On Subjective Uncertainty Quantification and Calibration in Natural Language Generation

Ziyu Wang **Chris Holmes**
University of Oxford

Abstract

Applications of large language models often involve the generation of free-form responses, in which case uncertainty quantification becomes challenging. This is due to the need to identify task-specific uncertainties (e.g., about the semantics) which appears difficult to define in general cases. This work addresses these challenges from a perspective of Bayesian decision theory, starting from the assumption that our utility is characterized by a similarity measure that compares a generated response with a hypothetical true response. We discuss how this assumption enables principled quantification of the model's subjective uncertainty and its calibration. We further derive a measure for epistemic uncertainty, based on a missing data perspective and its characterization as an excess risk. The proposed methods can be applied to black-box language models. We illustrate the methods on question answering and machine translation tasks. Our experiments provide a principled evaluation of task-specific calibration, and demonstrate that epistemic uncertainty offers a promising deferral strategy for efficient data acquisition in in-context learning.

1 INTRODUCTION

We are interested in uncertainty quantification (UQ) for language models (LMs) in *free-form* natural language generation (NLG): given instruction I, the model generates a response y' based on its predictive distribution $p_M(y \mid I)$, where y' could be *any* natural language passage that fits the instruction. An example of this is in question answering (QA): given a question from the user, the model may provide a brief answer, but it may also follow with supporting facts and explanations, which can vary in form and detail. The user can be satisfied by a wide variety of responses, irrespective of their style or (to some extent) the choice of supporting facts included.

Free-form NLG poses significant challenges to uncertainty quantification: some aspects of generation are irrelevant to the task's purpose and best excluded from uncertainty quantification, but it often appears that we are unable to characterize them precisely. If left unaddressed, however, the model's variation in the irrelevant aspects may dominate in standard uncertainty measures such as token-level entropy (Kuhn et al., 2023), making them uninformative about the model's actual performance on the task.

Starting from Kuhn et al. (2023), a recent line of work (Kuhn et al., 2023; Lin et al., 2024b; Zhang et al., 2023; Aichberger et al., 2024; Nikitin et al., 2024) studied this issue and proposed measuring the "semantic uncertainty" of generation; "semantics" is defined as the equivalence class of textual responses that logically entail one another. Empirical improvements in downstream tasks evidenced their contributions and highlighted the importance of *task-specific* uncertainty quantification, but important conceptual and practical issues remain. From a practical perspective, semantic equivalence is estimated using machine learning models, resulting in imprecise estimates that do not necessarily define an equivalence relation. The imprecision necessitates the introduction of heuristics to post-process the estimates or to aggregate them through other means. Conceptually, the notion of semantic equivalence does not always provide a valid or complete characterization of relevance for certain tasks. Style transfer (Jin et al., 2022) tasks provide an example of the first kind. For the second scenario, consider a QA task where the model is fully certain about the answer to a question, as well as a large set of supporting facts, each of which provides complete and independent justification for the answer; yet there is still the "uncertainty", or variation

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

across generations, about which facts are included in a response. Then very few of the model generations may logically entail one another, and there is a very high level of "semantic uncertainty" just as if there was complete uncertainty about the correct answer.

This paper is about the observation that the above challenges can be resolved in a more general setup from a perspective of *Bayesian decision theory* (Savage, 1954). We start by assuming that we can compute a similarity-type measure $S(y', y; I)$ that characterizes our *utility function* when the model generates y' in respond to I and y is a (true or hypothetical) correct response. This generalizes previous works based on semantic equivalence, which corresponds to defining S using entailment. It can also be applied to traditional structured prediction tasks where a choice of S is readily available (e.g., lexical scores for machine translation, Koehn, 2009), as well as any task where evaluation can be implemented using LMs that compute S based on few-shot demonstrations or detailed instructions (e.g., Yu et al., 2023; Wang et al., 2023; Bannur et al., 2024).

To quantify the subjective uncertainty defined by the LM, we assume the generation y is always chosen to maximize the expected utility under the LM's predictive distribution p_M (§2.1). This is a common, albeit implicit, assumption in NLG (Bickel and Doksum, 1997; Bertsch et al., 2023). It can be supported by a belief that high-capacity LMs may approximate Bayesian inference (e.g., Xie et al., 2021; Akyürek et al., 2022; Hahn and Goyal, 2023), and is also more broadly applicable whenever we view p_M as our best available model for $y \mid I$. Subjective uncertainty is then naturally characterized by the Bayes risk, or equivalently the maximum achievable expected utility, where the action space is defined by the candidate generations available. This simple observation allows us to understand that previous methods for "semantic uncertainty" can be adapted to a broader range of scenarios, regardless of whether semantic equivalence is relevant. It also highlights a unique, principled approach for aggregating similarity measures among generations (§2.2).

For subjective uncertainty measures to be useful, it will be ideal if the LM is more calibrated. It has been unclear in previous works how calibration can be evaluated in general NLG tasks, given the distinction between relevant and irrelevant differences for a task and our apparent inability to formally characterise it. The decision-theoretic view provides a natural answer to this question: an LM is deemed to possess a calibrated notion of uncertainty if the expected subjective utility of its action matches the expectation of the actually incurred utility w.r.t. the true data distribution (§2.3). This observation allows us to quantify the calibration of LMs through reliability diagrams (Murphy and Winkler, 1977) and a generalization of the expected calibration error (Naeini et al., 2015).

Bayesian modelling enables the decomposition of predictive uncertainty into *epistemic uncertainty* and *aleatoric uncertainty* (Der Kiureghian and Ditlevsen, 2009). Epistemic uncertainty refers to the reducible proportion of uncertainty and can guide data acquisition (Kendall and Gal, 2017), or more broadly, indicate when additional information can be most effectively used to improve prediction. The quantification of epistemic uncertainty is of interest in many LM applications, as predictions can often benefit from additional information (Lewis et al., 2020; Brown et al., 2020; Li et al., 2023), which may be costly to obtain (Agarwal et al., 2024; Min et al., 2024). Yet it appears challenging due to the black-box nature of LMs and the free-form generations involved. As we show in §2.4, the decision-theoretic view leads to a principled measure of task-specific epistemic uncertainty in in-context learning (ICL, Brown et al., 2020); key additional ingredients include a missing data perspective to Bayesian modelling (Fong et al., 2023) and a connection between epistemic uncertainty and an excess risk (Xu and Raginsky, 2022). The resulted measure is connected to and generalizes several recent methods, which provides independent justification for its use.

The methodological developments allow us to better understand the "probabilistic uncertainty" in modern LMs, that is, the uncertainty reflected in the variation of the model's predictive distribution. §4 provides experimental illustrations: We revisit past comparisons (Tian et al., 2023; Lin et al., 2024b) of probabilistic and non-probabilistic UQ for instruction-tuned LMs by conducting a principled evaluation of task-specific calibration (§4.1). In §4.2, we demonstrate that epistemic uncertainty provides a deferral strategy (Madras et al., 2018; Dohan et al., 2022) that facilitates efficient data acquisition in many-shot ICL (Agarwal et al., 2024).

The rest of this paper is structured as follows: §2 presents the proposed methodology, §3 reviews related work, §4 presents the experiments, and §5 provides concluding remarks.

2 QUANTIFYING UNCERTAINTY AND CALIBRATION

2.1 A Utilitarian Setup

The setup of this paper is as follows:

- we have a NLG task: given prompt $I \in \mathcal{I}$, generate a response $y' \in \mathcal{Y}$ where $\mathcal{Y}$ denotes the space of natural language responses;
- our utility can be quantified through a similarity

measure $S(y', y; I) \in \mathbb{R}$, which can be cheaply evaluated;

- we have an LM with predictive distribution p_M, and, in the absence of further evidence, consider p_M sufficiently trustworthy so that the ideal generation (i.e., *action*) y' for I should maximize the *expected utility*,

$$y' := \arg\max_{y' \in \mathcal{Y}'_I} \mathbb{E}_{y \sim p_M(\cdot|I)} S(y', y; I), \tag{1}$$

where $\mathcal{Y}'_I \subset \mathcal{Y}$ is the space of candidate generations.

In the above, p_M represents a subjective belief about the true data distribution. $\mathcal{Y}'_I \subset \mathcal{Y}$ can be determined based on computational constraints, e.g., as a separate set of samples from $p_M(\cdot \mid I)$.[1] The utility S determines a *risk* function $r(y', y; I) = -S(y', y; I)$, which can be scaled and shifted as desired.

As discussed in §1, this setup is very general. This decision-theoretic perspective is taken in the framework of minimum Bayes risk (MBR) decoding (Bickel and Doksum, 1997), which advocates for generating y based precisely on (1). MBR recovers many modern methods for natural language generation (Bertsch et al., 2023), which can be viewed as implicitly adopting the same view. Note that the decision-theoretic setup does not compel us to assume that p_M is strictly a Bayesian model (e.g., in the sense of Xie et al., 2021): although that will provide a good justification, it is also possible that a non-Bayesian LM better approximates our belief than any available Bayesian model does.

2.2 Task-Specific Measures of Subjective Uncertainty

It is well known that Bayesian uncertainty measures are connected to minimum achievable risks of the form

$$R_B(I) := \min_{y' \in \mathcal{Y}'_I} \mathbb{E}_{y \sim p_M(\cdot|I)} r(y', y; I). \tag{2}$$

For example, if $\mathcal{Y} = \mathcal{Y}'_I \subset \mathbb{R}$ and $r(y, y'; I) = (y - y')^2$, $R_B(I)$ will be equivalent to the predictive variance; if with an abuse of notation we redefine y' as a density function and $r(y', y; I) \leftarrow \log y'(y)$, $R_{B,\mathcal{Y}'_I}(I)$ will become the entropy. When the action space $\mathcal{Y}'_I \subsetneq \mathcal{Y}$ is a strict subset, (2) becomes connected to measures of "usable information" discussed in Xu et al. (2019).

It is thus natural to also adopt (2) as a measure of subjective uncertainty in NLG. By definitions, it represents the *task-specific* uncertainty we are interested in, and we recover measures of semantic uncertainty if we define r based on semantic equivalence. (2) can be implemented via Monte-Carlo estimation by drawing multiple samples $y \sim p_M$ from the LM.

A common generation strategy is defined by the Gibbs predictor, which corresponds to (1) with $\mathcal{Y}'_I := \{y_I\}$, where $y_I \sim p_M(\cdot \mid I)$ is a single random sample. In such cases, the expected uncertainty $\mathbb{E}_{\mathcal{Y}'_I} R_B(I)$ is connected to the degree statistic in Lin et al. (2024b, Eq. 8); see also Wang et al. (2022, p.8). The difference is that we use a similarity measure that we assume to define our utility. Compared with previous works (Kuhn et al., 2023; Lin et al., 2024b; Nikitin et al., 2024) which explored many uncertainty measures based on the aggregation of similarity measures, the decision-theoretic perspective highlights one unique method without any underspecified design choices, justifies its use beyond the scope of semantic equivalence, and shows that different choices of predictors ($\mathcal{Y}'_I$) should be matched with different uncertainty measures. Although being a measure of subjective uncertainty, it is not guaranteed to always outperform heuristic alternatives on downstream tasks, in particular if the LM is severely miscalibrated.

2.3 Evaluation of Task-Specific Calibration

Subjective uncertainty measures are most useful when the LM is reasonably calibrated. As discussed in Huang et al. (2024a), many previous works studying uncertainty in free-form NLG did not distinguish between calibration and predictive performance in their evaluation, which leads to a relative lack of understanding about the calibration in such settings.

A common definition of calibration is as follows (see e.g., Johnson et al., 2024): suppose the spaces $\mathcal{Y}, \mathcal{I}$ are discrete, and denote the true data distribution as p_0; a model p_M is considered calibrated w.r.t. some grouping function $G(I) : \mathcal{I} \to \mathcal{G}$ if we have

$$\begin{aligned} &\mathbb{E}_{p_0}(p_M(y = y' \mid I) \mid G(I) = g) \\ &= \mathbb{E}_{p_0}(p_0(y = y' \mid I) \mid G(I) = g) \quad \forall g \in \mathcal{G}, y' \in \mathcal{Y}. \end{aligned} \tag{3}$$

$G(I)$ can be defined based on uncertainty measures, or it may incorporate additional information such as input demographics. In NLG, however, (3) can be trivially violated if, e.g., the LM always generates a more verbose response compared with p_0, which may happen due to instruction tuning (Adlakha et al., 2023).

In light of the assumptions in §2.1, it is natural to relax (3) to require that for all $g \in \mathcal{G}, y' \in \mathcal{Y}'_I, s \in \mathbb{R}$

$$\begin{aligned} &\mathbb{E}(p_M(\{y : S(y', y; \mathcal{I}) = s\} \mid I) \mid G(I) = g) \\ &= \mathbb{E}(p_0(\{y : S(y', y; \mathcal{I}) = s\} \mid I) \mid G(I) = g) \end{aligned} \tag{3'}$$

(3') is equivalent to (3) if $\mathcal{Y}'_I = \mathcal{Y}$ and $S(y', y; \mathcal{I}) = \mathbf{1}\{y = y'\}$, or more generally if for all $I \in \mathcal{I}$ the map

[1] When $\mathcal{Y}'_I$ is stochastic, expectation quantities such as (1) should also be averaged over the randomness in $\mathcal{Y}'_I$. For brevity, we will omit all such (outmost) expectations.

from y to the function

$$s_{y,I} : \mathcal{Y}'_I \to \mathbb{R},\ y' \mapsto S(y', y; I)$$

is injective. (3') becomes weaker if S only depends on y and y' through a non-invertible function, e.g., a map from a natural language passage to its *semantics*. In such cases, it is clear that (3') remains sufficiently strong while focusing on task-specific calibration.

Remark 2.1. We can generally view the function $s_{y,I}(y')$ as a representation of the "task-specific semantics".[2] In fact, the replacement of $y' \in \mathcal{Y}'_I$ with $s_{y',I}(\cdot)$ has been advocated by Savage (1954), and $s_{y,I}$ is referred to as a *Savage act* (see e.g., Marinacci, 2015, §2.4).

A main reason we are interested in (3) or (3') is that they guarantee the calibration of uncertainty measures. It follows from (3') that $\mathbb{E}_{p_0(I|G(I)=g)p_M(y|I)} r(y', y; I) = \mathbb{E}_{p_0(I|G(I)=g)p_0(y|I)} r(y', y; I)$ for all y', and thus

$$\mathbb{E}_{p_0(I|G(I)=g)} R_B(I) = \mathbb{E}_{p_0(I|G(I)=g)p_0(y|I)} r(\hat{y}_I, y; I),$$

where $\hat{y}_I$ denotes the MBR generation (1). Plugging in $G(I) = R_B(I)$, we find that the calibration criterion (3') holds only if

$$\begin{aligned} &\mathrm{ECE}(p_M) := \mathbb{E}_s |f_M(s) - s| = 0, \quad \text{where} \\ &f_M(s) := \mathbb{E}_{I,y,\hat{y}_I}(r(\hat{y}_I, y; I) \mid R_B(I) = s) \end{aligned} \tag{4}$$

for all $s \in \mathbb{R}$. (4) generalizes the *expected calibration error* (ECE, Naeini et al., 2015) in classification, which it recovers with $r(y', y; I) \leftarrow \mathbf{1}\{y' \neq y\}$. As in the classification case, we can estimate (4) through histogram binning. The plot for f_M is called a *reliability diagram* (Murphy and Winkler, 1977) and provides insights on overconfidence ($f_M(s) < s$) or underconfidence. Importantly, these methods remain applicable in NLG without the need to introduce any binary notion of correctness. Moreover, they will automatically focus on the task-specific aspects of generation, as ECE and f_M only depend on model samples through the utility.

2.4 Representing and Eliciting Epistemic Uncertainty

We now turn to the quantification of epistemic uncertainty, using ICL as a key example. Epistemic uncertainty is the proportion of uncertainty that can be reduced given additional observations. It is particularly relevant when such observations are costly to obtain or incorporate into the generation process. In these scenarios, we can limit their use to queries with a higher level of *reducible* (i.e., epistemic) uncertainty.

[2] For the purposes of MBR generation and uncertainty quantification, it suffices to define the function $s_{y,I}$ on $\mathcal{Y}'_I$. For other purposes it may be helpful to extend its domain.

Epistemic uncertainty in ICL. In ICL the LM is prompted with $I = \langle z_{1:n}, x_* \rangle = \langle x_1, y_1, \ldots, x_n, y_n, x_* \rangle$, where $z_i := (x_i, y_i)$ are demonstration input-output pairs and x_* denotes the test query. n can be arbitrary. We assume that the user's utility does not depend on $z_{1:n}$ given x_*, so that the risk has the form $r(y, y'; I) \equiv r(y, y'; x_*)$. We propose the following measure of epistemic uncertainty, inspired by Fong et al. (2023); Xu and Raginsky (2022):

$$\begin{aligned} &\inf_{y' \in \mathcal{Y}'_I} \mathbb{E}_{p_x(x_{n+1:N}|I)} \mathbb{E}_{p_M(y_{n+1:N}|y_{1:n}, x_{1:N})} \\ &\qquad \mathbb{E}_{p_M(y_*|z_{1:N}, x_*)} r(y', y_*; x_*) \\ &- \mathbb{E}_{p_x(x_{n+1:N}|I)} \mathbb{E}_{p_M(y_{n+1:N}|y_{1:n}, x_{1:N})} \\ &\qquad \inf_{y' \in \mathcal{Y}'_I} \mathbb{E}_{p_M(y_*|z_{1:N}, x_*)} r(y', y_*; x_*). \end{aligned} \tag{5}$$

In the above, p_x denotes a user-specified distribution of inputs, $y_{n+1:N} \sim \prod_{j=n+1}^{N} p_M(y_j \mid z_{1:j-1}, x_j)$ follow the LM's predictive distribution, and $N > n$ is a hyperparameter. (5) can be estimated through autoregressive sampling and Monte-Carlo estimation; see §4.2 for implementation details.

Comparing with (2), we can see that (5) is the difference between the total uncertainty from two predictors; the second predictor has access to the additional demonstrations $z_{n+1:N} = (x_{n+1:N}, y_{n+1:N})$, where $y_{n+1:N}$ are model generations. Thus, (5) is a subjective measure of the *reducible uncertainty* given $x_{n+1:N}$. By convexity of the inf functional, (5) is always non-negative; so is its Monte-Carlo estimate as long as we use the same set of samples for its both terms.

Bayesian justifications. To further understand (5), consider a simplified setting as follows: suppose p_M is equivalent to a Bayesian model that assumes $z_{1:n}$ are i.i.d. conditional on a latent variable θ, $x_{n+1:N} \sim p_x(\cdot \mid I)$ are conditionally i.i.d., and $z_{n+1:N}$ are sufficiently informative so that the likelihood function $p(y = \cdot \mid x = x_*, \theta)$ determined by $\theta \mid z_{1:N}$ is identifiable given infinite samples. Then under mild technical conditions, as $N \to \infty$ (5) will become equivalent to

$$\begin{aligned} &\min_{y' \in \mathcal{Y}'_I} \mathbb{E}_{p_M(\theta|z_{1:n})} \mathbb{E}_{p(y|\theta, x=x_*)} r(y', y; x_*) - \\ &\qquad \mathbb{E}_{p_M(\theta|z_{1:n})} \min_{y' \in \mathcal{Y}'_I} \mathbb{E}_{p(y|\theta, x=x_*)} r(y', y; x_*). \end{aligned} \tag{5'}$$

The equivalence can be proved with the same idea as Fong et al. (2023); we provide a proof in App. A for completeness. (5') represents the amount of reducible risk should we have full knowledge about the distribution of $y \mid x = x_*$, which can be determined given either the latent θ or an infinite amount of missing data. As noted in Xu and Raginsky (2022), This is precisely the intuition behind Bayesian epistemic uncertainty, and we recover standard epistemic uncertainty measures (e.g., posterior variance for a regression mean, or mutual information) if we return to the standard choices of r such as the square or log loss.

Within the Bayesian framework, (5) with a finite choice of N will provide a lower bound for (5'): by convexity the second term will be larger. The intuition is that knowledge of the samples $z_{n+1:N}$ does not always enable the full reduction of risk. We can view the lower bound as an indication of the "true" epistemic uncertainty (5'), and it can be tightened if we choose $\{x_{n+i}\}$ to be more similar to x_*. The bound will also be tight for finite N if the uncertainty is "fully aleatoric" or "fully epistemic", i.e., if the risk at $x = x_*$ cannot be reduced by any amount of additional $z_{n+1:N}$ or if it can be fully reduced by a single z_{n+1}. Alternatively, we can observe that we do not always need the tightest bound for (5'): if all we can do is to collect $N - n$ real samples, the remaining uncertainty will be effectively irreducible.

As discussed in §1, this Bayesian view may be justified through a common belief that high-capacity LMs may approximate Bayesian inference. It is also supported by the results of Wen et al. (2022, §2.6, §4), which imply that in a multi-task setting, any p_M with a strong average-case performance for prediction will yield a similar value for (5). Falck et al. (2024) investigated this belief empirically on synthetic numerical datasets and found it to be valid for smaller choices of N. Taken together, these results suggest that the Bayesian perspective can be relevant when we restrict N to a regime where ICL is statistically efficient.

Connection to non-Bayesian methods. Independent to the above, Eq. (5) can also be understood through its connection to recent works. Johnson et al. (2024); Ahdritz et al. (2024) proposed methods that measure the correlation between consecutive predictions (y_{n+1}, y_*) for the same input $x_{n+1} = x_*$; as argued in Johnson et al. (2024), any learning system p_M with complete confidence should lead to conditionally i.i.d. (y_*, y_{n+1}) given I, and any p_M with a "fully epistemic" uncertainty should produce $y_* = y_{n+1}$. It is clear that (5) matches these behaviours at the extremes: it equals 0 in the former case, and in the latter case correctly indicate a fully reducible risk. The Bayesian perspective provides additional insights to the works of Johnson et al. (2024); Ahdritz et al. (2024). Eq. (5) generalizes their methodology by allowing for a wider range of choices for $\{x_{n+1:N}\}$, and the above discussion provides guidance on their choices.

Remark 2.2 (generalizations). (5) can be readily applied beyond ICL if we use $(x_{n+1:N}, y_{n+1:N})$ to represent more general queries, e.g., any clarifying questions and their answers (Hou et al., 2023). In such scenarios, (5) remains a non-negative measure of task-specific, reducible uncertainty. While we will focus on ICL in experiments, we note that this generalization can be useful when the real responses are more expensive to obtain compared to LM generations, for example when they require interaction with user (Li et al., 2023) or a third-party source (Min et al., 2024).

3 RELATED WORK

Probabilistic UQ for LMs. UQ has long been studied in LM applications involving a short or fixed-form response (e.g., Kamath et al., 2020; Jiang et al., 2021). Free-form generation introduces additional challenges. Our discussions are motivated by recent works on semantic uncertainty (Kuhn et al., 2023; Tian et al., 2023), which we seek to clarify and extend to a general scope. The measure (2) is a generalized entropy (DeGroot, 1962); similar to Xu et al. (2019, p.8), we deviate from that line of classical work by considering general loss functions and computational constraints.

All methods reviewed so far can be viewed as adopting a *probabilistic* perspective: they quantify uncertainty through the variation in the LM's predictive distribution. Other approaches to UQ include probing the internal representation of LMs (Lin et al., 2024a; Orgad et al., 2024), and prompting the LM to reason about its uncertainty (Xiao and Wang, 2021; Kadavath et al., 2022). Orthogonal to these developments, conformal prediction methods (e.g., Quach et al., 2023; Mohri and Hashimoto, 2024; Yadkori et al., 2024b) take a prespecified uncertainty measure and a calibration set as input and provide a predictor with coverage guarantees.

Calibration in free-form generation. Band et al. (2024) studied calibration in scenarios where the free-form generation y is used for downstream classification based on a known predictor $\phi(y)$. Our work generalizes their setup, as we can always define a utility for y based on ϕ and the standard utility functions for classification. Huang et al. (2024a) studied the *rank*-calibration of general uncertainty measures which, as discussed in their work, is a different property. We note that an LM could be arbitrarily overconfident or underconfident for the uncertainty measure (2) to remain rank-calibrated. Huang et al. (2024b) proposed several calibration metrics based on a similarity measure. In their work both the similarity measure and the method to aggregate calibration error are left underspecified, rendering the resulted metrics hard to interpret; in contrast, our Eq. (4) admits a straightforward interpretation as a generalized ECE.

Prompt-based methods, instruction-tuned LMs. A recent line of work proposes to quantify uncertainty through prompting, e.g., by asking the LM to output the probability that its generation is correct (Xiao and Wang, 2021; Mielke et al., 2022; Ren et al., 2023) or to answer a multiple-choice question (Kadavath et al., 2022). Compared with the probabilistic methods, this approach can be more efficient computationally as it avoids the need to draw multiple samples for each query. However, it requires the LM to have stronger reasoning capabilities, and is more difficult to apply when the user's utility involves factors beyond factual correctness or for the quantification of epistemic uncertainty.

Still, within the above scope, it is reasonable to ask whether the verbalized approach might be preferable for instruction-tuned LMs. The question remains largely open: previous studies have found that instruction tuning can affect probabilistic calibration (Achiam et al., 2023; Tian et al., 2023), but empirical comparisons between probabilistic and verbalized approaches yielded inconsistent results (Kuhn et al., 2023; Tian et al., 2023; Xiong et al., 2024). Moreover, in the free-form setting there is also a lack of principled evaluation of calibration: many prior works (Kuhn et al., 2023; Lin et al., 2024b) adopted evaluation metrics that conflate calibration with predictive performance (Huang et al., 2024a, p.5), and some studies (Jiang et al., 2021; Tian et al., 2023) evaluated token-level (rather than task-specific) calibration, which is less relevant. Our experiments in §4.1 will fill in this gap.

Epistemic uncertainty. Our method is inspired by Fong et al. (2023); Xu and Raginsky (2022). Similar ideas were discussed by Berti et al. (2021); Wen et al. (2022) in relation to Fong et al. (2023), and Grünwald and Dawid (2004); Kotelevskii and Panov (2024) in relation to Xu and Raginsky (2022). Similar notions of reducible uncertainty also appear in active learning (Smith et al., 2023), and in experiment design (Rainforth et al., 2024) in relation to Remark 2.2. All of these works either restricted to the estimation of a full data distribution, or required explicitly Bayesian models,[3] rendering them inapplicable to general LM applications. Falck et al. (2024) applied the method of Fong et al. (2023) to ICL, but similarly restricted to density estimation for numerical data and approximately Bayesian models. Hou et al. (2023) proposed to measure the change in log likelihood when clarifying questions are presented to the LMs; concurrent work of Yadkori et al. (2024a) also proposed a measure similar to (5), but based on the log loss. These methods thus cannot distinguish between task-specific and irrelevant uncertainties, and also requires access to the LM's predictive density; our work avoids such limitations.

[3] i.e., a p_M that corresponds to a Bayesian posterior predictive, defined by a known posterior over model parameters and a likelihood.

4 EXPERIMENTS

4.1 Uncertainty in Question Answering

We first illustrate the uncertainty and calibration measures by revisiting the question answering (QA) experiments in recent works (Lin et al., 2024b; Tian et al., 2023). As discussed in §3, our goal is to determine whether probabilistic UQ methods are suitable for modern instruction-tuned LMs, by providing a principled evaluation of task-specific calibration.

Experiment setup. We combine the QA datasets studied by Lin et al. (2024b); Tian et al. (2023): `CoQA` (Reddy et al., 2019), `TriviaQA` (Joshi et al., 2017), `NQOpen` (Kwiatkowski et al., 2019), `SciQ` (Welbl et al., 2017), and `TruthfulQA` (Lin et al., 2021). We subsample 1000 queries for each dataset and report bootstrap confidence intervals (CIs) for all metrics. The utility $S(y', y; I)$ is defined by instructing an LM to rate the consistency of y' with y as an answer to the user's question. Our setup generally follows Lin et al. (2024b), except that we added a system prompt instructing LMs to generate single-line responses and switched to `gpt-4o-mini` for evaluation.[4] As `TriviaQA` and `NQOpen` contain trivia questions that were annotated well before the knowledge cutoff date of current LMs, we further remove a subset of questions with clearly outdated answers; see App. B.1 for details.

We use LMs from the following model families: `gpt-4o` (OpenAI, 2024), `llama-3.1` (Dubey et al., 2024), `gemini-1.5` (Gemini Team, 2024), and `claude-3.5` (Anthropic, 2024). All LMs are instruction tuned unless otherwise noted. In a subset of experiments, we compare the uncertainty measure (2) (denoted as `Prob.`) with the prompt-based methods of Kadavath et al. (2022, denoted as `P(True)`) and Tian et al. (2023, `Verb.`). We apply all methods to the Gibbs predictor, and estimate uncertainty measures using Monte-Carlo estimation with 10 samples for each query. To evaluate calibration, we use the ECE metric (4) estimated through histogram binning (Naeini et al., 2015). For reference, we also compute the average test utility to quantify predictive performance, as well as the AURAC metric from Nadeem et al. (2009); Lin et al. (2024b), which measures the performance of a selective prediction procedure based on a given LM and uncertainty measure. Full setup details are deferred to App. B.1.

[4] Lin et al. (2024b) verified that LM-based evaluation aligns with human perception in these tasks when using `gpt-3.5`; we replicated their analysis for `gpt-4o-mini`. We also confirmed that except on `SciQ` the generations are varied in form. `SciQ` is a short-form dataset which we include for consistency with Tian et al. (2023).

Results and discussion. We defer full results to App. B.2 and summarize the main findings below.

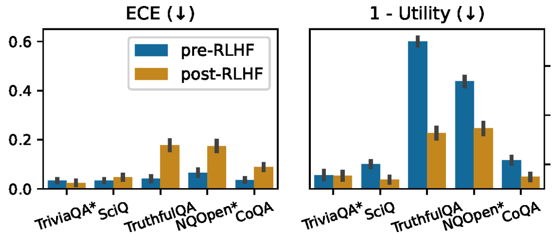


Figure 1: Question answering: ECE and utility for the `llama-3.1-70b` model before and after instruction tuning. Error bar denotes 95% bootstrap CI.

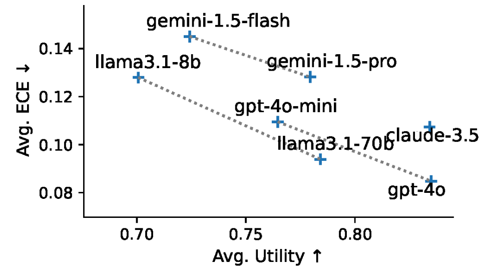


Figure 2: Question answering: test utility vs ECE for isntruction-tuned LMs, averaged over the datasets in Lin et al. (2024b). App. B.2 presents results for individual datasets.

How does instruction tuning affect calibration? We revisit an experiment in Tian et al. (2023) on the impact of instruction tuning on *token-level* calibration. Fig. 1 plots the ECE and test utility of `llama-3.1-70b` before and after instruction tuning. Its left subplot can be compared with Tian et al. (2023, Fig. 1 left, for the first three datasets). The comparison reveals a rather different phenomenon: whereas the calibration of token-level probability deteriorated significantly (an increase in ECE by 0.1–0.2) across all three datasets in Tian et al. (2023), we find the deterioration of *task-specific* calibration to be much smaller, except on `TruthfulQA`. It is interesting to note that among these datasets, `TruthfulQA` is the only one where instruction-tuning led to a significant improvement in utility; additional results on `NQOpen` and `CoQA` further confirm this trend. These results indicate that *the impact of instruction tuning on calibration may have been overestimated* in previous work. Results for AUARC are provided in App. B.2, where we also evaluate additional LMs. We find that across all datasets, the performance of selective prediction based on task-specific uncertainty is either approximately unchanged or improved after instruction tuning. This is similarly in contrast to Tian et al. (2023, Fig. 1 right), which shows that selective prediction based on token-level uncertainty consistently performs worse after instruction tuning.

We now compare the calibration of instruction-tuned LMs with their predictive performance. As shown in Fig. 2, calibration is generally correlated with predictive performance, but substantial variation remains. It is thus reasonable to expect that the calibration of many instruction-tuned LMs may still be improved with different tuning methodologies.

Comparing probabilistic and prompt-based uncertainty. We compare the calibration of the uncertainty measure (2) with prompt-based alternatives, for which we can define an analogous ECE metric. Fig. 3 visualizes ECE from different methods. We can see that the measure (2) generally leads to the best calibration error; the only exception is on `TruthfulQA`, which is "adversarially constructed" to evaluate LMs' tendency to repeat human falsehoods. Reliability diagrams in App. B.2 provide further insights. Taken together, the results suggest that *the probabilistic approach to UQ remains effective for modern instruction-tuned LMs.*

Note that in the above, only the probabilistic uncertainty measure makes use of the utility S. It is thus less applicable in scenarios where we do not have a good surrogate of the utility. However, as discussed before, in many tasks this is a reasonable assumption. Moreover, the ability of the probabilistic approach to incorporate different choices of utility can also be an advantage, especially when utility cannot be interpreted as a probability of correctness.

4.2 Epistemic Uncertainty in ICL

We now turn to epistemic uncertainty (EU) quantification in ICL, and investigate whether the proposed EU measure may provide an effective deferral strategy that routes selected queries to a more expensive predictor.

Background and setup. LM generations can often be improved given information retrieved from an external source. Such a process can become expensive if retrieval is from a third party (Min et al., 2024) or if it leads to a very long prompt (Agarwal et al., 2024). In such scenarios, it would be desirable to limit retrieval to queries with a higher level of reducible uncertainty.

Agarwal et al. (2024) showed that many-shot ICL can be effective in low-resource machine translation (Haddow et al., 2022). Informed by their findings, we evaluate the use of EU for choosing between a LM predictor given few-shot demonstrations and a more expensive predictor based on many-shot ICL. We compare this method (denoted as `EU`) to a `random` baseline, and an alternative based on total uncertainty (`TU`, Eq. (2)).

We follow the setup in Agarwal et al. (2024) and adopt

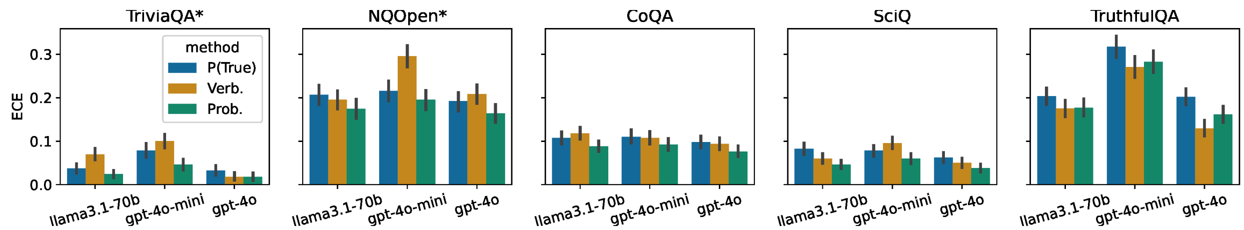


Figure 3: Question answering: calibration error from different methods. Error bar denotes 95% bootstrap CI.

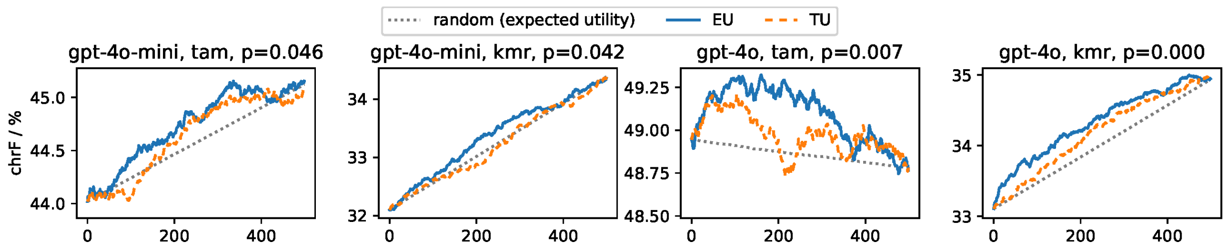


Figure 4: Machine translation: average utility vs the number of deferrals to the many-shot predictor. We also report the p-value of a permutation test that compares the AUC-DF from the `EU` method to `random`.

the chrF score (Popović, 2015) as the utility. chrF is a lexical similarity score that measure both semantic and syntactic differences; the latter provides a source of irreducible uncertainty. We define a base predictor using ICL with $n = 4$ shots, and a more expensive predictor with $n' = 128$. We defer to the expensive predictor on m test queries with the highest EU, and report the average utility (over all test samples) w.r.t. m. Following Gupta et al. (2024), we refer to this function as a *deferral curve*, and use its integral (AUC-DF) as a scalar summary.

The EU measure (5) is instantiated with $N-n=4$ and estimated as follows: for a query $I = \langle z_{1:n}, x_* \rangle$, we first draw l i.i.d. samples $\{z^{(i)}_{n+1:N} = (x^{(i)}_{n+1:N}, y^{(i)}_{n+1:N})\}_{i=1}^{l}$ where $x^{(i)}_{n+1:N} \sim p_x(\cdot|I)$ and $y^{(i)}_{n+1:N} \sim \prod_{k=n+1}^{N} p_M(y_k \mid \langle z_{1:n}, z^{(i)}_{n+1:k-1}, x^{(i)}_k \rangle)$. Then, for each i, we sample $\{y^{(i,j)}_*\}_{j=1}^{m} \sim p_M(y_* \mid \langle z_{1:n}, z^{(i)}_{n+1:N}, x_* \rangle)$. We use $l(m-1)$ samples to define a plug-in Monte-Carlo estimator, and the remaining samples, $\{y^{(i,m)}_*\}_{i=1}^{l}$, to define the action space in (5). We define p_x by using an LM to rewrite x_*, and choose $l=5, m=8$.

Following Agarwal et al. (2024), we adopt the translation tasks from English to Kurdish (`kmr`) and Tamil (`tam`) in the FLORES+ dataset (NLLB Team et al., 2022). Due to resource limitations, we subsample 500 data points from each dataset. We consider the following LMs: `gpt-4o-mini`, `gpt-4o` and `gemini-1.5-pro`. Full setup details are deferred to Appendix B.3.

Result and discussion. Fig. 4 plots the deferral curves from the GPT models. We can see that the `EU` method consistently outperforms the `random` baseline, and the improvement in AUC-DF is statistically significant. In contrast, the `TU` baseline is not always better than random, which is expected since a larger number of in-context demonstrations will not help if predictive uncertainty is irreducible. To further compare between the `EU` and `TU` methods, we compute the bootstrap distributions of their AUC-DF metrics, which are obtained by resampling the test queries without replacement. As shown in Table 1, for all but one experiment (`gemini-1.5-pro` on `kmr`) `EU` outperforms `TU` in more than 75% of bootstrap simulations.

Table 1: Machine translation: AUC-DF difference between `EU` and `TU` methods. We report the median, 25% and 75% percentiles across 1000 bootstrap simulations.

	gemini-1.5-pro	gpt-4o-mini	gpt-4o
tam	0.05 [0.00, 0.09]	0.09 [0.01, 0.16]	0.12 [0.06, 0.19]
kmr	0.02 [−0.02, 0.06]	0.15 [0.10, 0.21]	0.15 [0.08, 0.22]

In the `gemini-1.5-pro` experiments, both `EU` and `TU` demonstrate less improvement over the random baseline. This may be attributable to the fact that uncertainty from `gemini-1.5-pro` is less calibrated: the respective base predictor achieves an average ECE of 0.303 on the two datasets, compared with 0.104 using `gpt-4o-mini` and 0.083 using `gpt-4o`. Nonetheless, `EU` still appears to be the more effective method, as indicated by Table 1. See App. B.4 for full results.

5 CONCLUSION

This work studies uncertainty quantification in free-form natural language generation. Through a decision-theoretic perspective we derived principled methods to quantify the model's task-specific subjective uncertainty and to evaluate its calibration. These methods are also connected to previous work, which provides additional justification for their adoption. While the discussions are not necessarily novel outside the context of language modelling, within this domain the decision-theoretic perspective appears new and addresses important conceptual challenges in uncertainty quantification. Experiments demonstrated the practical utility of the proposed methods.

Acknowledgements

Ziyu Wang contributed to this work while at the University of Oxford and is now affiliated with Citadel LLC. Z. W. acknowledges support from Novo Nordisk during his time at the University of Oxford. Chris Holmes acknowledges support from the Medical Research Council Programme Leaders award MC_UP_A390_1107, The Alan Turing Institute, Health Data Research, U.K., and the U.K. Engineering and Physical Sciences Research Council through the Bayes4Health programme grant. This work is also supported by the Google Cloud Research Credits program (GCP339294252).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774.*

Adlakha, V., BehnamGhader, P., Lu, X. H., Meade, N., and Reddy, S. (2023). Evaluating correctness and faithfulness of instruction-following models for question answering. *arXiv preprint arXiv:2307.16877.*

Agarwal, R., Singh, A., Zhang, L. M., Bohnet, B., Chan, S., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., Behbahani, F., Faust, A., and Larochelle, H. (2024). Many-Shot In-Context Learning. arXiv:2404.11018 [cs].

Ahdritz, G., Qin, T., Vyas, N., Barak, B., and Edelman, B. L. (2024). Distinguishing the Knowable from the Unknowable with Language Models. arXiv:2402.03563 [cs].

Aichberger, L., Schweighofer, K., Ielanskyi, M., and Hochreiter, S. (2024). How many Opinions does your LLM have? Improving Uncertainty Estimation in NLG.

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. (2022). What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661.*

Anthropic (2024). Claude 3.5 sonnet model card addendum.

Band, N., Li, X., Ma, T., and Hashimoto, T. (2024). Linguistic Calibration of Long-Form Generations. In *International Conference on Machine Learning.*

Bannur, S., Bouzid, K., Castro, D. C., Schwaighofer, A., Bond-Taylor, S., Ilse, M., P'erez-Garc'ia, F., Salvatelli, V., Sharma, H., Meissen, F., Ranjit, M. P., Srivastav, S., Gong, J., Falck, F., Oktay, O., Thieme, A., Lungren, M. P., Wetscherek, M. T. A., Alvarez-Valle, J., and Hyland, S. L. (2024). Maira-2: Grounded radiology report generation. *arXiv*, abs/2406.04449.

Berti, P., Dreassi, E., Pratelli, L., and Rigo, P. (2021). A class of models for Bayesian predictive inference. *Bernoulli*, 27(1).

Bertsch, A., Xie, A., Neubig, G., and Gormley, M. R. (2023). It's MBR All the Way Down: Modern Generation Techniques Through the Lens of Minimum Bayes Risk. arXiv:2310.01387 [cs].

Bickel, P. J. and Doksum, K. A. (1997). *Mathematical statistics: basic ideas and selected topics.* Holden-Day Inc.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2024). JAX: composable transformations of Python+NumPy programs.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

DeGroot, M. H. (1962). Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419.

Der Kiureghian, A. and Ditlevsen, O. (2009). Aleatory or Epistemic? Does it Matter? *Structural safety*, 31(2):105–112.

Dohan, D., Xu, W., Lewkowycz, A., Austin, J., Bieber, D., Lopes, R. G., Wu, Y., Michalewski, H., Saurous, R. A., Sohl-Dickstein, J., et al. (2022). Language model cascades. *arXiv preprint arXiv:2207.10342.*

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783.*

Falck, F., Wang, Z., and Holmes, C. C. (2024). Are large language models bayesian? a martingale perspective on in-context learning. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Fong, E., Holmes, C., and Walker, S. G. (2023). Martingale posterior distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1357–1391. arXiv:2103.15671 [stat].

Gemini Team (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Gemma Team (2024). Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4).

Gupta, N., Narasimhan, H., Jitkrittum, W., Rawat, A. S., Menon, A. K., and Kumar, S. (2024). Language model cascades: Token-level uncertainty and beyond. *arXiv preprint arXiv:2404.10136*.

Haddow, B., Bawden, R., Barone, A. V. M., Helcl, J., and Birch, A. (2022). Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Hahn, M. and Goyal, N. (2023). A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*.

Hou, B., Liu, Y., Qian, K., Andreas, J., Chang, S., and Zhang, Y. (2023). Decomposing uncertainty for large language models through input clarification ensembling. *arXiv preprint arXiv:2311.08718*.

Huang, X., Li, S., Yu, M., Sesia, M., Hassani, H., Lee, I., Bastani, O., and Dobriban, E. (2024a). Uncertainty in Language Models: Assessment through Rank-Calibration. arXiv:2404.03163 [cs, stat].

Huang, Y., Liu, Y., Thirukovalluru, R., Cohan, A., and Dhingra, B. (2024b). Calibrating Long-form Generations from Large Language Models. arXiv:2402.06544 [cs].

Jiang, Z., Araki, J., Ding, H., and Neubig, G. (2021). How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Jin, D., Jin, Z., Hu, Z., Vechtomova, O., and Mihalcea, R. (2022). Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Johnson, D. D., Tarlow, D., Duvenaud, D., and Maddison, C. J. (2024). Experts Don't Cheat: Learning What You Don't Know By Predicting Pairs. arXiv:2402.08733 [cs].

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. (2022). Language Models (Mostly) Know What They Know. arXiv:2207.05221 [cs].

Kamath, A., Jia, R., and Liang, P. (2020). Selective question answering under domain shift. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Kendall, A. and Gal, Y. (2017). What Uncertainties do we Need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Kotelevskii, N. and Panov, M. (2024). Predictive Uncertainty Quantification via Risk Decompositions for Strictly Proper Scoring Rules. arXiv:2402.10727 [cs, stat].

Kuhn, L., Gal, Y., and Farquhar, S. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. arXiv:2302.09664 [cs].

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Li, B. Z., Tamkin, A., Goodman, N., and Andreas, J. (2023). Eliciting Human Preferences with Language Models. arXiv:2310.11589 [cs].

Lin, S., Hilton, J., and Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Lin, Z., Trivedi, S., and Sun, J. (2024a). Contextualized sequence likelihood: Enhanced confidence scores for natural language generation. *arXiv preprint arXiv:2406.01806*.

Lin, Z., Trivedi, S., and Sun, J. (2024b). Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Madras, D., Pitassi, T., and Zemel, R. (2018). Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31.

Marinacci, M. (2015). Model uncertainty. *Journal of the European Economic Association*, 13(6):1022–1100.

Mielke, S. J., Szlam, A., Dinan, E., and Boureau, Y.-L. (2022). Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Min, S., Gururangan, S., Wallace, E., Shi, W., Hajishirzi, H., Smith, N. A., and Zettlemoyer, L. (2024). SILO language models: Isolating legal risk in a nonparametric datastore. In *The Twelfth International Conference on Learning Representations*.

Mohri, C. and Hashimoto, T. (2024). Language Models with Conformal Factuality Guarantees. arXiv:2402.10978 [cs].

Murphy, A. H. and Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 26(1):41–47.

Nadeem, M. S. A., Zucker, J.-D., and Hanczar, B. (2009). Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In *Machine Learning in Systems Biology*, pages 65–81. PMLR.

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Nikitin, A., Kossen, J., Gal, Y., and Marttinen, P. (2024). Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *arXiv preprint arXiv:2405.20003*.

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Mejia-Gonzalez, G., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No language left behind: Scaling human-centered machine translation.

OpenAI (2024). Gpt-4o system card. https://cdn.openai.com/gpt-4o-system-card.pdf.

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., and Belinkov, Y. (2024). Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Post, M. (2018). A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. (2023). Conformal Language Modeling. arXiv:2306.10193 [cs].

Rainforth, T., Foster, A., Ivanova, D. R., and Bickford Smith, F. (2024). Modern bayesian experimental design. *Statistical Science*, 39(1):100–114.

Reddy, S., Chen, D., and Manning, C. D. (2019). Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Ren, J., Zhao, Y., Vu, T., Liu, P. J., and Lakshminarayanan, B. (2023). Self-evaluation improves selective generation in large language models. In Antorán, J., Blaas, A., Buchanan, K., Feng, F., Fortuin, V., Ghalebikesabi, S., Kriegler, A., Mason, I., Rohde, D., Ruiz, F. J. R., Uelwer, T., Xie, Y., and Yang, R., editors, *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 49–64. PMLR.

Robinson, N. R., Ogayo, P., Mortensen, D. R., and Neubig, G. (2023). Chatgpt mt: Competitive for high- (but not low-) resource languages. *arXiv preprint arXiv:2309.07423*.

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley.

Smith, F. B., Kirsch, A., Farquhar, S., Gal, Y., Foster, A., and Rainforth, T. (2023). Prediction-oriented bayesian active learning. In *International Conference on Artificial Intelligence and Statistics*, pages 7331–7348. PMLR.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. (2023). Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., Xu, J., Qu, J., and Zhou, J. (2023). Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Welbl, J., Liu, N. F., and Gardner, M. (2017). Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

Wen, Z., Osband, I., Qin, C., Lu, X., Ibrahimi, M., Dwaracherla, V., Asghari, M., and Van Roy, B. (2022). From Predictions to Decisions: The Importance of Joint Predictive Distributions. arXiv:2107.09224 [cs, stat].

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Xiao, Y. and Wang, W. Y. (2021). On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. (2021). An explanation of in-context learning as implicit Bayesian inference. *arXiv preprint arXiv:2111.02080*.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. (2024). Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *International Conference on Learning Representations*. arXiv. arXiv:2306.13063 [cs].

Xu, A. and Raginsky, M. (2022). Minimum Excess Risk in Bayesian learning. *IEEE Transactions on Information Theory*, 68(12):7935–7955.

Xu, Y., Zhao, S., Song, J., Stewart, R., and Ermon, S. (2019). A theory of usable information under computational constraints. In *International Conference on Learning Representations*.

Yadkori, Y. A., Kuzborskij, I., György, A., and Szepesvári, C. (2024a). To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*.

Yadkori, Y. A., Kuzborskij, I., Stutz, D., György, A., Fisch, A., Doucet, A., Beloshapka, I., Weng, W.-H., Yang, Y.-Y., Szepesvári, C., et al. (2024b). Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*.

Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., Fonseca, E. K. U. N., Lee, H. M. H., Abad, Z. S. H., Ng, A. Y., et al. (2023). Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Zhang, J., Li, Z., Das, K., Malin, B. A., and Kumar, S. (2023). SAC3: Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency. arXiv. arXiv:2311.01740 [cs].

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See App. A.
   (b) Complete proofs of all theoretical results. [Yes] See App. A.
   (c) Clear explanations of any assumptions. [Yes] See App. A.
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] See the supplemental material.

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] See figure and table captions, and App. B.

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] See App. B.1.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes] See App. B.

   (b) The license information of the assets, if applicable. [Yes] See App. B.

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

On Subjective Uncertainty Quantification and Calibration in Natural Language Generation Supplementary Materials

A DEFERRED PROOF

Claim A.1. *Suppose $\mathcal{X}, \mathcal{Y}$ are finite sets, and $\sup_{y',y,x} |r(y', y; x)| < \infty$. Suppose there exists a Bayesian model with prior π over a discrete parameter $\theta \in \Theta$ and a (conditional) likelihood function $p(y \mid x, \theta)$, s.t. for all $(n, x_{1:n}, y_{1:n}, x_*)$ we have*

$$p_M(y_{n+1} = \cdot \mid x_{n+1} = x_*, x_{1:n}, y_{1:n}) = \int \pi(d\theta \mid x_{1:n}, y_{1:n}) p(y = \cdot \mid x = x_*, \theta)$$

where $\pi(d\theta \mid x_{1:n}, y_{1:n}) \propto \pi(d\theta) \prod_{i=1}^n p(y = y_i \mid x = x_i, \theta)$ denotes the parameter posterior. Suppose (5) is defined using $x_{n+1:N} \overset{i.i.d.}{\sim} p_{x,1}$, and that for all $x_* \in \mathcal{X}$ and π-a.e. θ we have

$$\lim_{n\to\infty} \mathbb{E}_{x_{1:n}\sim p_{x,1}, y_i \sim p(y=\cdot|x=x_i,\theta)} \|p_M(y_{n+1} = \cdot \mid x_{n+1} = x_*, x_{1:n}, y_{1:n}) - p(y = \cdot \mid x = x_*, \theta)\|^2_{\ell_2(\mathcal{Y})} = 0, \tag{6}$$

where $\|f\|_{\ell_2(\mathcal{Y})} := \sqrt{\sum_{y\in\mathcal{Y}} f(y)^2}$ denotes the ℓ_2 norm. Then for $p_{x,1}$-a.e. $(x_{1:n}, x_*)$ and p_M-a.e. $y_{1:n}$, the $n \to \infty$ limit of (5) is equivalent to (5').

In the above, (6) is the assumed identifiability condition for the likelihood function. We expect the numerous technical restrictions to be relaxable, but refrain from a more general proof for brevity. The main conditions that enable the equivalence are that *(i)* p_M can be augmented to define an exchangeable (or c.i.d.) model for (x, y), and *(ii)* the likelihood function is identifiable.

Proof. We will show that each of the two terms in (5) and (5') are equivalent.

Let us introduce the following notations: define $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ and, for any x_i, y_i, $z_i := (x_i, y_i) \in \mathcal{Z}$. For all $n \geq 0$, define $\bar{p}_M(z_{n+1} = (x, y) \mid z_{1:n}) = p_{x,1}(x) p_M(y_{n+1} = y \mid x_{n+1} = x, x_{1:n}, y_{1:n}), \bar{p}(z = (x, y) \mid \theta) = p_{x,1}(x) p(y \mid \theta, x)$. Clearly, $\bar{p}_M$ is equivalent to a Bayesian model with prior π and the factorized likelihood $\bar{p}$. Thus, $z_{n+1:N+1} \sim \bar{p}_M(\cdot \mid z_{1:n})$ are exchangeable, and we have, for all $n, z_{1:n}, N > n, y' \in \mathcal{Y}$ and $p_{x,1}$-almost every x_*,

$$\mathbb{E}_{\bar{p}_M}(r(y', y_{N+1}; x_*) \mid z_{1:n}, x_{N+1} = x_*) = \mathbb{E}_{\bar{p}_M}(r(y', y_{n+1}; x_*) \mid z_{1:n}, x_{n+1} = x_*).$$

By definitions, the terms above equal the first term in (5) and (5'), respectively, with their outmost infimum removed. Retaking infimum proves the equivalence of the first terms.

For the second term, observe that by the boundedness of r and the discreteness of $\mathcal{Y}$, for all $x_* \in \mathcal{X}$ the function $R_{x_*}(\rho) := \inf_{y' \in \mathcal{Y}'_I} \mathbb{E}_{y\sim\rho} r(y', y; x_*)$ is Lipschitz continuous w.r.t. the $\ell_2(\mathcal{Y})$ norm.[5] Thus, for π-a.e. θ we have

$$\lim_{N\to\infty} \mathbb{E}_{z_{1:N-n} \overset{i.i.d.}{\sim} \bar{p}(z|\theta)} |R_{x_*}(p_M(y = \cdot \mid x = x_*, z_{1:N-n})) - R_{x_*}(p(y = \cdot \mid x = x_*, \theta))| = 0$$

following (6). It follows by the boundedness of $R_{x_*}(\cdot)$ and the dominated convergence theorem that

$$\lim_{N\to\infty} \mathbb{E}_{\theta\sim\pi, z_{1:N-n}\sim\bar{p}(z|\theta)} R_{x_*}(p_M(y = \cdot \mid x = x_*, z_{1:N-n})) = \mathbb{E}_{\theta\sim\pi} R_{x_*}(p(y = \cdot \mid x = x_*, \theta)). \tag{7}$$

[5] Note that any distribution ρ over $\mathcal{Y}$ can be identified with a probability mass function in $\ell_2(\mathcal{Y})$ since $\mathcal{Y}$ is discrete.

The RHS above equals the expectation of the second term in (5') over $z_{1:n} \sim \bar{p}_M$. Denote the second term in (5) as $u_N(z_{1:n})$. By definition of $\bar{p}_M$ and its exchangeability, as well as the convexity of $R_{x_*}(\cdot)$, we have

$$\begin{aligned}
\mathbb{E}_{\bar{p}_M(z_{1:n})} u_N(z_{1:n}) &= \mathbb{E}_{\bar{p}_M(z_{1:n})} \mathbb{E}_{\bar{p}_M(z_{n+1:N} \mid z_{1:n})} R_{x_*}(\bar{p}_M(y_{N+1} = \cdot \mid x_{N+1} = x_*, z_{1:N})) \\
&\le \mathbb{E}_{\bar{p}_M(z_{n+1:N})} R_{x_*}(\mathbb{E}_{\bar{p}_M(z_{1:n} \mid z_{n+1:N})} \bar{p}_M(y_{N+1} = \cdot \mid x_{N+1} = x_*, z_{1:N})) \\
&= \mathbb{E}_{\bar{p}_M(z_{1:N-n})} R_{x_*}(\bar{p}_M(y_{N-n+1} = \cdot \mid x_{N-n+1} = x_*, z_{1:N-n})) \\
&= \mathbb{E}_{\theta \sim \pi, z_{1:N-n} \sim \bar{p}(z \mid \theta)} R_{x_*}(p_M(y = \cdot \mid x = x_*, z_{1:N-n})).
\end{aligned}$$

Combining with (7) and applying the dominated convergence theorem yield

$$\mathbb{E}_{\bar{p}_M(z_{1:n})} \lim_{N \to \infty} u_N(z_{1:n}) = \lim_{N \to \infty} \mathbb{E}_{\bar{p}_M(z_{1:n})} u_N(z_{1:n}) \le \mathbb{E}_{\bar{p}_M(z_{1:n})} \mathbb{E}_{\pi(\theta \mid z_{1:n})} R_{x_*}(p(y = \cdot \mid x = x_*, \theta)). \tag{8}$$

On the other hand, we have

$$\begin{aligned}
u_N(z_{1:n}) &= \mathbb{E}_{\bar{p}_M(z_{n+1:N} \mid z_{1:n})} R_{x_*}(\bar{p}_M(y = \cdot \mid x = x_*, z_{1:N})) \\
&= \mathbb{E}_{\bar{p}_M(z_{n+1:N} \mid z_{1:n})} R_{x_*}(\mathbb{E}_{\pi(\theta \mid z_{1:N})} p(y = \cdot \mid x = x_*, \theta)) \\
&\ge \mathbb{E}_{\bar{p}_M(z_{n+1:N} \mid z_{1:n})} \mathbb{E}_{\pi(\theta \mid z_{1:N})} R_{x_*}(p(y = \cdot \mid x = x_*, \theta)) \\
&= \mathbb{E}_{\pi(\theta \mid z_{1:n})} R_{x_*}(p(y = \cdot \mid x = x_*, \theta)),
\end{aligned}$$

so the same holds for $\lim_{N \to \infty} u_N(z_{1:n})$. Comparing with (8) we find that equality must hold for $\bar{p}_M$-a.e. $z_{1:n}$. This proves the equivalence for the second terms in (5) and (5'), and consequently the original claim. □

B EXPERIMENTAL DETAILS

Code for the experiments are available in the supplemental material. It is based on the code release of Lin et al. (2024b, MIT License) and the following main software libraries: `scikit-learn` (Pedregosa et al., 2011, BSD 3-Clause License), `pandas` (Wes McKinney, 2010, BSD 3-Clause License) and `jax` (Bradbury et al., 2024, Apache-2.0 License).

B.1 Question Answering: Deferred Setup Details

We adopt the prompts in Lin et al. (2024b) for generation and evaluation; see their Appendix B. For evaluation, their prompt instructs the LM to "rate the level of consistency" between a provided answer and a reference based on in-context demonstrations. For generation, we add a system prompt that instructs the LM to generate single-line responses to avoid interference with the evaluation prompt template in Lin et al. (2024b); we find that without the system prompt, the more recent LMs we have evaluated have a greater tendency to generate multi-line responses. The generations can still contain multiple sentence, and we find them to be varied in format and content.

The `NQOpen` and `TriviaQA` datasets contain a number of outdated questions, for example about the current holder of a public office or the latest season of an ongoing TV series. We thus use an LM (`gpt-4o`) to identify and remove such questions. We provide the LM with a question and its (possibly outdated) reference answer, and asks it to evaluate whether the question may fall into the above categories; the LM is asked to deprioritize its internal knowledge over the reference answer during evaluation. The prompt is shown in Fig. 5. We note that *the filtering process does not affect any finding reported in the main text*; see e.g. Fig. 6 (a) and (b). we chose to apply this process so that the reported results more accurately reflect the LMs' performance on these tasks.

We access all LMs through API services, from OpenAI (for GPT models) or Google Cloud. The API versions are as follows: `gpt-4o-2024-08-06`, `gpt-4o-mini-2024-07-18`, `gemini-1.5-flash-001`, `gemini-1.5-pro-001`, `claude-3.5-sonnet@20240620` (versioning from Google cloud). We use the default sampling hyperparameters which always include a temperature of 1.

We implement the prompt-based uncertainty measures as follows: for the method of Kadavath et al. (2022) we adopt their prompt template and include model generations in the prompt, as recommended in their work; for the method of Tian et al. (2023), we use the two-stage variant (`Verb. 2S top-1`) which separates answer and confidence generation, so that the answer generation process can be comparable with the other methods; we use their prompt for confidence generation. For all test metrics we report bootstrap CIs computed using 10000 simulations.

All prompts and an illustrative subset of model generations can be found at the code repository.

```
I will provide you with a trivia question and its reference answer as of {year}. Please evaluate if
     the answer to this question is likely to have changed between {year} - 2023.

Examples of questions that might have changed answers include:
* Recent events or records
* The latest season of an ongoing TV series
* Current holders of public offices or titles

Use your best judgment in the evaluation. In scenarios where your internal knowledge could be
    inaccurate, rely on common sense and the provided reference answer. Please format your response
     in two lines, as follows:
1. A brief explanation of your reasoning (one sentence)
2. A "Yes" or "No" answer indicating whether the answer likely changed

Please adhere to the format of the example below:

Question: when did the botswana currency first come into circulation
Reference Answer: 1976
Reasoning: The date a currency first entered circulation is a historical fact that doesn't change.
Answer: No

Now please answer the following:

Question: {question}
Reference Answer: {answer}
```

Figure 5: Question answering: prompt template used to filter trivia datasets.

B.2 Question Answering: Additional Results

Full results in the setting of Fig. 1–3 are provided in Fig. 6–8, respectively, where we also provide results for AUARC. As expected, across LMs with a similar level of predictive performance, AUARC is generally correlated with calibration error: see the comparison between `llama-3.1-70b` and `gemini-1.5-pro` (or `gpt-4o-mini`) in Fig. 7. It is also interesting to note that in the setting of Fig. 1, the calibration of the `gemma2-27b` model (Gemma Team, 2024) often improves after instruction tuning; inspection on `TriviaQA` and `SciQ` appears to show that the instruction-tuned LM has a similar level of variability in its generations, but the quality of generation has generally improved.

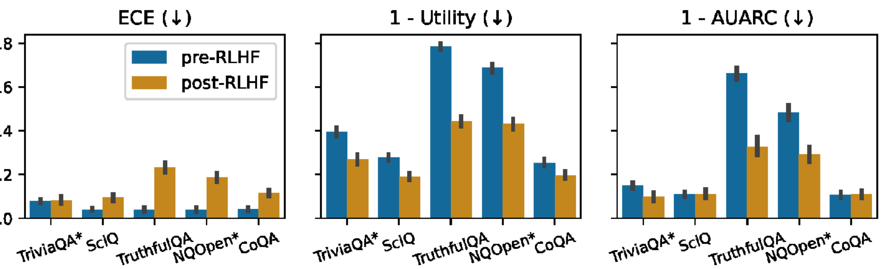
Fig. 9 compares the reliability diagrams obtained from different uncertainty measures, evaluated on `llama-3.1-70b` and `gpt-4o`. We can see that all uncertainty measures can be overconfident on inputs with low uncertainty, while the prompt-based measures can also be underconfident on high-uncertainty inputs.

B.3 Machine Translation: Setup Details

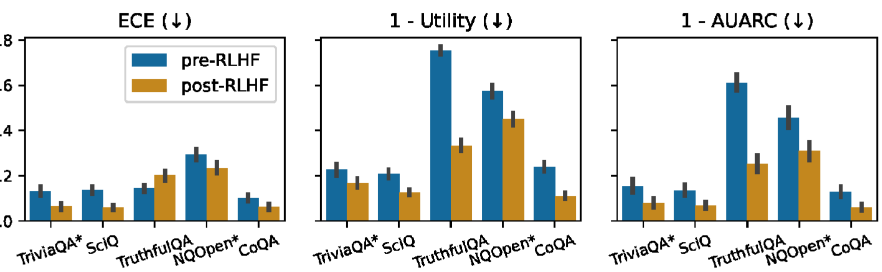
We use the prompt in Agarwal et al. (2024) for the translation task, and generate $x_{n+1:N}$ using `gemini-1.5-flash` with the prompt in Fig. 10. Details for the LMs involved are the same as App. B.1.

The deferral experiments involve two ICL predictors. The base predictor is implemented by sampling $n = 4$ demonstrations uniformly without replacement. For the more expensive predictor we append $n' - n = 124$ demonstrations that are closest to the test query in an embedding space, where embeddings are computed using the `text-embedding-004` model from Google Cloud.[6] In both cases we retrieve demonstrations from the `dev` split of the dataset and evaluate on the `devtest` split. We experimented with both (nontrivial) MBR and Gibbs

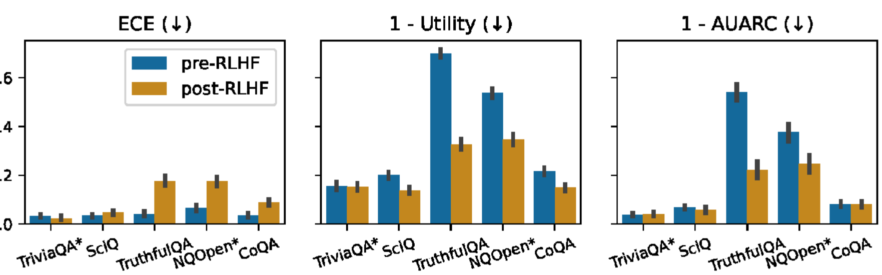
[6] `https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/text-embeddings-api`



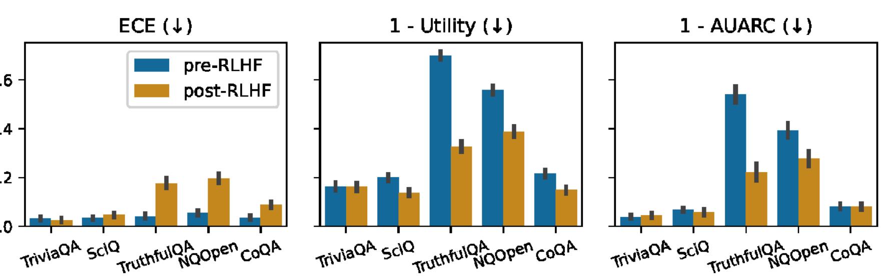
(a) `llama-3.1-8b`



(b) `gemma2-27b`



(c) `llama-3.1-70b`



(d) `llama-3.1-70b`, without filtering for trivia datasets

Figure 6: Question answering: full results in the setting of Fig. 1.

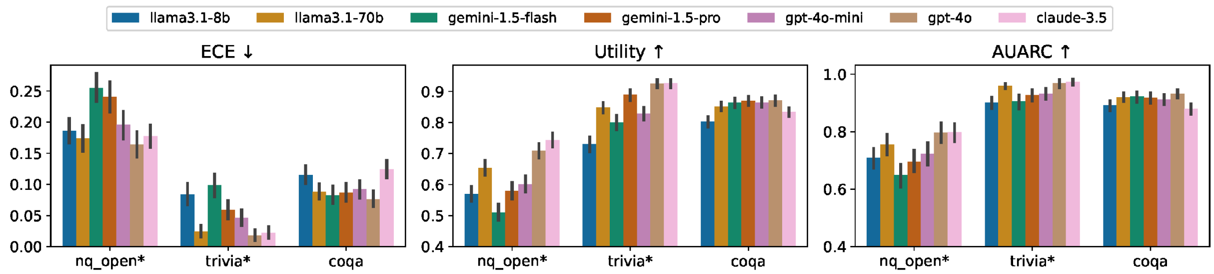


Figure 7: Question answering: full results in the setting of Fig. 2. Error bar denotes 95% bootstrap CI.

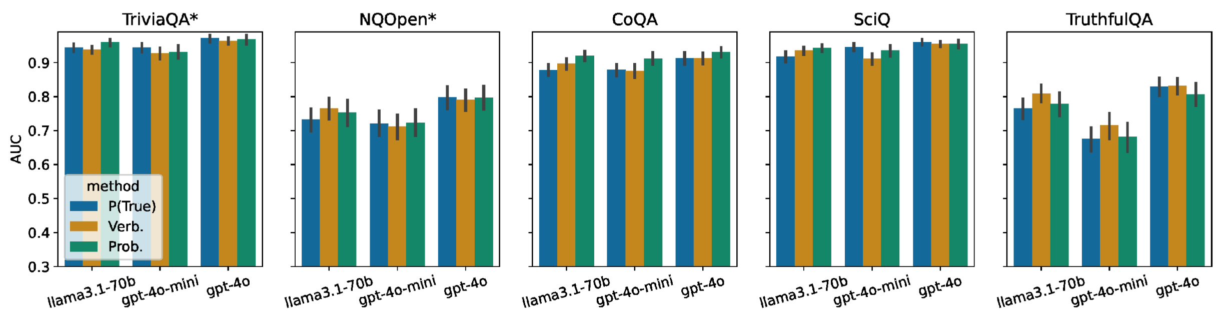


Figure 8: Question answering: results for AUARC in the setting of Fig. 3.

decoding;[7] the former is implemented by drawing 16 samples for each query, using half of them to define $\mathcal{Y}'_I$ and half to approximate the expectation in (1). The main text reports the results for the MBR predictors, as we find them to perform better across all experiments. However, the same EU measure also works well for the Gibbs predictors, as shown in Fig. 12.

Following Robinson et al. (2023), we use the chrF2++ score implemented by Post (2018). The p-values reported in Fig. 4 are calculated using 10000 replications for the `random` baseline.

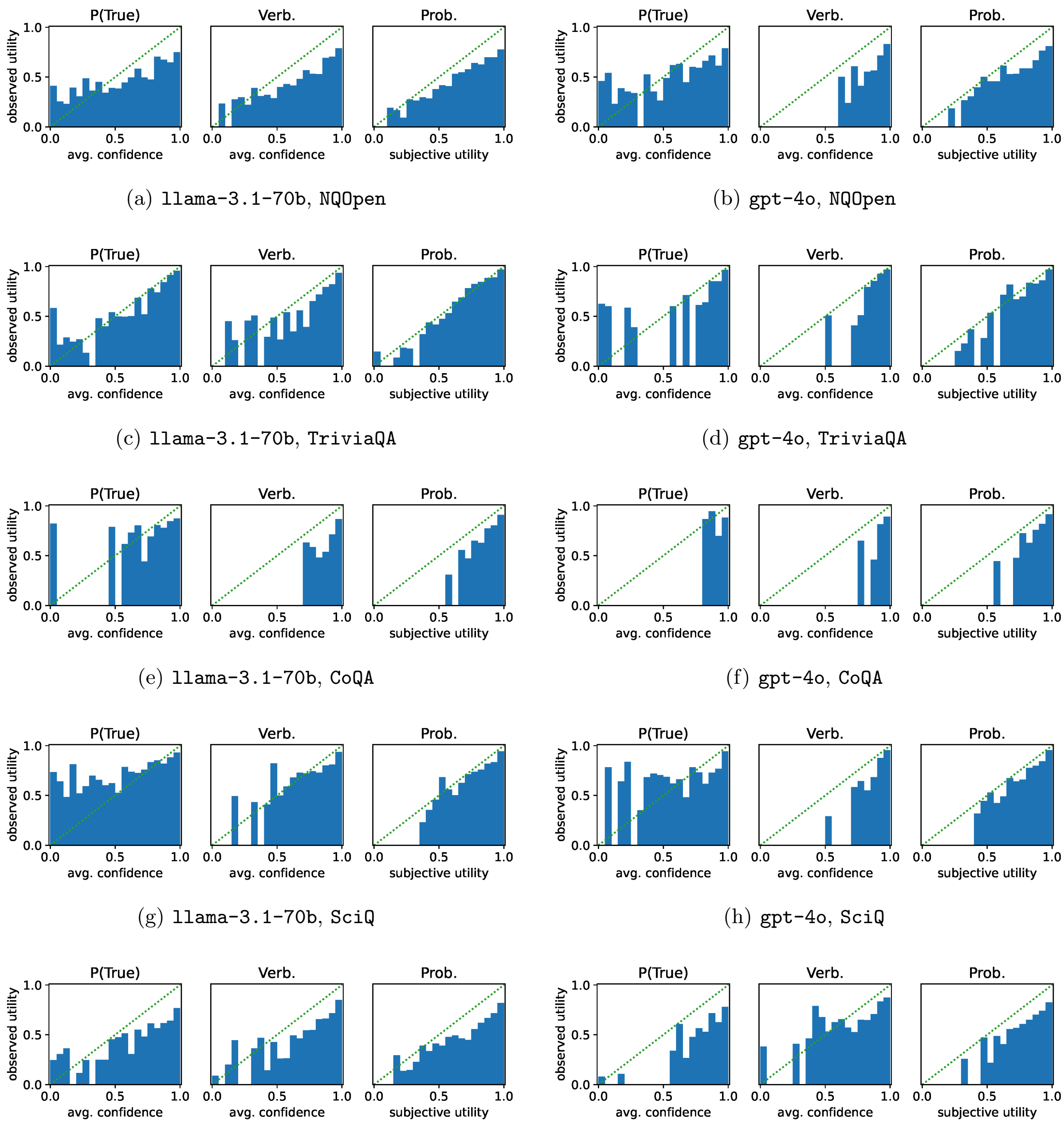
B.4 Machine Translation: Additional Results

Table 2 reports the calibration error from all models. We can see that `gemini-1.5-pro` consistently demonstrates worse calibration.

Table 2: Machine translation: ECE for all models.

	`gemini-1.5-pro`	`gpt-4o-mini`	`gpt-4o`
Tam	0.250	0.041	0.065
Kmr	0.356	0.017	0.010

[7]Note that these predictors do not coincide with the two predictors in (5). Instead, (5) is only used as a surrogate to approximate the performance improvement in the deferral process.



(a) `llama-3.1-70b`, `NQOpen`

(b) `gpt-4o`, `NQOpen`

(c) `llama-3.1-70b`, `TriviaQA`

(d) `gpt-4o`, `TriviaQA`

(e) `llama-3.1-70b`, `CoQA`

(f) `gpt-4o`, `CoQA`

(g) `llama-3.1-70b`, `SciQ`

(h) `gpt-4o`, `SciQ`

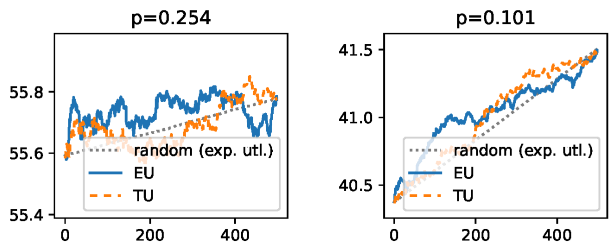
(i) `llama-3.1-70b`, `TruthfulQA`

(j) `gpt-4o`, `TruthfulQA`

Figure 9: Question answering: reliability diagrams for the `gpt-4o` and `llama-3.1-70b` models. Note that the x-axis is always equivalent to the negated uncertainty measure.

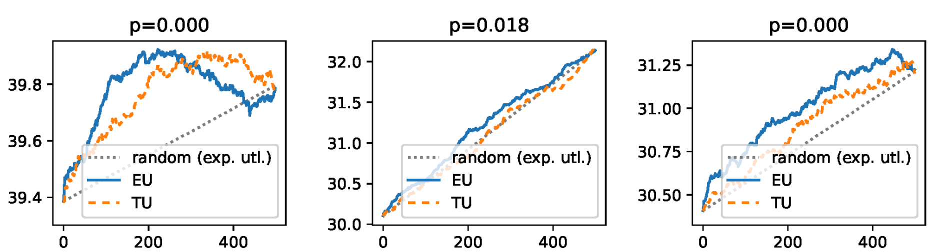
```
Please generate {K} sentences that convey a broadly similar meaning as the one provided, ensuring they
    are of similar length and use varied wording. Place a blank line after each sentence. Avoid adding
    any extra explanations.

{x_*}
```

Figure 10: Machine translation: prompt used to define p_x. x_* denotes the test query.



(a) English → Tamil (b) English → Kurdish

Figure 11: Machine translation: results for `gemini-1.5-pro` in the setting of Fig. 4.



(a) `gemini-1.5-pro` (b) `gpt-4o-mini` (c) `gpt-4o`

Figure 12: Machine translation: deferral curve for Gibbs predictors on the English → Kurdish task.