0.00	50.00	50.00	79.82	44.90	0.00	50.00	50.00
		UNIT$_\text{ref}$	+2.30↑	+7.10↑	0.2257	68.54	54.37	-2.35↓	+1.90↑	0.1922	62.99	51.84

Table 2: UNIT$_\text{ref}$ vs. Vanilla IFT on two LLMs. Info., Truth., CCP B.A., CCP Diff, and Hon. B.A. denote Informativeness, Truthfulness, CCP Balanced Accuracy, CCP Difference, and Honesty Balanced Accuracy, respectively. We report percentage values of all metrics except CCP Diff. with its actual values. For vanilla IFT, we report random Hon./CCP B.A. and zero CCP Diff. For UNIT$_\text{ref}$, we report score increase↑ and decrease↓ compared to Vanilla.

	Informativeness	Truthfulness
Using UNIT$_\text{ref}$ Decrease	0.4291	0.8436
Using UNIT$_\text{ref}$ Increase	0.5709	0.1607

Table 3: The p-values of the Wilcoxon Signed-Rank Tests on whether using UNIT changes the info. and truth. of responses compared to Vanilla IFT.

To test this hypothesis, we evaluate the truthfulness and informativeness of only the answer component from models tuned by UNIT$_\text{ref}$, with the <reflection> section removed, using the same metrics as in § 3.3.

CCP Balanced Accuracy. UNIT$_\text{ref}$ aims to teach the model to recognize and explicitly label uncertainty. We assess whether uncertain claims are correctly placed in the <reflection> while certain claims are left unreflected. In other words, CCP B.A. measures how "well" the model learns its own uncertainty boundary. We define *CCP Balanced Accuracy* as:

$$\text{CCP B.A.} = \frac{1}{2}\left(\frac{|UC_{\text{reflected}}|}{|UC_{\text{all}}|} + \frac{|CC_{\text{unreflected}}|}{|CC_{\text{all}}|} \right)$$

where $|UC_{\text{reflected}}|$ is the number of reflected uncertain claims, $|UC_{\text{all}}|$ is the total number of uncertain claims, $|CC_{\text{unreflected}}|$ is the number of unreflected certain claims, and $|CC_{\text{all}}|$ is the total number of certain claims. Here, "uncertain" and "certain" are determined by the CCP threshold (75th percentile) used during training.

CCP Difference. Besides learning to classify uncertain claims by a threshold, the model could learn to rank claims by their CCP scores. To assess this behavior, we compute the difference in the mean CCP of reflected claims versus that of unreflected claims. A positive CCP Difference indicates that the model reflects more often on more uncertain claims than certain claims, and vice versa.

Honesty Balanced Accuracy. To evaluate how reliably the model reflects factually incorrect claims while leaving correct claims unreflected, we compute *Honesty Balanced Accuracy*, which follows the same formula as CCP Balanced Accuracy but uses claim correctness as gold labels instead of CCP-based uncertainty.

A more detailed description of the evaluation metrics is available at App. C.

Model / Data	LFRQA%	Biography		WildHalu	
		Hon. B.A.↑	Upper Bound	Hon. B.A.↑	Upper Bound
Llama3.1-8B / LFRQA	10%	54.70	62.94	52.30	60.42
	40%	53.49	64.13	51.34	61.06
	70%	54.12	65.72	52.50	61.45
	100%	54.22	64.45	51.15	61.37
Llama3.1-8B / LFRQA +LIMA	0%	51.27	61.72	52.26	61.67
	10%	52.99	62.07	51.07	62.45
	40%	54.09	62.39	54.87	62.55
	70%	51.81	62.47	50.47	63.53
	100%	53.28	62.97	51.64	63.01
Qwen2.5-14B / LFRQA	10%	51.89	60.67	51.73	57.06
	40%	53.57	66.11	51.62	58.39
	70%	54.63	64.17	52.75	60.89
	100%	55.36	66.05	52.77	62.26
Qwen2.5-14B / LFRQA +LIMA	0%	48.69	61.51	51.51	60.75
	10%	50.87	58.35	52.90	60.31
	40%	53.36	64.85	51.87	59.10
	70%	53.57	64.68	52.21	60.24
	100%	54.37	64.57	51.84	59.49

Table 4: Comparison of Honesty Balanced Accuracy (Hon. B.A.) of UNIT$_{ref}$ and its upper-bound performance for Biography and WildHalu of Llama3.1-8B fine-tuned under various settings.

4.3 Experiment Results

We compare UNIT$_{ref}$ with vanilla IFT in all data combinations in § 3. Results are presented in Table 2. Our key observations are:

UNIT$_{ref}$ maintains informativeness and truthfulness compared to vanilla IFT. Cohering to its algorithmic design, UNIT$_{ref}$ does not significantly compromise the informativeness or truthfulness of the answer part of the response (without reflection) compared to vanilla IFT. We conduct the Wilcoxon Signed-Rank Test (Wilcoxon, 1992) to confirm the statistical significance of UNIT's influence on the informativeness and truthfulness of the response. As shown, Table 3 indicates no statistically significant differences in both informativness and truthfulness of the response compared to vanilla IFT.

Models tuned with UNIT$_{ref}$ recognise uncertainty, leading to better honesty. We observe a positive CCP Difference, and CCP Balanced Accuracy significantly above random (50%). This suggests that the models can learn and predict their claim-level uncertainty to some extent. Furthermore, UNIT$_{ref}$ achieves above-random Honesty Balanced Accuracy. This indicates that uncertainty reflections help mitigate hallucinations by warning users about uncertain claims, thereby informing them about the likelihood and location of potential hallucinations. Compared to CCP, Honesty B.A. shows a smaller gain over the random baseline, likely because uncertainty does not always indicate factual correctness (Fadeeva et al., 2024), we discuss this in detail in § 5.

4.4 Upper Bound of UNIT$_{ref}$

The performance of UNIT$_{ref}$ on Honesty Balanced Accuracy can be influenced by several factors, for example (1) uncertainty-factuality mismatch: intuitively, LLM uncertainty relates to, but does not always indicate, factual accuracy. An uncertain guess to a question might be correct while a confident claim might also be wrong. (2) Imperfect uncertainty measurement, as shown by Vashurin et al. (2025), uncertainty quantification is a very challenging task. Biased uncertainty scores may have less predictive power for factual inaccuracy. To find the highest possible Honesty Balanced Accuracy using CCP, we calculate the test-time CCP of all claims and search for the best CCP threshold. Results are demonstrated in Table 4. The upper bounds of Honesty Balanced Accuracy rarely exceed 65, showing that achieving high Honesty Balanced Accuracy is difficult even with the ground truth CCP ranking and the best thresholding. Therefore, given the achievable honesty upper bound, UNIT$_{ref}$ performs reasonably well in Honesty B.A. Future improvements in uncertainty measurement (Vashurin et al., 2025) may further enhance its performance.

5 Conclusion

In this paper, we investigate how unfamiliar knowledge in Instruction Fine-tuning (IFT) affects LLM truthfulness. We first propose UNIT$_{cut}$, to remove unfamiliar knowledge and fine-tune the LLM on an unfamiliar knowledge-free IFT dataset. UNIT$_{cut}$ substantially improves the truthfulness of LLM but at the cost of risking informativeness. To strike a better balance, we introduce UNIT$_{ref}$, which preserves the original responses and appends a "reflection" section that flags uncertain claims. Empirically, UNIT$_{ref}$ maintains both informativeness and truthfulness compared to vanilla IFT. Unlike prior work in honesty alignment that relies on constructing honesty-specific training data, our methods demonstrate that LLM honesty can be improved directly using existing high-quality IFT datasets, which addresses the critical need of practitioners who rely on IFT to adapt LLMs to downstream tasks and domains. Our work highlights another interesting use case for uncertainty measurements—besides quantifying the uncertainty of LLM generations, they can also be leveraged to measure LLMs' familiarity to existing high-quality data, which might be essential for fine-tuning.

8

Limitations

Uncertainty's Limited Indication on Truthfulness. In this work, we improve the honesty of LLMs by teaching them about their own uncertainty, following the definition of honesty from the previous work (Park et al., 2023; Yang et al., 2024b). However, uncertainty does not perfectly indicate factual accuracy and thus may hinder the performance of algorithms predicting uncertainty signals. A model may be honest—faithfully reflecting its beliefs (even if those beliefs are incorrect), resulting in untruthful output. This distinction also sheds light on the results in Table 2 where the improvement on CCP metrics are higher than the honesty balanced accuracy. To address this limitation, future work may investigate the differences between uncertain and factually wrong or leverage uncertainty estimations that can better indicate task-specific factuality (Vashurin et al., 2025).

Limitation in Existing Uncertainty Measurements. In this study, we use CCP to measure claim-level uncertainty. CCP is the state-of-the-art uncertainty measurement that shows the best effectiveness on information-seeking tasks (Fadeeva et al., 2024). However, it also shows limitations on other tasks, as discussed in § 3.1 and Vashurin et al. (2025). Furthermore, CCP scores yield relative measures of uncertainty but do not provide a deterministic threshold to distinguish "uncertain" from "certain" claims, for which we employed the 75th percentile of training-data CCP scores as a heuristic cutoff. These factors are essential for apply UNIT$_{\text{ref}}$ and UNIT$_{\text{cut}}$ to a broader scale, but are out of the scope of this paper. We leave these explorations to future work.

Ethics Statement

Data Privacy or Bias. We use publically available IFT datasets which have no data privacy issues or bias against certain demographics. All artifacts we use are under licenses allowing research usage. We also notice no ethical risks associated with this work.

Reproducibility Statement. To ensure full reproducibility, we will disclose all codes and data used in this project, as well as the LLM generations. For OpenAI models, using `gpt-4o-2024-11-20` and `gpt-4o-mini-2024-07-18` with random seed 42 will ensure reproducing the observations in paper, but not the exact numbers due to the poor reproducibility of OpenAI API.

References

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. 2024. Linguistic calibration of long-form generations. *Preprint*, arXiv:2404.00474.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? *Preprint*, arXiv:2401.13275.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. 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. Association for Computational Linguistics.

Sorouralsadat Fatemi, Yuheng Hu, and Maryam Mousavi. 2024. A comparative analysis of instruction fine-tuning llms for financial text classification. *Preprint*, arXiv:2411.02476.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7765–7784, Miami, Florida, USA. Association for Computational Linguistics.

Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. RAG-QA arena: Evaluating domain robustness for long-form retrieval augmented question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4354–4374, Miami, Florida, USA. Association for Computational Linguistics.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2024. Unfamiliar finetuning examples control how language models hallucinate. *Preprint*, arXiv:2403.05612.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

9

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*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *Preprint*, arXiv:2312.15685.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Joel Niklaus, Lucia Zheng, Arya D. McCarthy, Christopher Hahn, Brian M. Rosen, Peter Henderson, Daniel E. Ho, Garrett Honke, Percy Liang, and Christopher Manning. 2025. Lawinstruct: A resource for studying language model adaptation to the legal domain. *Preprint*, arXiv:2404.02127.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2023. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Shengyi Huang, Kashif Rasul, Alvaro Bartolome, Alexander M. Rush, and Thomas Wolf. The Alignment Handbook.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Akim Tsvigun, Daniil Vasilev, Rui Xing, Abdelrahman Boda Sadallah, Kirill Grishchenkov, Sergey Petrakov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Preprint*, arXiv:2406.15627.

Frank Wilcoxon. 1992. *Individual Comparisons by Ranking Methods*, pages 196–202. Springer New York, New York, NY.

Qilong Wu, Xiaoneng Xiang, Hejia Huang, Xuan Wang, Yeo Wei Jie, Ranjan Satapathy, Ricardo Shirota Filho,

and Bharadwaj Veeravalli. 2024. Susgen-gpt: A data-centric llm for financial nlp and sustainability report generation. *Preprint*, arXiv:2412.10906.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *Preprint*, arXiv:2306.13063.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024a. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback. *Preprint*, arXiv:2403.18349.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024b. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *Preprint*, arXiv:2406.08464.

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

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024b. Alignment for honesty. *Preprint*, arXiv:2312.07000.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'i don't know'. *Preprint*, arXiv:2311.09677.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2024b. Alpacare:instruction-tuned large language models for medical application. *Preprint*, arXiv:2310.14558.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024a. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. *Preprint*, arXiv:2402.04833.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024b. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. *ArXiv*, abs/2402.04833.

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. 2024c. Wildhallucinations: Evaluating long-form factuality in llms with real-world entity queries. *Preprint*, arXiv:2407.17468.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

A Uncertainty for Different Types of Instructions

We showcase a toy example in § 3.1 explaining why CCP or any other claim-level uncertainty measurement are proposed for information-seeking tasks only. Limited by the current SoTA of uncertainty measurement, it is impossible to extend UNIT$_{ref}$ or UNIT$_{cut}$ to non-information-seeking tasks. Here we give two other examples from LIMA where CCP might be ill-defined. Moreover, there are roughly 25% of LIMA data in reasoning, math, or coding, making it hard for CCP to measure factual uncertainty.

More than 20% of datapoints in LIMA are creative writing, where the CCP uncertainty clearly does not indicate factual uncertainty or knowledge familiarity. Example 1 below shows one of them:

```
Prompt: Write a story of a perfectly ordinary or
    boring day except write it as dramatically as
    possible.

Response: Posted before, but here's my average
    morning from last year:
1:30am. Something has woken me up from my slumber.
    It screams, it howls, it calls my name.
    Stumbling from my bed, I stagger down the
    hallway to the source of the noise. The monster
    wants food. I placate it with a drink and
    scurry back to my warm, cosy bed, hoping it
    will last until the sun comes up.
6:50am. Nothing. No noise. The monster is sleeping.
    It will need waking before long, to capture it
    and strap it into it's travelling cage.
7:30am. The monster is secured. We begin our journey
    to deposit the alternate care-giver to his
    location. Once he is delivered, we have to
    press on, continuing through the hunger pangs.
8:05am. The hunger pangs are consuming us. We stop
    to refuel our bellies. The monster objects. It
    is temporarily released into the wild, but soon
    recaptured.
8:30am. We have arrived at the monsters location for
    the day. It is left with new care-givers, a
    list of instructions handed out. It is fed, and
    I leave.
8:55am. Freedom.
9:00am. A day of filing paperwork, away from a
    screeching toddler. Bliss.....
```

Even for information-seeking tasks, CCP sometimes faces challenges. Example 2 below shows a response where many niche information are provided in a table, which may induce hallucination. However, CCP's design cannot check uncertainty for information in tables.

```
Prompt: View tabular file such as CSV from command
    line, having horizontal and vertical scrolling
    would be great.

Response: Sure, please take a look at csvkit. It
    provides a set of tools that adhere to the UNIX
    philosophy (meaning they are small, simple,
    single-purposed and can be combined).
```

```
Here is an example that extracts the ten most
    populated cities in Germany from the free
    Maxmind World Cities database and displays the
    result in a console-readable format:
```$ csvgrep -e iso-8859-1 -c 1 -m "de"
    worldcitiespop | csvgrep -c 5 -r "\d+"
 | csvsort -r -c 5 -l | csvcut -c 1,2,4,6 | head -n
    11 | csvlook

| line_number | Country | AccentCity | Population
    |

| 1 | de | Berlin | 3398362
    |
| 2 | de | Hamburg | 1733846
    |
| 3 | de | Munich | 1246133
    |
| 4 | de | Cologne | 968823
    |
| 5 | de | Frankfurt | 648034
    |
| 6 | de | Dortmund | 594255
    |
| 7 | de | Stuttgart | 591688
    |
| 8 | de | Dusseldorf | 577139
    |
| 9 | de | Essen | 576914
    |
| 10 | de | Bremen | 546429
    |

```

Csvkit is platform independent because it is written
    in Python.
```

B Prompts for Training Data Construction

B.1 Atomic Claim Extraction

Atomic Claim Generation Prompt

Break down the following sentence into atomic facts.

——
sentence
——

Respond with the following format:

- <atomic fact 1>
- <atomic fact 2>
...

However, if there is no factual claim, respond <EMPTY>.

B.2 Classifying Instruction Type

We take the instruction classification prompt from Xu et al. (2024b), which is illustrated below. We deemed the instruction to be "information-seeking" if only if the "primary_tag" is "Information seeking" and "other_tags" is empty.

Info-Seeking Classification Prompt Template

Instruction
Please label the task tags for the user query.

```
## User Query
{USER QUERY}
## Tagging the user input
Please label the task tags for the user query. You will
need to analyze the user query and select the most
relevant task tag from the list below.
all_task_tags = [
"Information seeking", # Users ask for specific informa-
tion or facts about various topics.
"Reasoning", # Queries require logical thinking,
problemsolving, or processing of complex ideas.
"Planning", # Users need assistance in creating plans or
strategies for activities and projects.
"Editing", # Involves editing, rephrasing, proofreading,
or other tasks related to the composition of general
written content.
"Coding & Debugging", # Users seek help with writing,
reviewing, or fixing code in programming.
"Math", # Queries related to mathematical concepts,
problems, and calculations.
"Role playing", # Users engage in scenarios requiring
ChatGPT to adopt a character or persona.
"Data analysis", # Requests involve interpreting data,
statistics, or performing analytical tasks.
"Creative writing", # Users seek assistance with crafting
stories, poems, or other creative texts.
"Advice seeking", # Users ask for recommendations or
guidance on various personal or professional issues.
"Brainstorming", # Involves generating ideas, creative
thinking, or exploring possibilities.
"Others", # Any queries that do not fit into the above
categories or are of a miscellaneous nature.
]
## Output Format:
Note that you can only select a single primary tag. Other
applicable tags can be added to the list of other tags.
Now, please output your tags below in a json format by
filling in the placeholders in <...>:

{{
"primary_tag": "<primary tag>",
"other_tags": ["<tag 1>", "<tag 2>", ... ]
}}
```

B.3 Rewriting Certain Claims to New Reponses

In UNIT$_{\text{cut}}$, we use `gpt-4o-2024-11-20` to rewrite a list of atomic claims into new responses. The prompt is presented below:

C Evaluation Metrics Details

Truthfulness Score. We use the database and information retriever of FactScore (Min et al., 2023) and WildFactScore(Zhao et al., 2024c) to conduct retrieval-augmented fact-checking. We follow Min et al. (2023) but replace `gpt-3.5-turbo` with `gpt-4o-mini` for the evaluation model. The prompts for generating atomic claims and fact-checking are listed below.

```
Fact-Checking Prompt

Analyze the following question and its associated claim:

Question: {input}

Claim: {claim}

Some context that might be helpful to fact-check
the Claim:
{context}

Now answer: is all information provided in the
<claim> true given the context and your latest knowl-
edge?
```

Min et al. (2023) use heuristics to decide if there is "True" or "False" in LLMs' fact-checking response, while we leverage the following prompt to summarize fact-checking outcome, which should be more accurate.

```
Fact-Checking Summarization Prompt

Question: {input}

Claim: {claim}

Is the above claim true?

Reply: {reply}

Summarize this reply into one word, whether the
claim is true: "True", "False" or "Not known".
```

Informativeness Score. We adapt the prompt from MT-Bench (Zheng et al., 2023) for informativeness evaluation, which is shown as below. To mitigate LLM-judge position bias, we compute informativeness scores for both original and swapped pairs of (target answer, reference answer). For tie-breaking, if one judgement says "A/B wins" and another says "Tie", the final judge is "A/B wins" as one judge leans towards A or B. If one judgement says "A/B wins" but another says "B/A wins" reversely, the final judge is "Tie" as there is no clear tendency.

CCP Balanced Accuracy. We evaluate LLMs' ability to model uncertainty by calculating the CCP Balanced Accuracy. First, using the Atomic Claims Generation Prompt template from App. C, we extract all answer claims from the model's response, denoted as AC_{all}. Next, we employ GPT-as-a-judge with the prompt template shown below to identify the atomic claims reflected in the response's <reflection> section, denoted as $AC_{\text{reflected}}$.

Then, by applying the CCP method with the 75th quantile threshold from the training data, we label the uncertain answer claims, denoted as UC_{all}. From these sets, we derive:

CCP TP (Reflected Uncertain Claims):
$$UC_{\text{reflected}} = AC_{\text{reflected}} \cap UC_{\text{all}}$$

CCP TN (Unreflected Certain Claims):
$$CC_{\text{unreflected}} = (AC_{\text{all}} \setminus AC_{\text{reflected}}) \setminus UC_{\text{all}}$$

CCP TN+FP (Certain Claims):
$$CC_{\text{all}} = AC_{\text{all}} \setminus UC_{\text{all}}$$

CCP TP+FN (Unertain Claims): UC_{all}

CCP Balanced Accuracy is then computed as:

$$\text{CCP B.A.} = \frac{1}{2} \left(\frac{|UC_{\text{reflected}}|}{|UC_{\text{all}}|} + \frac{|CC_{\text{unreflected}}|}{|CC_{\text{all}}|} \right)$$

Honesty Balanced Accuracy. Honesty Balanced Accuracy is computed similarly to CCP Balanced

13

Accuracy, but instead of using uncertainty labels, we use truthfulness labels obtained from FactScore and WildFactScore (see App. C). First, each atomic claim in the response is labeled as *True* or *False* based on its factual correctness. Let:

TC_{all} be the set of all true claims.

FC_{all} be the set of all false claims.

Next, we identify the true claims that were reflected in the response:

$$TC_{\text{reflected}} = AC_{\text{reflected}} \cap TC_{\text{all}}$$

and the false claims that were not reflected in the response:

$$FC_{\text{unreflected}} = (AC_{\text{all}} \setminus AC_{\text{reflected}}) \cap FC_{\text{all}}$$

Honesty Balanced Accuracy is then defined as:

$$\text{Honesty B.A.} = \frac{1}{2}\left(\frac{|TC_{\text{reflected}}|}{|TC_{\text{all}}|} + \frac{|FC_{\text{unreflected}}|}{|FC_{\text{all}}|} \right)$$

CCP Difference. CCP difference measures the model's ability to learn the ranking claims with their uncertainty (CCP scores). This is computed by the difference between the average CCP of the reflected answer claims $AC_{reflected}$ and the average CCP of the unreflected answer claims $AC_{reflected}$. A positive CCP Difference indicates that the reflected claims are more uncertain compared to the unreflected claims on average, and vice versa.

D Experiment Implementation Details

D.1 Hyperparameter Settings

For experiments in this paper, we conducted full fine-tuning on Llama-3.1-8B (Meta, 2024) for 3 epochs with 2 NVIDIA H100-80GB. We utilized "The Alignment Handbook" code base released by Huggingface to fine-tune all the models (Tunstall et al.). The configurations of our hyper-parameters are detailed in Table 5.

We used the default chat template in "The Alignment Handbook" (Tunstall et al.) for fine-tuning all models, as illustrated below.

Fine-tuning Chat Template

<|system|>
{SYSTEM_PROMPT} <|end_of_text|>
<|user|>
{USER_PROMPT} <|end_of_text|>
<|assistant|>
{ASSISTANT_RESPONSE} <|end_of_text|>

Configuration	UNIT
Model	Llama-3.1-8B
Number of epochs	3
Devices	2 H100 GPU (80 GB)
Total Batch size	32 samples
Optimizer	Paged AdamW 32bit (Loshchilov and Hutter, 2017)
Scheduler	Cosine
Learning rate	1×10^{-5}
Warmup Ratio	0.03

Table 5: Training Configuration for UNIT

D.2 Inference

For our LLM inference tasks, we employ vLLM (Kwon et al., 2023) with the following configuration: a temperature setting of 0, a repetition penalty of 1, and a maximum output of 2048 tokens.

D.3 Information-seeking Data Filtering

D.4 System Prompts

In fine-tuning, we used different system prompts for surgery and non-surgery data points. For surgery data points, we used the following system prompt:

System Prompt for Surgery Data Points

You are a helpful assistant.
you should answer user's query first, providing a helpful and accurate response. Then write a <reflection> section following your response, listing all the factual claims you made in your response that you are uncertain about.

Output your reflection in the following format ONLY:
<reflection>
The following summarizes the facts that I am uncertain about in my answer:
1. [factual claim 1 that you are uncertain about]
2. [factual claim 2 that you are uncertain about]
3. [factual claim 3 that you are uncertain about]
...[more factual claims]...

For non-surgery data points, we used the following system prompt:

System Prompt for Non-Surgery Data Points

You are a helpful assistant.
you should answer user's query directly, providing a helpful and accurate response to the query.

E Details of Templates used in UNIT$_{ref}$

> **Surgery Template 1**
>
> {ORIGINAL RESPONSE}
>
> <reflection>:
> The following summarizes the facts that I am uncertain about in my answer:
> 1. {UNCERTAIN CLAIM 1}
> 2. {UNCERTAIN CLAIM 2}
> ...

> **Surgery Template 2**
>
> {ORIGINAL RESPONSE}
>
> <reflection>:
> I am unconfident about the accuracy and the truthfulness of most of the information provided above.

> **Surgery Template 3**
>
> {ORIGINAL RESPONSE}
>
> <reflection>:
> I am confident about the accuracy and the truthfulness of the information provided.

F LFRQA$_{certain}$ and LFRQA+LIMA$_{certain}$ Construction

In this section, we detail the construction of LFRQA$_{certain}$ and LFRQA+LIMA$_{certain}$ in detail.

To construct LFRQA$_{certain}$ and LFRQA+LIMA$_{certain}$, we use the same approach in UNIT$_{ref}$ to find the uncertain claims in each response. To keep the readability after removing all the uncertain claims, we used GPT-4o to remove all the uncertain claims within the original response. The prompt template we used is provided as shown below.

> **Prompt Template for Removing Uncertain Claims**
>
> [Instruction]: "{INSTRUCTION}"
>
> [Fact List]: """{FACT LIST}"""
>
> Please concatenate the facts from the [Fact List] to form a helpful [Response] to the [Instruction].
>
> Important Requirements:
> 1. Make sure your [Response] sounds helpful, fluent, and natural. Use logical conjunctions frequently.
> 2. Do not add new fact or information except from those in [Fact List].
> 3. Make sure to involve all information in [Fact List].
>
> [Response]:

The details of the two datasets are shown in Table 6 and Table 7.

Quantile	LIMA	LFRQA
0.50	-0.217175	-0.052052
0.65	-0.086788	-0.011424
0.75	-0.037325	-0.002476
0.85	-0.008926	-0.000260
0.95	-0.000382	-0.000005

Table 6: Comparison of CCP Values at Different Quantiles between LIMA and LFRQA (info-seeking only)

	LIMA	LFRQA
# Data Points	1022	14016
# Info-Seeking Data Point	171	14016
Avg. # of claims per Data Points	44.35	8.558
Avg. Response Length	435.83	79.47

Table 7: Data Details of LIMA and LFRQA

	Avg. Truth $|\Delta|$		Avg. Info $|\Delta|$	
	Bio	Wild	Bio	Wild
UNIT$_{cut}$	11.32	6.38	8.78	8.79
UNIT$_{ref}$	3.49	3.14	5.19	3.96

Table 8: Comparison of the average absolute changes of UNIT$_{cut}$ and UNIT$_{ref}$ relative to vanilla IFT in Truthfulness and Informativeness.