Conformal Linguistic Calibration: Trading-off between Factuality and Specificity

Zhengping Jiang
Johns Hopkins University
zjiang31@jhu.edu

Anqi Liu
Johns Hopkins University
aliu.cs@jhu.edu

Benjamin Van Durme
Johns Hopkins University
vandurme@jhu.edu

Abstract

Language model outputs are not always reliable, thus prompting research into how to adapt model responses based on uncertainty. Common approaches include: *abstention*, where models refrain from generating responses when uncertain; and *linguistic calibration*, where models hedge their statements using uncertainty quantifiers. However, abstention can withhold valuable information, while linguistically calibrated responses are often challenging to leverage in downstream tasks. We propose a unified view, Conformal Linguistic Calibration (CLC), which reinterprets linguistic calibration as *answer set prediction*. First we present a framework connecting abstention and linguistic calibration through the lens of linguistic pragmatics. We then describe an implementation of CLC that allows for controlling the level of imprecision in model responses. Results demonstrate our method produces calibrated outputs with conformal guarantees on factual accuracy. Further, our approach enables fine-tuning models to perform uncertainty-aware adaptive claim rewriting, offering a controllable balance between factuality and specificity.[1]

1 Introduction

Large language models (LLMs) can provide knowledge responses based on their comprehensive training sets [48, 59, 78], but these responses may not be trustworthy [41, 26], and can be stated in an over-confident tone [42]. One solution is to communicate a calibrated uncertainty measure to the user, so as to enable more informed decision-making [80]. This uncertainty information can be expressed in different ways, which we hereby classify into two groups. The first set of approaches marks a response explicitly with the model's level of confidence [66, 12, 65]. When such uncertainty expressions are conveyed through text generation and are adjusted to faithfully represent accuracy of the response, this is called *linguistic calibration* [42, 3, 6, 69]. For example, a model uncertain in its response may state: "*Possibly* the largest city in the US is Los Angeles" (Fig 1).[2] However, how to interpret linguistic uncertainty can be unclear and subjective. Also, the outer structure introduced by linguistic calibration hampers the downstream task / evaluation [34], as it is unclear what special treatment should be given to these uncertain assertions. Another set of approaches have the model generate prediction sets instead of a single response, such that with high probability some are acceptable [50, 32, 55]. A special case of this is when a model abstains from responding when it is not confident [58, 28, 39, 45, 9, 18]. E.g., "I don't know" is trivially equivalent to predicting the

[1] Code available at: https://github.com/zipJiang/CLC.

[2] Answer by population is *New York City*, by area is *Sitka*.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

https://doi.org/10.52202/085713-4338

set of all possible answers. However, this breaks the normal human-model interaction paradigm, and the set prediction is not directly actionable [11].

To address these limitations, we propose a novel unifying view, *Conformal Linguistic Calibration*: having the model express uncertainty levels by generating less specific claims until they are confident. We achieve this by reinterpreting uncertainty quantifiers as set prediction using recent advances [19] in possible world semantics [30, 29]. At a high level, being more confident in a claim means that a subject would only regard it as false in counterfactual worlds that are drastically different from the actual one. These gradually less similar counterfactual worlds form a series of nested sets corresponding to the subject's belief [36, 63]. From this perspective, linguistic calibration can be seen as a form of *soft abstention*, whereas traditional abstention [28, 44] corresponds to committing to the universal set of all possible worlds. As shown in Figure 1, this allows us to improve the flexibility and controllability over an abstention-based approach. Also, our formulation naturally supports conformal prediction techniques [67, 1], offering probabilistic guarantees on factuality while producing confident, clear claims that are easy for downstream use. In summary, we:

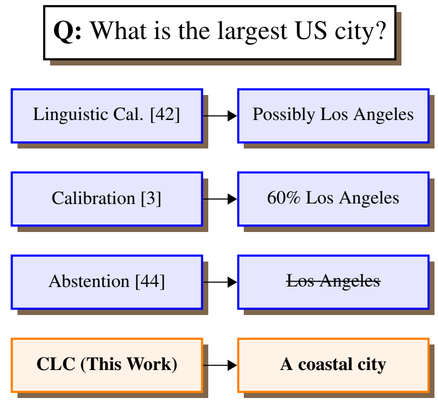

Figure 1: Distinct from previous approaches CLC allows generating a less precise but more confident statement.

1. Propose a new paradigm for model to communicate their uncertainty.
2. Provide an algorithmic instantiation of the process by probing semantic uncertainty [31], building nested subsets, and summarizing beliefs, which can be post-hoc calibrated through learn-then-test (LTT) [1].
3. Show that our approach enables a smooth trade-off between factuality and specificity [74]—for example, allowing a 7B model to outperform GPT-4o in factuality on a challenging QA dataset by answering less specifically [74].
4. Train an amortized, uncertainty-aware claim rewriter that explicitly trades off specificity for factuality following user instructions. We release both our rewriting model and dataset.

2 Related Work

LLM Factuality Evaluation Large language models enable rich long-form generation [5, 46], yet hallucinations remain persistent [41, 26, 24]. To monitor factual precision, evaluation pipelines typically decompose model outputs into atomic claims and verify each against trusted sources or retrieval-augmented evidence [43, 71, 72]. Subsequent work sharpens these pipelines by improving claim extraction and coverage [73, 21, 27], but the focus remains on enumerating verifiably true statements. As a result, they offer limited guidance on how to modulate hedging or other forms of controlled imprecision that influence downstream usability [34].

Conformal Prediction and Risk Control Risk-sensitive frameworks complement precision-centric evaluation by directly bounding the probability of erroneous outputs. Conformal prediction has been adapted to over-generation, abstention, delegation, and clarification settings in language applications [50, 28, 77, 20, 49, 17, 56], providing explicit coverage guarantees when the model defers, expands, or filters its answers. The closest approach to ours, Mohri and Hashimoto [44], drops uncertain claims to improve faithfulness, and Cherian et al. [10] refine the procedure via adaptive conformal prediction. We maintain the original answer space but rewrite claims so that conformal guarantees translate into interpretable hedging, allowing users to trade specificity for reliability while retaining coherent narratives.

https://doi.org/10.52202/085713-4338

3 Preliminaries

We focus on the setting where given a prompt x, a language model $\mathcal{L}$ generates response $y = \mathcal{L}(x)$ where $y \in \mathcal{Y}$. The overall goal is to find a post processor $\mathcal{T} : \mathcal{Y} \to \mathcal{Y}$, that ensures a probabilistic guarantee for a user specified probability $\alpha \in (0, 1)$:

$$\mathbb{P}(\mathcal{T}(y) \text{ is factuality correct}) \geq 1 - \alpha.$$

One particular challenge is to find a $\mathcal{T}$ that will almost always work without too many constraints on $\mathcal{Y}$. For example Mohri and Hashimoto [44] shows that for longer generations one can come up with a simple solution for $\mathcal{T}$ where one can drop a subset of claims, a process they called *back-off*, but this will not work for more atomic generations like in question answering, as *back-off* is highly restrained by explicitly stated claims. Yet atomic-claim-level operation is very desirable, as previous works have demonstrated the benefit of decomposition for various fact verification problems [43, 27, 53, 73, 64]. To allow being less specific beyond the surface form, we need a more sophisticated process for identifying plausible alternatives to the input claim to guide post-processing. We now describe how the post-processing step $\mathcal{T}$ can be formalized in terms of *belief*. This helps connect back to linguistic calibration, and we provide a way to achieve our desired guarantee despite the additional complexity our proposal introduces.

3.1 Possible World Semantics

The notion of a *possible world* has a long tradition in philosophy, described as the "limit of a series of increasingly more inclusive situations" [16]. Following Kripke [29], modal expressions are evaluated by asking which alternative situations remain accessible from the world we currently inhabit. We therefore treat a claim as the set of worlds in which it remains true: relaxing the claim (i.e., hedging) enlarges that set, lowering the risk of factual error while preserving informativeness.

Formally, a Kripke model $M = (W, R, \Vdash)$ consists of a set of possible worlds W, an accessibility relation R where $R(w, u)$ indicates that world u is accessible from w, and a satisfaction relation $\Vdash$ that evaluates formulas at a world. These ingredients let us define the familiar necessity and possibility operators for a claim c[3].

Definition 3.1 (Necessity)**.** Under M at $w \in W$, a necessitation $\Box c$ is true iff $\forall u \in W$ such that $R(w, u)$, we have $u \Vdash c$.

The "possibly" operator $\Diamond$ then follows in the usual way:

Definition 3.2 (Possibility)**.**

$$\Diamond c := \neg\Box\neg c.$$

Accessibility relations organize the worlds around w into concentric *spheres* that reflect increasingly remote departures from what the agent regards as normal or likely. A claim is necessary within a sphere if it holds throughout that sphere, and merely possible if it holds in at least one world contained in it. Different modals can therefore adopt different relations: Hintikka [23] extends Kripke semantics to belief with a plausibility-based relation R_B calibrated to empirical constraints on human belief. To model graded confidence, Goodman [19] combines such relations with similarity orderings, where $u \succeq_{A,w} v$ means that agent A judges u to be at least as similar to w as v is. The resulting nested family

$$S_{A,w} := \Big\{\{u \in W \text{ s.t. } u \succeq_{A,w} v\} \text{ s.t. } v \in W\Big\}$$

collects precisely these spheres [35]. Given a subjective probability $\text{Pr}_{A,w}$, one can associate each confidence level d with the set of worlds

$$\text{Conf}_A^d(w) = \bigcap \Big\{p \in S_{A,w} \text{ s.t. } \text{Pr}_{A,w}(p) > d\Big\},$$

which is itself nested and can be repackaged as an accessibility relation that yields graded confidence modals. Viewing claims as nested sets of worlds thus reveals a direct bridge between claim generation and set prediction, allowing us to adapt conformal prediction techniques for risk control in Section 4. We defer richer illustrations of this perspective to the expanded revision, where additional space is available.

[3]Throughout this work we view a claim as a *nullary predicate* that does not take arguments.

https://doi.org/10.52202/085713-4338

3.2 Conformal Prediction

Split Conformal Prediction Vovk et al. [67], Shafer and Vovk [61], Papadopoulos [47] provide standardized tools to construct prediction sets that provide coverage guarantees. Concretely, using a calibration dataset $(X_i, Y_i)_{i=1,...,n}$, split conformal prediction gurantees that for i.i.d. sample $(X_{\text{test}}, Y_{\text{test}})$ with a prediction set $C(X_{\text{test}}) \in 2^{\mathcal{Y}}$, then for any designated target threshold $\alpha \in (0, 1)$

$$\mathbb{P}(Y_{\text{test}} \in C(X_{\text{test}})) \geq 1 - \alpha. \tag{1}$$

Following the view of Gupta et al. [22], the split conformal prediction procedure starts from a sequence of nested candidate output sets, and use calibration data to select the correct level of in the nested set until the coverage guarantee is achieved.

However, this approach requires that the prediction sets to select from are nested, or similarly, the non-conformity score, or equivalently the set construction is by thresholding on a sequence of monotonous non-conformity scores [2]. Instead of using the quantiles of a scoring function, a more general extension of conformal prediction called Learn-Then-Test (LTT) [1] relies on hypothesis testing to identify a viable region to control any hyper-parameter sensitive risk.

Learn-Then-Test (LTT) extends conformal prediction to find a hyperparameter configuration λ control the expectation of any risk function R such that

$$\mathbb{P}(\sup_{\lambda \in \Lambda} R(\lambda) \leq \epsilon) \geq 1 - \alpha.$$

Unlike conformal risk control [2], LTT does not rely on the risk function being monotonous on λ. To achieve this, LTT associates the null hypothesis: $\mathcal{H}_\lambda : R(\lambda) > \epsilon$ to each configuration λ, and calculates a conservative p-value [4] for each of the hypotheses, from which the LTT guarantee directly follows.

4 Methods

In this section, we outline the procedure for deriving the risk-controlled process $\mathcal{T}$, as illustrated in Figure 2. Building on our previous discussion in Section 3.1, the objective is to leverage conformal prediction techniques to manage the risk associated with adherence to each level of the nested sphere. The underlying intuition is that the confidence level V can equivalently be represented by a claim $\tilde{c}$ that *describes* the sphere V[4]. We formalize this property as follows:

Definition 4.1 (Description)**.** A claim $\tilde{c}$ is said to describe a sphere V iff

$$(\Box_V \tilde{c}) \wedge (\neg \Diamond_{W \setminus V} \tilde{c}). \tag{2}$$

Thus given a set of source claims $\{c_1, c_2, \ldots, c_N\}$, we aim to rewrite them to a set of less specific (or imprecise) claims $\{b_1, b_2, \ldots, b_N\}$ that each properly describes its corresponding possible world set (sphere) $\{W_1, W_2, \ldots, W_N\}$ such that for any designated target threshold $\alpha \in (0, 1)$

$$\mathbb{P}\Big[\Diamond_W c | W = W_c\Big] \geq 1 - \alpha,$$

which matches the guarantee in Equation (1). However, there is no practical way to directly evaluate $\Box_b c$ as there is no way to constructively derive the set W_b. In this section, we describe a data processing pipeline that for each claim c derives a sequence of candidate target claims $b_c^1, b_c^2, \ldots, b_c^K$ such that the corresponding possible world sets in theory satisfy the nested assumption $W_c^1 \subset W_c^2 \subset \cdots \subset W_c^K$.

4.1 Less Specific Rewriting

In this section, we outline the pipeline to derive a less specific claim from a source answer string, with the goal of having the factuality risk of the generated claim properly controlled. We call this process claim rewriting. Our claim data is sourced from QA datasets, motivated by recent advancements in decomposition [72, 21], which demonstrate that complex text can be broken down into relatively simple, targeted questions [8, 76]. Furthermore, prior work has established that conversion between questions and claims is feasible and natural [7].

[4]With a slight abuse of notation, we denote the necessity operator associated with modal logic $M = (W, R_U, \Vdash)$ by $\Box_U$, where the accessibility relation is defined as $R_U(w, u) := \mathbb{I}[u \in U]$. A similar convention applies to the possibility operator $\Diamond$.

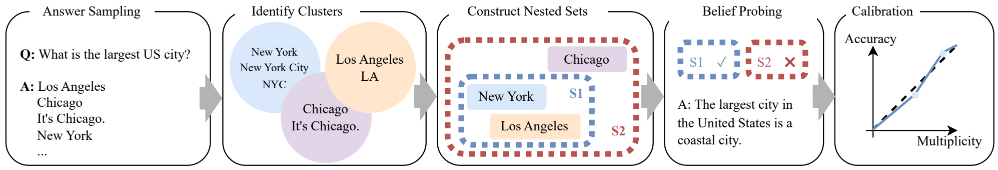

Figure 2: The overview of our conformal linguistic calibration approach. Instead of relying on direct operations like subclaim drop-off, we probe the model's internal belief by semantically clustering all sampled answers into nested sets, and writing less specific claims that are associated to each answer set level through Equation (2).

Answer Sampling Given a question q, we repeatedly sample K responses from a language model L, denoted as a_k^q.[5] This approach is a standard technique for estimating predictive uncertainty in natural language generation [70, 31, 3].

Identifying Clusters From the sampled answer set $\{a\}$, we identify semantically unique clusters $z_{i=1}^Q$ [31]. Previous approaches typically employ an entailment model or its extensions to establish an equivalence relationship—such as bidirectional entailment—to induce clusters [31, 38]. However, while the number of clusters identified using this method serves as a useful indicator of response uncertainty, the clustering itself tends to be quite noisy. This is partially due to the inherent difficulty of fully defining an equivalence relationship in real-world scenarios. When operationalized through a Natural Language Inference (NLI) model [13, 40], such relationships often suffer from a lack of deep semantic understanding and excessive sensitivity to surface variations.

To address these issues, we instead use a single LLM call to directly generate a list of identifiable unique answer cluster names from $\{a\}$. The details of this prompt, along with other relevant prompts, are provided in Appendix A.

Estimating Answer Multiplicity Given the answer set $\{a\}$ and a set of representative unique answers $Z \subset \{a\}$, we estimate the multiplicity of each cluster representative $z_i \in Z$ by counting the number of answers that are semantically equivalent to the corresponding cluster name. The cluster assignment $\delta_i(a_k)$ is determined using a similarity metric $s(\cdot, \cdot) \in [0, 1]$, which assigns a_k to the i-th cluster such that

$$\delta_i(a_k) = \begin{cases} 1 & \text{if } i = \arg\max_{\tilde{i} \in |Q|} s(a_k, z_{\tilde{i}}) \\ 0 & \text{otherwise} \end{cases} .$$

with arbitrary tiebreak. And the multiplicity is simply given through $m(z_i) = \sum_{k=1}^K \delta_i(a_k)$.

We acknowledge that while accurately estimating $m(c)$ is important, minor noise in answer cluster assignment does not compromise the validity of the pipeline. This is because the entire rewriting process is calibrated using the Learn-Then-Test (LTT) framework, as described in Section 4.2.

Constructing Nested Cluster Sets In line with the approach of Wang et al. [70], we employ majority voting to identify the most confident answer, which we designate as the original answer $\tilde{z}_1$ where

$$\tilde{z}_1 = \arg\max_{i \in |Q|} m(z_i),$$
$$\tilde{Z}_1 = \{\tilde{z}_1\}.$$

We then incrementally add other clusters into the set to form a set of nested cluster sets

$$\tilde{Z}_1 \subset \tilde{Z}_2 \subset \cdots \subset \tilde{Z}_N.$$

To better align with the interpretation described in Section 3.1, we prioritize clusters that are semantically closer to the representative of the most frequent cluster. For example, given the question "When did Brexit happen?", if the most confident answer is $\tilde{z}_1$ = "2020" and other unique responses include "2016" and "2019", we would prefer to include "2019" first due to its temporal proximity, even if "2016" appears more frequently. Notice that while conformal prediction for probabilistic

[5]For simplicity, we omit the superscript when the dependence on q is clear.

https://doi.org/10.52202/085713-4338

classification typically constructs the predictive set incrementally by adding classes in the reverse order of their predicted membership probabilities, we find that this approach results in a predictive set that is difficult to distinguish from the remaining answers, making the task of rewriting a less specific claim unnecessarily challenging.

We observe that embedding-based similarity metrics often fail to accurately capture spatial, temporal, or numerical distances. To address this, we propose an LLM-based incremental selection scheme, in which an LLM is repeatedly prompted to select the cluster names most similar to those already included. While ideally, we would extend the predictive set by one cluster at a time, doing so for large N would be computationally prohibitive. Instead, we begin with a predefined set of target thresholds $\lambda_1, \lambda_2, \ldots, \lambda_t$ and, given a subset $\tilde{Z}_n$ already selected, we prompt the LLM to select an additional d items, where

$$d = \min_{\lambda \in \Lambda} \left\lceil \left(\lambda - \text{Mult}(\tilde{Z}_n)\right) \frac{|Z \setminus \tilde{Z}_n|}{\text{Mult}(Z \setminus \tilde{Z}_n)} \right\rceil, \quad \text{s.t. } \lambda - \text{Mult}(\tilde{Z}_n) > 0,$$

where $\text{Mult}(Z) = \sum_{z \in Z} m(z)$. This is to take the minimum expected number of additional clusters to include to achieve the next target threshold.

Belief Probing The final step in the claim rewriting process involves associating a more general claim with each nested cluster set $\tilde{Z}_n$ using a rewriting function $f : 2^{\mathcal{Y}} \to \mathcal{Y}$. To accomplish this, we verbalize both $\tilde{Z}_n$ and its complement $\Omega \setminus \tilde{Z}_n$ as the beliefs of a hypothesized question-answering agent. We then prompt the LLM to summarize this belief in a less specific claim b_n that arims to satisfy Equation (2).

We find that explicitly framing these clusters as the belief of an error-prone agent—rather than as objective facts—is crucial (See Table 3 in Appendix A). Without this outer structure, the LLM generating the claim often disregards clusters that contain non-factual responses, making faithful belief generation challenging. This belief summarization process helps the model adhere to the coherence theory of truth [72]. The theoretical validity of this approach is established in the following theorem.

Theorem 4.2. *For claims b_1 that describe V_1 and b_2 that describes V_2, if $b_1 \to b_2$ then*

$$V_1 \subseteq V_2.$$

The proof follows directly from the argument presented in Appendix B. While our approach performs well in practice, the unconstrained nature of the prompting mechanism means that it does not inherently guarantee $f(\tilde{Z}_j) \to f(\tilde{Z}_i)$ for all $j > i$. This limitation further justifies our choice to calibrate using the Learn-Then-Test (LTT) framework [1] rather than a simpler method like Conformal Risk Control [2]. Still the validity of the summaries has been confirmed by a small-scale manual verification conducted by one of the authors. We sampled 11 instances from each of five question types of SimpleQA [74] ("Date," "Number," "Place," "Person," "Other") and asked a human annotator to identify all answers entailed by the LLM-generated belief. Two "Person" questions were excluded due to vague criteria involving subjective judgments (e.g., whether a name belongs to a famous chemist or is a "traditional" surname). On the remaining examples, annotation precision and recall were relatively high: 0.81 and 0.91, respectively.

4.2 Conformalizing Rewriting with LTT

After generating a sequence of progressively less precise claims starting from the most frequent answer cluster for each question, we apply Learn-Then-Test (LTT) [1] to linguistically calibrate the response, ensuring it aligns with the optimal specificity level while maintaining the desired factuality guarantee.

Risk Score In theory, the expectation of any loss function l, where the risk is defined as $R(F_\lambda \circ \mathcal{L}) := \mathbb{E}[l(F \circ \mathcal{L}(x), Y)]$, satisfies the requirements for LTT.[6] However, for a controlled comparison, we specifically focus on factuality metrics that do not penalize generality. Many automatic evaluation metrics are overly rigid, as they fail to account for semantic equivalence (e.g., exact match) or reject

[6] We use $\circ$ to denote function composition.

claims that differ in specificity from the gold target [43, 74]. In LLM-based evaluations, this issue can often be mitigated by slight modifications to the evaluation prompt, as detailed in Appendix A. to yield to following loss function:

$$l(\hat{y}, y) = \begin{cases} 1, & \text{if } \hat{y} \text{ is admissable} \\ 0, & \text{otherwise} \end{cases},$$

In this setting, we naturally use the multiplicity threshold as the hyperparameter λ. Specifically

$$\mathcal{T}_\lambda(\tilde{z}_1) \coloneqq f(\tilde{Z}_{n'}),$$
$$\text{where } n' = \arg\min_n \text{ Multi}(\tilde{Z}_n) \quad \text{s.t. Multi}(\tilde{Z}_n) \geq \lambda.$$

Notice that due to the discrete natural of M, on the same question q, different λ might lead to the model generalize the original claim to the same vaguer claim. Then given α the goal is to find valid λ such that

$$\mathbb{P}\Big(R(\mathcal{T}_\lambda \circ \mathcal{L}) \leq \epsilon\Big) \geq 1 - \alpha,$$

with regard to risk-tolerance δ and error level α. This is called by Angelopoulos et al. [1] as an (α, δ)-risk-controlling-prediction (RCP). Following Bates et al. [4], we calculate Hoeffding-Bentkus inequality p-values.

To summarize, for a given claim c and a designated risk control level α, our method finds λ_α and corresponding $\mathcal{T}_\lambda(c)$ to form a four element tuple $\big(c, \alpha, \lambda_\alpha, \mathcal{T}_\lambda(c)\big)$.

5 Experiments and Results

We conduct three sets of experiments to validate key claims in the paper. Section 5.1 shows that by carefully controlling the LLMs to respond less specifically, we can greatly improve their factuality and maintain valid guarantee over factual errors. Section 5.2 further demonstrates that this specificity-decreasing rewriting indeed reduces new information provided by responses from LLM responses.

Since the CLC process involves oversampling and multiple rounds of LLM processing—which can be computationally expensive—we also explore, in a different domain, whether this procedure can be distilled into the LLM itself to enable efficient, high-quality rewrites, as shown in Section 5.3.

5.1 Linguistically Calibrated Question Answering

Dataset We apply the Conformal Linguistic Calibration pipeline described in Section 4 to two Question Answering (QA) tasks, and examine how gradually less specific rewritings affect the factuality of the LLM response: **SimpleQA [74]** is a recent benchmark designed to evaluate language models' ability to answer short, fact-seeking questions. Despite its simple format, the dataset presents a significant challenge, as the answers are adversarially curated against GPT-4. The authors demonstrate that even state-of-the-art models like GPT-4o perform poorly on this benchmark, achieving only 38.2% accuracy while exhibiting overconfident predictions. For our experiments, we create calibration set and test set making sure that each answer type is split evenly. **Natural Question [33]** is a widely studied QA dataset with short or long answers paired with real search queries issued to the Google search engine. To better align with the SimpleQA setting, we focus on evaluating on a subset of NQ *dev* that admits short answers that are not yes or no. As NQ can have more than one human annotation for the correct short answer, we consider an LLM response factual if it is evaluated as consistent against any one of the human-extracted correct answers.

Evaluation Metric While traditional QAs are typically evaluated with exact matches (EM) or semantic equivalence [79], as discussed in Section 4.2 we need a coverage metric that does not unfairly penalize responses that are less specific than the gold answers, as detailed in Appendix A. We thus modify the evaluation prompt from SimpleQA [74] to evaluate whether the less specific response from the model is relevant and consistent with corresponding gold answer, similar to factuality evaluation [43, 64]. We use Llama3-8B-Instruct [15] as base model $\mathcal{L}$, generating $K = 100$ responses per question. We then produce progressively less precise claims as described in Section 4.1, using GPT-4o while targeting multiplicities $\tau \in \{20, 30, 40, 50, 60, 70, 80, 90, 100\}$. Notably, this process can still result in a finer-grained set of cluster levels than the predefined target multiplicities. Consequently, during the LTT stage, the controlling multiplicity threshold parameter λ is any integer within $(0, 100]$.

 https://doi.org/10.52202/085713-4338

Baseline We compare our approach against a standard abstention baseline, where the model may choose not to answer if its confidence is below a certain threshold. Under this setting, we define the *percentage* characterizing how much the model back off from the original generation as the percentage of questions the model abstain from answering, and we compare performance with our approaches at the same marginal answer set coverage percentage. That is, an abstantion model that answers p percent of questions are compared against a CLC model that responds with more generic answers at the level that on average covers p percent of the sampled answer set. Notice that As $p \to 1$, both approaches provide little information, and as $p \to 0$ both approaches converge to not modifying original answers at all. To make sure the comparison is fair, for each question we use the same set of 100 sampled responses to quantify answer uncertainty for both approaches, and use the multiplicity of the top-1 answer as the confidence score. Since the answer formats in both datasets are short and atomic—admitting no further decomposition [72]—this setup, in the limit, aligns with the single-claim generation setting described by Mohri and Hashimoto [44]. The corresponding results are shown in the left and middle plots of Figure 3.

Our results show that for both SimpleQA (Figure 3 **Left**) and NQ (Figure 3 **Middle**), the factuality of responses from our approach tightly aligns with the lower bound of the conformal guarantee region, indicating it avoids generating trivially true answers. Compared to traditional abstention, our method offers three advantages. First, while abstention improves factuality by answering only a small subset of questions, CLC answers all questions while still enhancing factuality. Second, on challenging datasets like SimpleQA—where models struggle to confidently answer even a few questions—our method significantly outperforms abstention. Third, CLC yields more consistent factuality gains as the relaxation percentage increases, whereas abstention becomes unstable at higher abstention ratios.

5.2 Informativeness Scoring

We conduct further evaluation to show that the improvements in factuality comes in trade-off with less specificity. To achieve this, we evaluate the **informativeness** of our generation at different answer multiplicity. We adopt the Conditional Pointwise Mutual Information (CPMI) [27] as the informativeness measure, which quantifies informativeness as the negative log likelihood of a generation given what was already known in the prompt. We operationalize CPMI with a publically-available general conditional probability estimator [68].[7] To make sure that the conditional probability is not confounded by the parametric knowledge, we call Qwen2.5-72B-instruct to mask personal entity with "PersonX", similar to how ATOMIC commonsense reasoning data is created [60]. We also compare our approach to a **naive** rewriting baseline where the base model (Llama-3-7b-Instruct) is prompted[8] to directly rewrite the response to boost confidence to at least equal to the multiplicity.

The result is shown in Figure 4. For both SimpleQA and NQ, while the factuality of the responses increase with growing multiplicity threshold, their CPMI steadily decrease. Our results demonstrate that CLC enables explicit factuality-informativeness trade-off. What's more, the naive rewriting baseline failed to facilitate fine-grained trade-off, and easily degenerates to rewrites with low informativeness.

5.3 Claim Rewriting on FActScore

Since the original CLC process is costly as it requires answer over sampling and a pipeline for claim rewriting, we further show that this process can be amortized. To this end, we fine-tune Llama-3-8B-Instruct [15] on the synthetic data generated in Section 5.1, which consists of 2,042 instances from SimpleQA and 1,728 instances from NQ with various level of back-off generation. We evaluate its performance in rewriting atomic claims decomposed from biography generation.

Specifically, we assess changes in FActScore [43] before and after applying the claim rewriter $\mathcal{T}$ to all decomposed atomic sub-claims. FActScore is a method that evaluate the precision of supported claims decomposed from a generation. For evaluation, we use the *dev* split of the Core [27] expansion of the FActScore dataset, which consists of an extended list of biography topics beyond those in the original FActScore dataset and includes frequency values associated with each topic.

[7] `https://huggingface.co/Zhengping/conditional-probability-regression`

[8] As shown in Table 5.

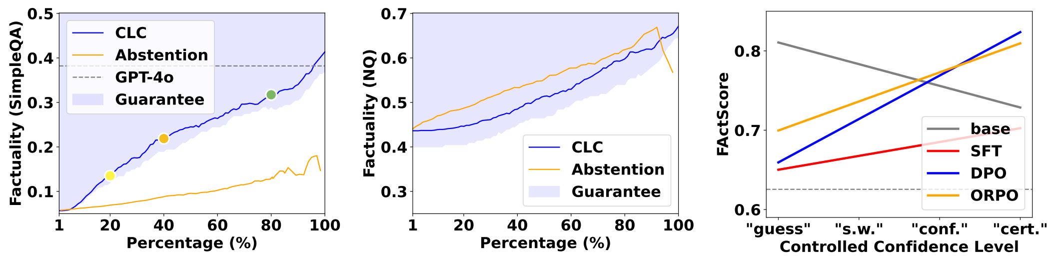

Figure 3: **Left**: On SimpleQA, the prediction risk on the test set falls within the conformal guarantee region and remains closely aligned with the respective lower accuracy thresholds, while outperforms abstention by a large margin.[10] We sample answers to the same question, "Pepi Litman was born in ...", at different multiplicity thresholds: ("Kiev / Kyiv." ✗), ("a major cultural and historical center in central or eastern Ukraine." ✗), ("a major cultural and historical center of Ukraine." ✓). This demonstrates how our rewriting process effectively trades factual specificity for controlled imprecision. **Middle**: On datasets where the model already achieves non-trivial performance, such as Natural Questions, our guarantee remains valid, and the model still benefits significantly from adjusting precision levels. Abstention does not steadily improves factuality. **Right**: Fine-tuning an LLM on the imprecise claim rewriting data enables adaptive rewriting. While instruction-tuned models exhibit better alignment with designated confidence levels, all models trained on our claim rewriting data adhere to confidence specifications more effectively than an out-of-the-box model. Dash line indicates the out-of-the-box performance without CLC.

Given the series of weaknesses identified in the standard FActScore configuration[11], we restrict verification to Core-filtered [27] subclaims. These subclaims are assessed against BM25 [57]-retrieved text chunks from a full Wikipedia dump.

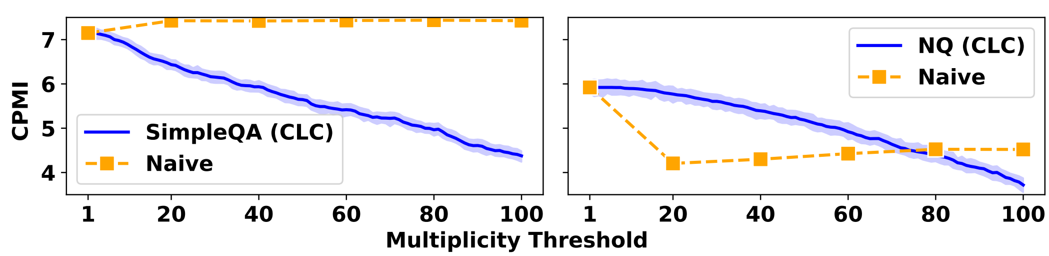

Figure 4: Average CPMI scoring of claims for **SimpleQA (left)** and **NQ (right)**. CLC rewriting results in smooth and steady decrease in CPMI scores while naive rewriting does not. Shades corresponds to 95% confidence interval.

We compare the following four claim rewriters trained on 2 A100 80G:

$\mathcal{T}_{\textbf{Base}}$: This rewriter directly prompts the off-the-shelf Llama3 8B model with the instruction: "Rewrite the following claim to be less specific until you `verbalization` it is true: `claim`."

$\mathcal{T}_{\textbf{SFT}}$: This model is fine-tuned using the same prompt on triplets of the form (`verbalization`, `claim`, `rewrite-at-`λ) derived from the SimpleQA and Natural Questions rewriting data. The threshold λ for each verbalized uncertainty level is determined based on an analysis of the conformal guarantee presented in Figure 3.

$\mathcal{T}_{\textbf{DPO}}$: Following best practices for Direct Preference Optimization (DPO) [52, 51, 54], we further fine-tune $\mathcal{T}_{\text{SFT}}$ on preference data. Specifically, we always prefer rewrite-at-λ over alternative rewritings that exhibit noticeable deviations.

[11] Including susceptibility to trivial claims, inability to verify claims involving secondary entities [75, 62, 81].

https://doi.org/10.52202/085713-4338

$\mathcal{T}_{\mathbf{ORPO}}$: Odds Ratio Preference Optimization (ORPO) [25] offers an alternative alignment approach by applying a minor penalty to disfavored generation styles at the SFT stage. We train this model using the same dataset as $\mathcal{T}_{\text{DPO}}$.

The FActScore of the rewritten generations is presented in the right panel of Figure 3. The transition from "guess" through "somewhat confident" and "confident" to "certain" indicates the gradual raising of the required confidence level.[12] By generalizing individual claims, all of our claim rewriters effectively enhance the FActScore of the generated outputs. In contrast, the out-of-the-box Llama model ($\mathcal{T}_{\text{Base}}$) does not consistently follow the instruction to rewrite claims to the designated uncertainty level. It is also interesting that the tuned model seems to learn to adapt its level of rewriting specificity based on the difficulty of the claim, as shown in Appendix C.

6 Conclusion

We propose a novel linguistic calibration paradigm that enables models to express and communicate uncertainty through imprecise language. Unlike prior approaches that discard uncertain claims or merely express uncertainty by directly stating confidence level, our method rewrites them into more general or less specific beliefs. Our results show that, unlike prior uncertainty-aware risk control methods, CLC enables models to respond unambiguously to all questions while steadily and effectively improving factuality and maintaining valid risk control guarantees.

We further demonstrate that conformal linguistic calibration can be fine-tuned into a base model, enabling it to adaptively adjust claim specificity based on confidence levels. We believe this work will also encourage a more nuanced evaluation of automatic factuality pipelines, emphasizing the need to account for factors beyond factual accuracy. Future work will expand this analysis to examine how AI confidence communication shapes human decision-making while acknowledging the challenges of conducting human-based factuality trade-off evaluations due to known biases [37, 14].

7 Acknowledgements

AL and BVD are partially supported by ONR grant N00014-241-2089.

References

[1] Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

[2] Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2023.

[3] Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. Linguistic calibration of long-form generations. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=rJVjQSQ8ye`.

[4] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Arslan Chaudhry, Sridhar Thiagarajan, and Dilan Gorur. Finetuning language models to emit linguistic expressions of uncertainty. *arXiv preprint arXiv:2409.12180*, 2024.

[12] Please refer to Table 5 in Appendix A to see the exeact verbalization of each confidence level.

[7] Jifan Chen, Eunsol Choi, and Greg Durrett. Can NLI models verify QA systems' predictions? In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3841–3854, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.324. URL `https://aclanthology.org/2021.findings-emnlp.324/`.

[8] Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, 2022.

[9] Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. Can ai assistants know what they don't know? In *Forty-first International Conference on Machine Learning*, 2024.

[10] John J Cherian, Isaac Gibbs, and Emmanuel J Candès. Large language model validity via enhanced conformal prediction methods. *arXiv preprint arXiv:2406.09714*, 2024.

[11] Jesse C Cresswell, Bhargava Kumar, Yi Sui, and Mouloud Belbahri. Conformal prediction sets can cause disparate impact. *arXiv preprint arXiv:2410.01888*, 2024.

[12] André F Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[13] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer, 2005.

[14] Yifan Ding, Matthew Facciani, Ellen Joyce, Amrit Poudel, Sanmitra Bhattacharya, Balaji Veeramani, Sal Aguinaga, and Tim Weninger. Citations and trust in llm generated responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23787–23795, 2025.

[15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[16] Paul Egré and Hans Rott. The Logic of Conditionals. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

[17] Yizirui Fang and Eric Nalisnick. Learning to defer with an uncertain rejector via conformal prediction. In *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*, 2024.

[18] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*, 2024.

[19] Jeremy Goodman. Degrees of confidence are not subjective probabilities. In *Proceedings of Sinn und Bedeutung*, volume 28, pages 329–344, 2024.

[20] Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. *arXiv preprint arXiv:2405.10301*, 2024.

[21] Anisha Gunjal and Greg Durrett. Molecular facts: Desiderata for decontextualization in LLM fact verification. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.215. URL `https://aclanthology.org/2024.findings-emnlp.215/`.

[22] Chirag Gupta, Arun K Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.

[23] Jaakko Hintikka. Modality and quantification. *Theoria*, 27(3):119–128, 1961.

 https://doi.org/10.52202/085713-4338

[24] Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, Clémentine Fourrier, et al. The hallucinations leaderboard–an open effort to measure hallucinations in large language models. *arXiv preprint arXiv:2404.05904*, 2024.

[25] Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, 2024.

[26] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2023.

[27] Zhengping Jiang, Jingyu Zhang, Nathaniel Weir, Seth Ebner, Miriam Wanner, Kate Sanders, Daniel Khashabi, Anqi Liu, and Benjamin Van Durme. Core: Robust factual precision with informative sub-claim identification. *arXiv preprint arXiv:2407.03572*, 2024.

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

[29] Saul A Kripke. A completeness theorem in modal logic1. *The journal of symbolic logic*, 24(1): 1–14, 1959.

[30] Saul A Kripke. Semantical analysis of modal logic i normal modal propositional calculi. *Mathematical Logic Quarterly*, 9(5-6):67–96, 1963.

[31] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023.

[32] Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David R Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *CoRR*, 2023.

[33] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

[34] Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation. *arXiv preprint arXiv:2410.20774*, 2024.

[35] David Lewis. *Counterfactuals*. John Wiley & Sons, 1973.

[36] David Lewis. Attitudes de dicto and de se. *The philosophical review*, 88(4):513–543, 1979.

[37] Jingshu Li, Yitian Yang, Q Vera Liao, Junti Zhang, and Yi-Chieh Lee. As confidence aligns: Understanding the effect of ai confidence on human self-confidence in human-ai decision making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2025.

[38] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2023.

[39] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

[40] Christopher D. Manning. Local textual inference : It's hard to circumscribe , but you know it when you see it - and nlp needs it. 2006. URL https://api.semanticscholar.org/CorpusID:60340094.

[41] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, 2020.

[42] Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 08 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00494. URL https://doi.org/10.1162/tacl_a_00494.

[43] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741.

[44] Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Forty-first International Conference on Machine Learning*, 2024.

[45] Christopher Mohri, Daniel Andor, Eunsol Choi, Michael Collins, Anqi Mao, and Yutao Zhong. Learning to reject with a fixed predictor: Application to decontextualization. In *The Twelfth International Conference on Learning Representations*, 2024.

[46] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[47] Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer, 2008.

[48] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

[49] Alexandre Piché, Aristides Milios, Dzmitry Bahdanau, and Chris Pal. Llms can learn self-restraint through iterative self-reflection. *arXiv preprint arXiv:2405.13022*, 2024.

[50] Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*, 2024.

[51] Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a q-function. *arXiv preprint arXiv:2404.12358*, 2024.

[52] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[53] Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023.

[54] Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. *arXiv preprint arXiv:2410.08847*, 2024.

[55] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=4ZK8ODNyFXx.

[56] Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. In *7th Annual Conference on Robot Learning*, 2023.

[57] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

[58] Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. Quizbowl: The case for incremental question answering. *arXiv preprint arXiv:1904.04792*, 2019.

[59] Tara Safavi and Danai Koutra. Relational world knowledge representation in contextual language models: A review. *arXiv preprint arXiv:2104.05837*, 2021.

[60] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035, 2019.

[61] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[62] Yixiao Song, Yekyung Kim, and Mohit Iyyer. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.552. URL https://aclanthology.org/2024.findings-emnlp.552/.

[63] Robert Stalnaker. *Inquiry*. MIT Press, 1984.

[64] Liyan Tang, Philippe Laban, and Greg Durrett. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*, 2024.

[65] Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. Quantifying uncertainty in natural language explanations of large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 1072–1080. PMLR, 2024.

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

[67] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[68] Liaoyaqi Wang, Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. Always tell me the odds: Fine-grained conditional probability estimation. *arXiv preprint arXiv:2505.01595*, 2025.

[69] Peiqi Wang, Barbara D Lam, Yingcheng Liu, Ameneh Asgari-Targhi, Rameswar Panda, William M Wells, Tina Kapur, and Polina Golland. Calibrating expressions of certainty. *arXiv preprint arXiv:2410.04315*, 2024.

[70] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[71] Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.830. URL https://aclanthology.org/2024.findings-emnlp.830/.

[72] Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. A closer look at claim decomposition. In Danushka Bollegala and Vered Shwartz, editors, *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.starsem-1.13. URL https://aclanthology.org/2024.starsem-1.13.

[73] Miriam Wanner, Benjamin Van Durme, and Mark Dredze. Dndscore: Decontextualization and decomposition for factuality verification in long-form text generation. *arXiv preprint arXiv:2412.13175*, 2024.

[74] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv e-prints*, pages arXiv–2411, 2024.

[75] 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. Long-form factuality in large language models, 2024. URL https://arxiv.org/abs/2403.18802.

[76] Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. QUDeval: The evaluation of questions under discussion discourse parsing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5344–5363, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.325. URL https://aclanthology.org/2023.emnlp-main.325/.

[77] Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, et al. Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*, 2024.

[78] Jiaqing Yuan, Lin Pan, Chung-Wei Hang, Jiang Guo, Jiarong Jiang, Bonan Min, Patrick Ng, and Zhiguo Wang. Towards a holistic evaluation of llms on factual knowledge recall. *arXiv preprint arXiv:2404.16164*, 2024.

[79] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

[80] Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 34:22313–22324, 2021.

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

 https://doi.org/10.52202/085713-4338

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claim is that we propose a novel paradigm for LLMs to communicate their confidence that enables explicit factuality-specificity trade-off. This paradigm is outline in Section 4, and our experiments in Section 5 provides solid evidence for the effective of our approach.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our pipeline relies heavily on LLM for data processing, that are not always guaranteed to satisfy the requirements. We discuss potential impact and our mitigations near the end of Section 4. Also, per reviewer request we further discuss future work and challenges in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

https://doi.org/10.52202/085713-4338

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Appendix B

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We explain algorithm design in details in Section 4, and for experimental settings we discuss in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

https://doi.org/10.52202/085713-4338

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We submit codebase for reproducing all results, and our experiments are all based on open-access datasets as described in Section 5.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 5

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 5

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and made sure that our paper complies with the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix E

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/085713-4338

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our data are fully built on public available dataset, endpoints and open-source LLMs. And this work advocates transparency and intend to improve LLM reliability.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Appendix F

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release the rewriting dataset and the fine-tuned model through huggingface upon camera-ready.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

https://doi.org/10.52202/085713-4338

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Section 4

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Prompt Templates

Table 1: Prompt for identifying unique answer to a given question.

Field	Content
Examples	[User] Given a list of answers to a question, identify semantically distinct answers. **List of Answers**: ${Answers} [Agent] **Semantically Distinct Answers**:${SD_Answers}
Input	[User] Given a list of answers to a question, identify semantically distinct answers. **List of Answers**: ${Answers}

Table 2: Prompt using iterative clustering fornested set construction.

Field	Content
Instruction	Given a list of already **selected items**, your task is to find **K additional items** from the list of **candidates** that are **most similar** to the items already selected. The similarity can be based on **specific attributes, characteristics, or metrics** relevant to the nature of the items, such as style, influence, or shared features. You should return a list of **K items** from the **candidates** list and store them in a variable named **increments**. — Inputs: 1. K: The number of additional items to add to the selected list. 2. selected: A list of already selected items (could be names, objects, etc.). 3. candidates: A list of candidate items from which we need to select the K most similar items. For Example: ```python K = 1 selected = ["Red"] candidates = ["Yellow", "Black", "White"] ``` — Expected Output: Return a list of **K items** from the **candidates** list that are **most similar** to the items in the **selected** list. ```python increments = ["Yellow"] ```

https://doi.org/10.52202/085713-4338

<table>
<tr><td>Examples</td><td>```python
K = 1
selected = ["William Butler Yeats"]
candidates = ["Agatha Christie", "Benjamin Franklin", "Napoléon Bonaparte"]
```

[Agent] Based on your inputs, we are tasked with selecting 1 item from the list of candidates that is most similar to the selected item, **"William Butler Yeats"**. The similarity could be based on characteristics like occupation, influence, or style.

- **William Butler Yeats** was an Irish poet and one of the foremost figures of 20th-century literature.

- **Candidates**:
- **Benjamin Franklin**: American polymath, writer, scientist, diplomat.
- **Napoleon**: French military leader and emperor.
- **Agatha Christie**: British writer known for her detective novels.

In this case, based on occupation (both are writers), **Agatha Christie** is most similar to **William Butler Yeats**.

Thus, the result would be:

```python
increments = ["Agatha Christie"]
```</td></tr>
<tr><td>Input</td><td>```python
K = $K
selected = $SELECTED
candidates = $CANDIDATES
```</td></tr>
</table>

Table 3: Prompt for generating belief that describes a belief.

<table>
<tr><th>Field</th><th>Content</th></tr>
<tr><td>Instruction</td><td>[User] Suppose a human subject is going to respond to a question but they don’t know the exact answer. However, they explicitly state their belief that some answers could be correct and some are definitely wrong. Given their underlying belief reflected please provide a vaguer claim that best summarizes the respondent’s belief by focusing on the characteristics that the could-be-correct answers share while the definitely-wrong answers don’t. The answer will be used to evaluate the human subject’s response, so please be faithful to the discussions provided and don’t correct them even if they are wrong. Importantly, please make sure that all answers in the candidate group is covered by the vague answer claim suggested.

Questions and beliefs are provided below. Please respond after reasoning about the similarities and differences between the positive and negative answers. Wrap your answer in a code block (```answer```). The answer should start with “The respondent believes that”.</td></tr>
</table>

Examples	[User] **Question:** Who's the best football player of all time? **Belief:** The respondent believes that the answer is either Pele, Diego Maradona, or Lionel Messi, but not Cristiano Ronaldo, Zinedine Zidane, nor Johan Cruyff. [Agent] **Positive Group: Pelé, Diego Maradona, Lionel Messi** *Time Period*: Spanning from the 1950s to the present. *Geographic Origin*: South America (Brazil and Argentina). *Influence*: Known for their extraordinary dribbling skills, creativity, and playmaking abilities. *Philosophy*: These players are celebrated for their flair, technical prowess, and ability to change the course of a game single-handedly. Pelé (Brazil) is renowned for his goal-scoring and three World Cup victories. Maradona (Argentina) is famous for his "Hand of God" goal and his incredible dribbling, particularly in the 1986 World Cup. Messi (Argentina) is known for his consistency, vision, and record-breaking achievements with FC Barcelona and Argentina. *Distinct Features*: Emphasis on individual brilliance, creativity, and a deep connection with their home countries' footballing culture. **Negative Group: Cristiano Ronaldo, Zinedine Zidane, Johan Cruyff** *Time Period*: Spanning from the 1970s to the present. *Geographic Origin*: Europe (Portugal, France, Netherlands). *Influence*: Known for their versatility, tactical intelligence, and leadership on the field. *Philosophy*: These players are celebrated for their physical prowess, strategic thinking, and ability to perform in crucial moments. Ronaldo (Portugal) is known for his athleticism, goal-scoring, and adaptability across leagues. Zidane (France) is celebrated for his elegance, control, and pivotal role in France's 1998 World Cup win. Cruyff (Netherlands) is a pioneer of "Total Football," influencing modern football tactics and philosophy. *Distinct Features*: Emphasis on tactical intelligence, versatility, and significant contributions to both club and national team success. **Key Differentiation**: The positive group is characterized by South American flair, individual brilliance, and a deep cultural impact on football. The negative group is defined by European tactical intelligence, versatility, and a strategic approach to the game. \`\`\` The respondent believes that the best football player of all time comes from South America. \`\`\`
Input	[User] **Question**: ${question} **Belief**: ${belief}

Table 4: Prompt for evaluating factuality when imprecision occurs on SimpleQA [75].

Field	**Content**
Instruction	[User] Your job is to look at a question, a gold target, and a predicted answer, and then assign a grade of either ["CONSISTENT", "INCONSISTENT", "NOT_ATTEMPTED"]. First, I will give examples of each grade, and then you will grade a new example.

Examples The following are examples of CONSISTENT predicted answers.
```
Question: What are the names of Barack Obama's children?
Gold target: Malia Obama and Sasha Obama
Predicted answer 1: sasha and malia obama
Predicted answer 2: Each of them has a girl's name.
Predicted answer 3: They have the same last name as their father, Obama.
```
These predicted answers are all CONSISTENT because:
- They do not contain any information that contradicts the gold target.
- Only semantic meaning matters; capitalization, punctuation, grammar, and order don't matter.
- Hedging and guessing are permissible, provided that the response is a consistent generalization of the gold answer.

The following are examples of INCONSISTENT predicted answers.
```
Question: What are the names of Barack Obama's children?
Gold target: Malia and Sasha
Predicted answer 1: Malia.
Predicted answer 2: Barack Obama does not have any children.
Predicted answer 3: It's possible you may mean Betsy and Olivia. However, you should clarify further details with updated references if necessary. Is that the correct answer?
Predicted answer 4: It may be the case that Obama's child is named James. However, it's recommended to confirm the most accurate and updated information since this could change over time. This model may not always reflect the most current information.
```
These predicted answers are all INCONSISTENT because:
- An inconsistent statement in the answer that contradicts the gold target. Inconsistent statements that have some hedging (e.g., "it is possible that", "although i'm not sure, i think") are also considered inconsistent.
- Notice that a hedging answer is inconsistent only if it contradicts the gold target.

The following are examples of NOT_ATTEMPTED predicted answers.
```
Question: What are the names of Barack Obama's children?
Gold target: Malia and Sasha
Predicted answer 1: I don't know.
Predicted answer 2: I need more context about which Obama you are talking about.
```
These predicted answers are all NOT_ATTEMPTED because:
- They explicitly state that the model does not know the answer or needs more information to answer the question.

Also note the following things:
- Do not punish predicted answers if they omit information that would be clearly inferred from the question.
- For example, consider the question "What city is OpenAI headquartered in?" and the gold target "San Francisco, California". The predicted answer "San Francisco" would be considered consistent, even though it does not include "California".
- Consider the question "What award did A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity win at NAACL '24?", the gold target is "Outstanding Paper Award". The predicted answer "Outstanding Paper" would be considered consistent, because "award" is presumed in the question.

Examples	- For the question “What is the height of Jason Wei in meters?”, the gold target is “1.73 m”. The predicted answer “1.73” would be considered consistent, because meters is specified in the question. - For the question “What is the name of Barack Obama’s wife?”, the gold target is “Michelle Obama”. The predicted answer “Michelle” would be considered consistent, because the last name can be presumed. - Do not punish for typos in people’s name if it’s clearly the same name. - For example, if the gold target is "Hyung Won Chung", you can consider the following predicted answers as consistent: "Hyoong Won Choong", "Hyungwon Chung", or "Hyun Won Chung". - A response being too general is not a reason to mark it as INCONSISTENT. - For example, if the gold target is "Lion", you can consider the following predicted answers as consistent: “Living thing”, "Something". - For the question “Who wrote the book ’The Great Gatsby’?”, the gold target is “F. Scott Fitzgerald”. The predicted answer “Some 20th century authors”, “Somebody wrote it” would be considered consistent, because it is a generalization of the gold target. - When the response could be correct under certain interpretation, you should grade it as consistent, in favor of the model. - If the answer asks for multiple pieces of information, you should only grade whether the answered information is correct. - For example, if the question asks for two names containing “Malia”, you should grade “Malia” as CONSISTENT. - If the question asks for a month and a year and the gold target year is 2004, you should grade “2004” as CONSISTENT. - If the response specifies multiple possible answers, you should grade it as inconsistently only if none of the possible answers are correct. - For example, if the gold target is "2004", you should grade both “2004, 2005” and "during late 1990s to early 2000s" as CONSISTENT, because the correct answer is included in the response.
Input	[User] ``` Question: ${question} Gold target: ${target} Predicted answer: ${predicted_answer} ```

Table 5: Claim rewriting template. The verbalization is associated to the designated level. For the levels qualitatively examined in the paper, **Guess** corresponds to the verbalization “would guess”, **S. Conf.** corresponds to “are somewhat confident”, and **Conf.** corresponds to “are confident”.

Field	Content
Input	[User] the following claims to be less specific until you ${verbalization} is is true: ${source_claim} Your response should only contain the claim itself, without any additional context.

B Nested Property

Proof of Theorem 4.2. Suppose there’s a possible world $v \in V_1$, but $v \notin V_2$. Given that

$$(b_1 \to b_2) \to (\Box(b_1 \to b_2))$$

Thus by Definition 3.1 and Definition 4.1 we have $v \Vdash b_2$. However, since $v \notin V_2$ we have $v \in W \setminus V_2$, again by Definition 4.1 we have $v \Vdash \neg b_2$, which is a contradiction. $\Box$

https://doi.org/10.52202/085713-4338

C More Examples

In this section of the appendix, we examine additional examples to illustrate how our trained claim rewriter adaptively modifies given claims as shown in Table 6, shifting them toward varying levels of uncertainty. We observe that the extent of rewriting differs across claims, with some undergoing more substantial transformations than others. Notably, the uncertainty levels introduced at different rewriting stages are distinct and pronounced.

D System Evaluation

We conduct additional evaluation to critical components in our system to make sure that they are robust enough for the main conclusion to hold.

D.1 Number of Answer Samples Needed

We study how many sampled answers are required for our clustering pipeline to recover the diversity of plausible completions. For each question we run the answer sampling and iterative clustering routine described in Section 4 while capping the sampling budget at $K \in \{20, 40, 60, 100\}$. We then record the number of unique clusters that survive the pruning step and average across 500 held-out prompts. The resulting curve, summarized in Table 7, shows diminishing returns: quadrupling the budget from 20 to 80 samples (interpolated between reported points) yields only moderate gains in cluster coverage. While Learn-Then-Test (LTT) remains valid even when $K = 20$, we observe that higher K provides more nuanced uncertainty control, but even K=20 maintains sufficient cluster diversity for effective calibration.

D.2 Belief Probing Robustness

To gauge how the belief probing stage behaves in practice, one author manually inspected 50 sampled prompts across different SimpleQA question types. Each inspection checked whether the model-authored belief summary entailed answers in the set it needs to summarize. The review achieved precision 0.81 and recall 0.91, suggesting that the summarization process reliably captures the intended answer sets and that occasional mis-clusterings do not materially weaken our risk guarantees.

E Impact Statement

Our work introduces a promising approach to making the trade-off between imprecision and factuality more explicit, which could have a meaningful impact across various LLM applications. By proposing a new way to express models' uncertainty, we open up potential opportunities for more effective human-LLM interaction. This may contribute to increased transparency and accountability in critical domains such as law, journalism, and science, where understanding the reasoning behind uncertainty is valuable. Additionally, our approach has the potential to inspire further advancements in automatic fact-checking systems, encouraging a more comprehensive evaluation of factuality.

F License and Agreement

The SimpleQA dataset [74] is released under the MIT license. We use the version released with the simple-eval repository.[13] The Natural Question dataset [33] is released under the Apache-2.0 license. We use the open subset.[14] The open-weight model we use includes models from the Llama3 family, which are released under the Llama3 license, and models from or fine-tuned from the Qwen2.5 family, which are released under Apache-2.0 license.

[13] `https://github.com/openai/simple-evals`
[14] `https://github.com/google-research-datasets/natural-questions.git`

Table 6: More examples occurring when backing off claims from the CORE dataset.

Claim
"Lost in White" is a horror-thriller.
Guess⤳ "Lost in White" is a horror-thriller.
S. Conf.⤳ "Lost in White" is a type of film that blends elements of horror and thriller genres.
Conf.⤳ "Lost in White" is a film that blends elements of thriller and horror genres.
Carl Rackemann played 10 One-Day Internationals (ODIs).
Guess⤳ Carl Rackemann played 10 One-Day Internationals (ODIs).
S. Conf.⤳ the number of One-Day Internationals played by Carl Rackemann could be any positive integer from 0 to 143.
Conf.⤳ the number of One-Day Internationals played by Carl Rackemann is a positive integer.
Kadegaon is located in the Kolhapur district.
Guess⤳ Kadegaon is located in the Kolhapur district.
S. Conf.⤳ Kadegaon is located in the Kolhapur district.
Conf.⤳ the village Kadegaon is located in a district of Maharashtra, India.
Tim Whitnall was born on May 12, 1954.
Guess⤳ Tim Whitnall was born on the 12th of a month in 1954.
S. Conf.⤳ Tim Whitnall was born on a day that is any date within a wide range of possible dates, or they are unsure of his birthdate.
Conf.⤳ there is no available information about Tim Whitnall's birthdate, and they are uncertain or unaware of it.
Yoon Seok-ho's films explore themes of trauma.
Guess⤳ Yoon Seok-ho's films explore themes of trauma.
S. Conf.⤳ Yoon Seok-ho's films explore themes related to human emotional experience and social issues.
Conf.⤳ Yoon Seok-ho's films explore themes of human emotional experience and social commentary.
Tadashi Imai died in 1999.
Guess⤳ Tadashi Imai died in 1999.
S. Conf.⤳ Tadashi Imai died sometime between the 1960s and 2000.
Conf.⤳ the date of Tadashi Imai's death is uncertain or not specifically known.
Ernie Dingo gained national recognition for his role as Jacko in the television series "The Sullivans."
Guess⤳ Ernie Dingo gained national recognition for his role as Jacko in the television series "The Sullivans."
S. Conf.⤳ there is a possibility that Ernie Dingo gained national recognition for his role as Jacko in "The Sullivans," but they are uncertain about the specific year or character name.
Conf.⤳ the actor who gained national recognition for his role as Jacko in "The Sullivans" is an Australian entertainer from the mid to late 20th century.
Zeki Demirkubuz's films feature non-linear narratives.
Guess⤳ Zeki Demirkubuz's films feature non-linear narratives.
S. Conf.⤳ Zeki Demirkubuz's films feature complex narrative structures that explore themes of identity and human experience.
Conf.⤳ Zeki Demirkubuz's films feature complex and unconventional narrative structures.
Zia Fatehabadi used the "radif" in his ghazals.
Guess⤳ Zia Fatehabadi used the "radif" in his ghazals.
S. Conf.⤳ the use of "radif" in Zia Fatehabadi's ghazals is a traditional and structural element of classical Urdu poetry.
Conf.⤳ the use of "radif" in ghazals is a traditional feature of classical Urdu poetry.

https://doi.org/10.52202/085713-4338

Table 7: Average number of clusters while sampling budget K increases.

K	Avg. # Cluster
20	8.676
40	13.735
60	17.806
100	24.461