Entropy Is Not Enough: Uncertainty Quantification for LLMs fails under Aleatoric Uncertainty

Tim Tomov Dominik Fuchsgruber Tom Wollschläger Stephan Günnemann
School of Computation, Information and Technology & Munich Data Science Institute
Technical University of Munich
{tim.tomov,d.fuchsgruber,t.wollschlaeger,s.guennemann}@tum.de

Abstract

Accurate uncertainty quantification (UQ) in Large Language Models (LLMs) is critical for trustworthy deployment. While real-world language is inherently ambiguous, existing UQ methods implicitly assume scenarios with no ambiguity. Therefore, a natural question is how they work under ambiguity. In this work, we demonstrate that current uncertainty estimators only perform well under the restrictive assumption of no aleatoric uncertainty and degrade significantly on ambiguous data. Specifically, we provide theoretical insights into this limitation and introduce two question-answering (QA) datasets with ground-truth answer probabilities. Using these datasets, we show that current uncertainty estimators perform close to random under real-world ambiguity. This highlights a fundamental limitation in existing practices and emphasizes the urgent need for new uncertainty quantification approaches that account for the ambiguity in language modeling.

1 Introduction

Many linguistic tasks that are solved by Large language models (LLMs) can be framed as *question-answering* (QA): a user poses a query, and the model provides an answer. As LLMs are increasingly deployed in high-stakes domains—such as medical diagnosis, legal advice, or autonomous decision-making it becomes critical not only to obtain correct answers but also to have reliable estimates of how well the model understands the data, also referred to as *epistemic uncertainty*. An important consideration when assessing model reliability in this context is that some questions permit more than one answer. Consider these two examples:

Single-answer (No ambiguity): "*Which hormone do I lack if I have type 1 diabetes?*" $\rightarrow$ *Insulin*.

Multi-answer (Ambiguity): "*Which medication should I take for type 2 diabetes?*" $\rightarrow$ *Metformin, Sulfonylureas, DPP-4 Inhibitors, ...* (all plausible, but with different probabilities).

Since in the first example there is only one correct answer, any model that predicts a distribution over possible replies should put all mass on this one answer. In the second example, multiple answers are correct, and they may be associated with different probabilities. This is known as *aleatoric uncertainty*: It refers to the randomness that is intrinsic to the distribution of true answers itself. Most uncertainty–quantification (UQ) methods for LLMs, however, are evaluated on data resembling the first question, where aleatoric uncertainty is zero (Devic et al., 2025). In this restrictive setting, a variety of UQ methods show satisfactory performance in estimating *epistemic uncertainty* (Kuhn et al., 2023; Duan et al., 2024; Yadkori et al., 2024). However, many realistic applications involve non-trivial aleatoric uncertainty. This motivates a critical question: *How do current UQ approaches perform under realistic conditions of ambiguity?*

In this work, we demonstrate that current UQ methods fail when answers have non-trivial aleatoric uncertainty. Specifically, we:

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

- Provide theoretical explanations why entropy-based methods are effective in settings of zero aleatoric uncertainty but fail in more general settings (Section 3.1).
- Extend existing ambiguous QA dataset with novel ground-truth answer distributions that are estimated from factual co-occurrence statistics (Section 3.2).
- Empirically confirm that existing UQ methods perform nearly at random in distinguishing and ranking high and low uncertainty questions when they are inherently ambiguous (Section 4.1).

Our findings fundamentally challenge the suitability of existing uncertainty quantification methods for the practical deployment of LLMs. We release our new benchmark with empirical answer distributions to support future research on UQ methods that explicitly account for non-trivial ambiguity.

2 Background

Uncertainty quantification (UQ) in machine learning (ML) characterizes the uncertainty in a model's predictive distribution for a given input x. This uncertainty, often referred to as *total uncertainty*, stems from two distinct sources: *epistemic uncertainty*, reflecting uncertainty in the model itself due to limited training data, model misspecification, or artifacts of optimization, and *aleatoric uncertainty*, which represents intrinsic randomness in the true data-generating process (Hüllermeier and Waegeman, 2021; Gawlikowski et al., 2022). Epistemic uncertainty can be reduced with sufficient data and a well-specified model, whereas aleatoric uncertainty is irreducible by definition. Importantly, when both sources are present, they jointly shape the model's predictive distribution, and naive uncertainty estimates may confuse epistemic uncertainty for genuine data ambiguity. As such, disentangling these sources of uncertainty is a central challenge in reliable ML.

With the general capability of LLMs to address diverse tasks by framing them as question-answering (QA) problems (Sanh et al., 2022), a natural approach to uncertainty quantification in LLMs is assessing the model's certainty in the answers it provides. Since LLMs often produce syntactically diverse yet semantically equivalent answers, it is useful to group answers into semantic equivalence classes (Kuhn et al., 2023). For instance, to the question "What is the capital of France?", the answers "Paris" or "The capital is Paris" represent the same semantic class. We focus on the distribution over these semantically distinct classes, denoted p in the remainder, with details given in Section B. Importantly, this perspective enables studying uncertainty quantification for LLMs as a classification problem, enabling us to build on established theory; see Section D for related approaches.

3 Methods

Following Kotelevskii et al. (2025), we define the *total uncertainty (TU)* as the cross-entropy between the true distribution p^* over semantic classes and the semantic distribution predicted by the model p. This allows a natural decomposition: *Aleatoric uncertainty (AU)* is the entropy of the true distribution p^* and *epistemic uncertainty (EU)* the Kullback-Leibler divergence between p^* and predicted distribution p:

$$\underbrace{\mathrm{CE}(p^*, p)}_{\text{Total}(TU)} = \underbrace{H(p^*)}_{\text{Aleatoric}(AU)} + \underbrace{\mathrm{KL}(p^* \| p)}_{\text{Epistemic}(EU)}$$

Unlike the widely used information-theoretic decomposition for sampling-based methods (Depeweg et al., 2018), which has faced criticism for conflating distinct sources of uncertainty (Wimmer et al., 2023; Smith et al., 2025), this formulation makes use of a reference distribution p^*. This is critical for principled evaluation (Smith et al., 2025) and provides a powerful tool to study uncertainty as we show in Section 3.1.

Many existing methods for LLMs estimate EU by measuring the variability in the predictive distribution p, with predictive entropy $H(p)$[1] being the most prominent example (Vashurin et al., 2025). In this work, we investigate under which conditions this variation-based EU estimation is a good approximation with high probability, while focusing on $H(p)$ as our central example. The corresponding insights translate to other variability-based uncertainty measures as well.

[1] In LLMs, where p is the semantic class distribution $H(p)$ is known as semantic entropy (Kuhn et al., 2023).

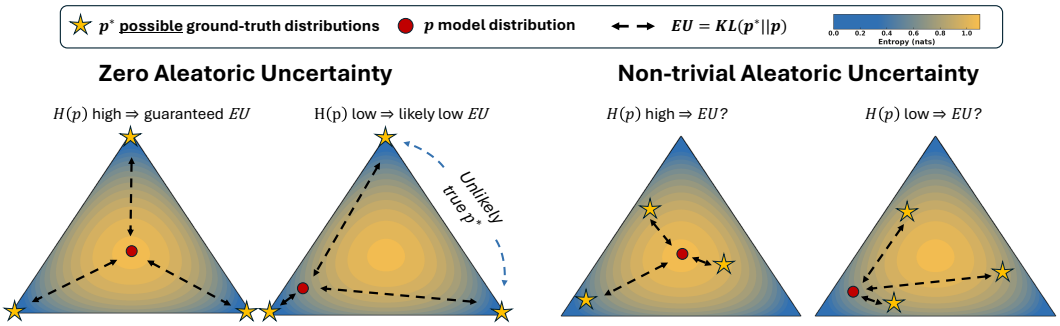

Figure 1: *Theoretical Insights on 3-class simplex* **Left:** Under zero aleatoric uncertainty, high entropy $H(p)$ guarantees EU, since all possible p^* are far away (Theorem 1). Assuming a well-trained model, observing low entropy likely indicates low EU as the model cannot often be confidently incorrect (Theorem 2). **Right**: Under non-trivial aleatoric uncertainty, observing high or low entropy does not provide information about the EU, since the ground-truth p^* has no constraint on its location.

3.1 When Does Model Entropy Reflect Epistemic Uncertainty?

When aleatoric uncertainty is zero, the epistemic uncertainty reduces to the negative log-probability the model assigns to the correct semantic class, given as $-\log(p(y = y^*))$ (Proposition 2). Therefore, it can be understood as the model's confidence in the correct answer. Building on this insight, we introduce two theoretical arguments that explain why entropy-based uncertainty estimation methods perform well in this case. Critically, neither of these arguments holds in the case of non-trivial AU, and we show that a faithful estimation in this case is impossible.

Theorem 1 (High Entropy ⇒ High Epistemic Uncertainty)**.** *Let there be $K \geq 2$ classes and $\delta \in [0, \log K]$ be a threshold on the entropy. Furthermore, let α_δ be the maximal possible probability on some class s.t. $H(p) \geq \delta$. Then the epistemic uncertainty with $H(p) \geq \delta$ is at least:*

$$EU \;\geq\; -\log \alpha_\delta.$$

Intuitively, high entropy $H(p) \geq \delta$ implies that the predictive distribution must become increasingly flatter. This implies that the maximum probability assigned to any class can be at most α_δ - naturally, also for the correct class y^*. Since epistemic uncertainty is quantified as $-\log p(y = y^*)$, such a flat predictive distribution hence leads to large epistemic uncertainty (Figure 1). Thus, Theorem 1 explicitly shows that *high predictive entropy necessarily implies high epistemic uncertainty*.

Theorem 2 (Low Entropy ⇒ Low Epistemic Uncertainty with High Probability)**.** *Let there be $K \geq 2$ classes and $\delta \in [0, \log 2]$ be a threshold on the entropy. Furthermore let $\mathcal{L} = \mathbb{E}_{(x,y)}[-\log p_y]$ the average loss and γ_δ be the minimal possible maximum probability on some class s.t $H(p) \leq \delta$. Then the probability that the epistemic uncertainty with $H(p) \leq \delta$ will be less than $-\log(\gamma_\delta)$ satisfies:*

$$\mathbb{P}(EU \leq -\log(\gamma_\delta) \mid H(p) \leq \delta) \geq 1 - \frac{\mathcal{L}}{-\log(1-\gamma_\delta) * \mathbb{P}(H(p) \leq \delta)}$$

Theorem 2 complements Theorem 1: If the entropy of a predictive distribution is small ($H(p) \leq \delta$), some class must have high probability γ_δ. This creates a dichotomy: if that class is correct, the EU is small ($\leq -\log(\gamma_\delta)$), whereas if it is incorrect, the EU is large ($\geq -\log(1-\gamma_\delta)$)(Figure 1). Noting that the training loss $-\log p_y$ coincides with EU under zero AU, the average loss $\mathcal{L}$ is the expected EU. As such, for a well-trained model (small $\mathcal{L}$), frequent high-EU errors would contradict the low average loss. Hence, confident but wrong predictions must be rare. The bound also depends on the frequency of confident predictions; while it may loosen when such cases are rare, the negative log likelihood objective encourages models to be confident, making this unlikely in practice. Overall, Theorem 2 formalizes the intuition that for models that make confident predictions and perform well on average, *observing a low predictive entropy is likely to correspond to low epistemic uncertainty*.

Non-trivial Aleatoric Uncertainty The preceding theoretical arguments show that under zero aleatoric uncertainty, high entropy indicates epistemic uncertainty, whereas low entropy predominantly reflects epistemic confidence. These insights use the fact that under zero aleatoric uncertainty,

Table 1: Examples of question-answer-distribution pairs

Dataset	Question	Answer(s)	# Counts in Data	$\mathbf{p}^*$	Entropy $\mathbf{H}(\mathbf{p}^*)$
TriviaQA	Where in England was Dame Judi Dench born?	$\{Yorkshire\}$	n/a	$[1.00]$	0.0
MAQA*	What is one essential component of the fire triangle?	$\{Heat, Fuel, Oxygen\}$	$\{31, 32, 25\}$	$[0.35, 0.36, 0.29]$	1.1
AmbigQA*	What is the name of one princess in Frozen?	$\{Elsa, Anna\}$	$\{188, 91\}$	$[0.67, 0.33]$	0.63

the ground-truth is constrained to be an indicator distribution.[2]Allowing for arbitrary aleatoric uncertainty lifts this restriction on p^*. Consequently, a high-entropy prediction no longer necessarily indicates high EU as the entropy may also arise from an inherently uncertain ground-truth (Figure 1 right) that is, at the same time, well reflected by the model (low EU). More generally, no function of the predictive distribution p alone can distinguish epistemic uncertainty from intrinsic ambiguity:

Proposition 1 (Non-Identifiability of Epistemic Uncertainty)**.** *Let $K \geq 2$ and Δ^{K-1} be the probability simplex over K classes. For any function $f : \Delta^{K-1} \to \mathbb{R}$ and any $p \in \Delta^{K-1}$, there exist $p_1^*, p_2^* \in \Delta^{K-1}$ such that*

$$\mathrm{KL}(p_1^* \,\|\, p) = 0 \quad \textit{and} \quad \mathrm{KL}(p_2^* \,\|\, p) = -\log \min_i p_i \geq \log K,$$

Thus, any function $f(p)$ can both indicate zero or high ($\geq \log(K)$) epistemic uncertainty.

As such, any estimator that is a function of p — e.g. semantic entropy — cannot faithfully estimate EU without restrictions on AU. Since assumptions about the inherent ambiguity poorly reflect many linguistic problems, these approaches are rendered unreliable measures of EU in practice.

3.2 A Novel QA benchmark for Unrestricted Aleatoric Uncertainty

To study how current estimators perform under aleatoric uncertainty, we introduce two novel ambiguous QA datasets, MAQA* & AmbigQA*, that are equipped with explicit estimates of the ground-truth distribution p^*. Concretely, we augment the ambiguous QA datasets MAQA (Yang et al., 2025) and AmbigQA (Min et al., 2020) by estimating the answer probabilities from the co-occurrence of question-answer pairs in the training data. This choice is well supported by previous work: Empirically, co-occurrence statistics correlate strongly with model performance: models score higher on samples with frequent co-occurrence (Kandpal et al., 2023; Mallen et al., 2023), and recently, Wang et al. (2025) demonstrates that, particularly in factual QA, LLM output probabilities correlate with co-occurrence statistics. Theoretically, as $n \to \infty$, the model will reproduce the pretraining distribution p_{train}, and epistemic uncertainty will vanish (Smith et al., 2025). The remaining uncertainty is thus purely aleatoric, reflecting the intrinsic variability in p_{train} itself. Consequently, estimating p^* from statistics of p_{train} is more principled than relying on external annotations that may diverge from p_{train}.

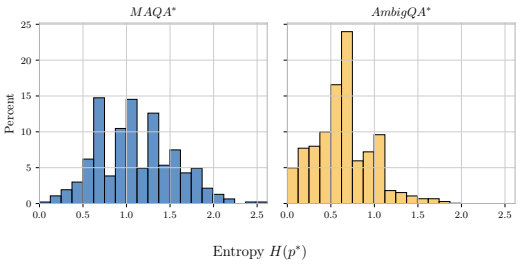

Figure 2: Distribution of ground-truth entropy $H(p^*)$ across questions in MAQA* (left) and AmbigQA* (right)

Since the pre-training datasets for LLMs are not publicly available, we instead employ the English Wikipedia (Wikimedia Enterprise, 2024) as a proxy for the pre-training corpus due to its widespread use in LLM pre-training and comprehensive coverage of factual knowledge. To perform the co-occurrence search, we use keywords extracted from the question alongside candidate answers. The keywords represent the most important words in the question, e.g., the question's subject, and importantly, both keywords and answers are stemmed to their base forms to ensure robustness against surface-form variation. Elsahar et al. (2018) demonstrate that subject–object co-occurrence is a reliable indicator for the presence of a subject–relation–object triplet, making it suitable for fact counting. We further improve the precision of these counts by using an entailment model to verify the factual occurrence of each candidate co-occurrence. The resulting datasets contain 468 and 2553 samples, spanning diverse cases of aleatoric uncertainty (Figure 2) with examples shown in Table 1.[3]

[2]In Section E.1 we discuss that also other restrictions are useful.

[3]Dataset: `https://huggingface.co/collections/ttomov/llm-uncertainty-under-ambiguity`

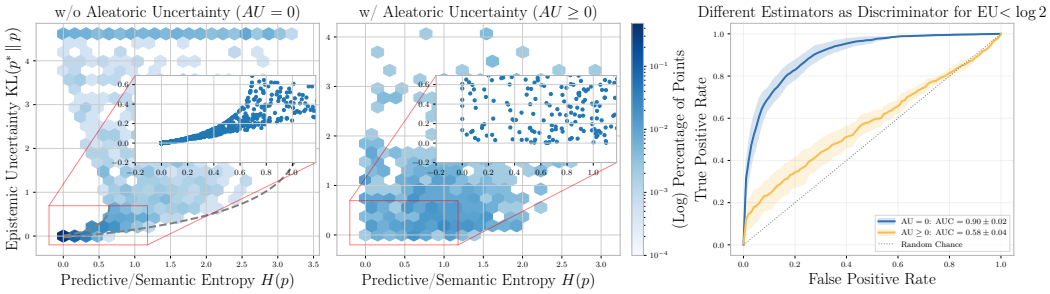

Figure 3: *Relationship between estimators and epistemic uncertainty (EU) for Gemma 3-12B on MAQA**. **Left**: Relationship between $H(p)$ and EU. In the absence of aleatoric uncertainty ($\mathrm{AU} = 0$), predictive entropy and EU are correlated, while in the presence of variable aleatoric uncertainty ($\mathrm{AU} \geq 0$), predictive entropy and EU are uncorrelated. Lines indicate theoretical bounds on the EU. **Right** ROC curves showing different estimators as a discriminator for identifying certain predictions ($\mathrm{EU} < \log(2)$). Shaded regions represent one standard deviation over different estimators.

To validate the robustness of our estimation strategy, we confirm these ground-truth distributions using two alternative co-occurrence search strategies. (i) Similarly as above, using keywords and answers but using as corpus the RedPajama-V1 dataset (Weber et al., 2024) via infini-gram (Liu et al., 2024), and (ii) through entity linking on the Pile dataset (Gao et al., 2020) using DPBedia Spotlight (Kandpal et al., 2023; Daiber et al., 2013). We find that the distributions from all strategies strongly align, which validates our approach for estimating ground-truth distributions (Section C).

4 Experiments

To confirm our theoretical analysis, we compare different UQ estimators across models on our novel MAQA* and AmbigQA* datasets and contrast that to performance on data with no AU.

Setup. We evaluate uncertainty estimators on three datasets: TriviaQA[4] (Joshi et al., 2017), containing questions with zero aleatoric uncertainty, and our novel MAQA* and AmbigQA* datasets, which feature varying degrees of ambiguity. The estimators include Semantic Entropy (SE) (Kuhn et al., 2023), Maximum Sentence Probability (MSP), Shifting Attention to Relevance (SAR) (Duan et al., 2024), and Iterative Prompting (IP) (Yadkori et al., 2024). We evaluate these across several models: LLaMA3.1 8B (Grattafiori et al., 2024), Gemma3 12B (Team et al., 2025), Qwen2.5 14B (Qwen et al., 2025)—each in both base and instruct variants—as well as GPT-4.1-mini, which serves as state-of-the-art LLM. We additionally experiment with different model sizes in Section A.4.

Metrics. We study how well the estimated EU represents the true EU. As both are continuous quantities, we report the following metrics: (i) For a given threshold δ we measure the separation between uncertain ($EU \geq \delta$) and certain ($EU < \delta$) samples using AUCROC. (ii) As threshold-independent metric, we employ the concordance statistic AUC_c(Therneau and Atkinson, 2024) ,which estimates $\mathbb{P}(EU_i > EU_j \mid \text{Estimator}_i > \text{Estimator}_j)$. It quantifies how reliably the estimator assigns higher ranks to samples with greater epistemic uncertainty.

4.1 Key Observations

I. Predictive Entropy is a reliable estimator of EU only under zero AU This observation (Figure 3, left) aligns well with both theorems of Section 3.1. The lower bound on the EU from Theorem 1 ensures that models with substantial predictive entropy cannot have low epistemic certainty. The primary sources of errors in the zero AU case are confident yet incorrect predictions (left top). However, given a sufficiently well-trained model, these occur with low probability (Theorem 2), which is reflected in the sparsity of that region (low bin counts). Conversely, for non-trivial AU, predictive entropy has no connection to EU. Pathological cases include predictions with high predictive entropy despite low EU, which can be seen in the $AU \geq 0$ case of Figure 3 (middle

[4]We use the first 2000 samples as this is enough to demonstrate the phenomenon.

Table 2: Concordance scores AUC_c for all estimators of three models on TriviaQA (AU=0) and on AmbigQA* & MAQA* (AU≥0). An $AUC_c = 0.50$ corresponds to random chance.

Model	AU = 0				AU ≥ 0							
	TriviaQA				MAQA*				AmbigQA*			
	SE	MSP	SAR	IP	SE	MSP	SAR	IP	SE	MSP	SAR	IP
Llama 3.1-8B	0.80	0.74	0.79	0.80	0.52	0.49	0.51	0.53	0.61	0.58	0.60	0.60
Gemma 3-12B	0.91	0.79	0.86	0.90	0.55	0.53	0.58	0.60	0.66	0.64	0.66	0.66
Qwen 2.5-14B	0.87	0.74	0.82	0.86	0.59	0.56	0.62	0.59	0.67	0.63	0.67	0.66

plot) in the right/middle bottom section. Another case are predictions exhibiting higher EU but low predictive entropy, which are located on the left-middle top section. These patterns demonstrate that predictive entropy is not a suitable quantity for accurately estimating EU under non-trivial AU.

II. Current estimators perform nearly at random in distinguishing and ranking high and low EU samples under non-trivial AU To quantify these effects, we set the uncertainty threshold to $\delta = \log(2)$[5]. In the zero AU case, different estimators well separate epistemically certain and uncertain (Figure 3, right). However, in scenarios involving non-zero AU, the discriminative power of the estimators reduces to close to random performance.

These findings are supported by the concordance statistic: Semantic Entropy fails to rank the true EU better than random chance in the presence of AU across models and both MAQA* and AmbigQA* (Table 2). The slightly higher scores on AmbigQA* can be attributed to its large proportion of near-zero entropy samples rather than genuine estimator effectiveness. Similar results also hold for the other estimators. Our findings are consistent with the instruct versions of the models as well; additionally, we note that these models, other than the corresponding base models, collapse their predictions frequently to only one semantic answer (Section A.3). Our claims also generalize to different model sizes, with the exception that very small models, which surprisingly perform better than random. We defer a discussion to Section A.4. Finally, this pathological pattern arises for all strategies for estimating p^*(Section A.2).

5 Discussion

Limitations Our new benchmark quantifies p^* as factual occurrences in Wikipedia. Although evidence suggests that such occurrences correlate well with model performance (Kandpal et al., 2023; Mallen et al., 2023; Wang et al., 2025), there is, to our knowledge, no work that empirically shows LLMs approach this distribution in the infinite data limit. We aim to mitigate potential inaccuracies by supplementing experiments using Dirichlet-distributed perturbations around the estimated ground truth (Section A.1) and find our claims to be robust. Furthermore, our evaluation requires models to provide a single answer for each question. While this is consistent with prior work (Kuhn et al., 2023; Aichberger et al., 2024), settings in which multiple answers are generated simultaneously require a fundamentally different theoretical framework for modeling uncertainty.

Current Estimators are not reliable The observations of Section 4.1 confirm the theoretical insights of Section 3.1. With our novel benchmark, we demonstrate that in the general scenario of ambiguity, current estimators fail to adequately assess *epistemic uncertainty*. Consequently, their use in general language tasks is problematic and not reliable.

Toward Reliable Estimators Accurately estimating epistemic uncertainty based on a function of p indeed appears impossible. Hence, more information than p is needed, and it seems, e.g., sensible to account for model uncertainty during training itself. For example, in classical UQ, evidential deep learning (Sensoy et al., 2018) learns a second-order distribution to represent epistemic uncertainty. More recent approaches train models on joint distributions to capture epistemic uncertainty (Johnson et al., 2024; Ahdritz et al., 2024). We hope that our work encourages the development of such estimators and that our benchmark enables a more holistic study of UQ in LLMs.

[5]In a binary classification scenario with $AU = 0$, an EU of $\log(2)$ corresponds precisely to $p_1 = p_2 = \frac{1}{2}$. Choosing a threshold $\delta > \log(2)$ would incorrectly label genuine misclassifications as epistemically certain.

References

Ahdritz, G., Gollakota, A., Gopalan, P., Peale, C., and Wieder, U. (2024). Provable uncertainty decomposition via higher-order calibration.

Aichberger, L., Schweighofer, K., and Hochreiter, S. (2024). Rethinking uncertainty estimation in natural language generation.

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., and Udluft, S. (2018). Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning.

Devic, S., Srinivasan, T., Thomason, J., Neiswanger, W., and Sharan, V. (2025). From calibration to collaboration: Llm uncertainty quantification should be more human-centered.

Duan, J., Cheng, H., Wang, S., Zavalny, A., Wang, C., Xu, R., Kailkhura, B., and Xu, K. (2024). Shifting attention to relevance: Towards the uncertainty estimation of large language models.

Elsahar, H., Vougiouklis, P., Remaci, A., Gravier, C., Hare, J., Laforest, F., and Simperl, E. (2018). T-REx: A large scale alignment of natural language with knowledge base triples. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. (2020). The pile: An 800gb dataset of diverse text for language modeling.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., and Zhu, X. X. (2022). A survey of uncertainty in deep neural networks.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., and et al., A. K. (2024). The llama 3 herd of models.

He, P., Liu, X., Gao, J., and Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Hou, B., Liu, Y., Qian, K., Andreas, J., Chang, S., and Zhang, Y. (2024). Decomposing uncertainty for large language models through input clarification ensembling.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

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.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., and Raffel, C. (2023). Large language models struggle to learn long-tail knowledge.

Kotelevskii, N., Kondratyev, V., Takáč, M., Éric Moulines, and Panov, M. (2025). From risk to uncertainty: Generating predictive uncertainty measures via bayesian estimation.

Kuhn, L., Gal, Y., and Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation.

Lin, J., Ma, X., Lin, S.-C., Yang, J.-H., Pradeep, R., and Nogueira, R. (2021). Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Liu, J., Min, S., Zettlemoyer, L., Choi, Y., and Hajishirzi, H. (2024). Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling*.

Liu, X., Chen, T., Da, L., Chen, C., Lin, Z., and Wei, H. (2025). Uncertainty quantification and confidence calibration in large language models: A survey.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Min, S., Michael, J., Hajishirzi, H., and Zettlemoyer, L. (2020). AmbigQA: Answering ambiguous open-domain questions. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

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

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., and et al. (2025). Qwen2.5 technical report.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., Datta, D., Chang, J., Jiang, M. T.-J., Wang, H., Manica, M., Shen, S., Yong, Z. X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Fevry, T., Fries, J. A., Teehan, R., Bers, T., Biderman, S., Gao, L., Wolf, T., and Rush, A. M. (2022). Multitask prompted training enables zero-shot task generalization.

Sensoy, M., Kaplan, L., and Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty.

Smith, F. B., Kossen, J., Trollope, E., van der Wilk, M., Foster, A., and Rainforth, T. (2025). Rethinking aleatoric and epistemic uncertainty. In *Forty-second International Conference on Machine Learning*.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., and et al. (2025). Gemma 3 technical report.

Therneau, T. M. and Atkinson, E. (2024). Concordance. Vignette of the `survival` R package. Accessed: 2025-08-29.

Vashurin, R., Fadeeva, E., Vazhentsev, A., Rvanova, L., Vasilev, D., Tsvigun, A., Petrakov, S., Xing, R., Sadallah, A., Grishchenkov, K., Panchenko, A., Baldwin, T., Nakov, P., Panov, M., and Shelmanov, A. (2025). Benchmarking uncertainty quantification methods for large language models with LM-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Wang, X., Antoniades, A., Elazar, Y., Amayuelas, A., Albalak, A., Zhang, K., and Wang, W. Y. (2025). Generalization v.s. memorization: Tracing language models' capabilities back to pretraining data. In *The Thirteenth International Conference on Learning Representations*.

Weber, M., Fu, D. Y., Anthony, Q., Oren, Y., Adams, S., Alexandrov, A., Lyu, X., Nguyen, H., Yao, X., Adams, V., Athiwaratkun, B., Chalamala, R., Chen, K., Ryabinin, M., Dao, T., Liang, P., Ré, C., Rish, I., and Zhang, C. (2024). Redpajama: an open dataset for training large language models. *NeurIPS Datasets and Benchmarks Track*.

Wikimedia Enterprise, W. F. (2024). Structured wikipedia.

Wimmer, L., Sale, Y., Hofman, P., Bischl, B., and Hüllermeier, E. (2023). Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?

Xie, J., Zhang, K., Chen, J., Lou, R., and Su, Y. (2024). Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts.

Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y., and Xu, W. (2024). Knowledge conflicts for llms: A survey.

Yadkori, Y. A., Kuzborskij, I., György, A., and Szepesvári, C. (2024). To believe or not to believe your llm.

Yang, Y., Yoo, H., and Lee, H. (2025). Maqa: Evaluating uncertainty quantification in llms regarding data uncertainty.

A Additional Experiments

A.1 Accounting for uncertainty in estimating p^*

In practice, our estimate of the ground-truth distribution p^* is itself uncertain due to limited or noisy co-occurrence counts. To explicitly capture this uncertainty, we use a Dirichlet prior $p^* \sim \mathrm{Dir}(\alpha)$, with parameters $\alpha = (\alpha_1, \ldots, \alpha_C)$. We start with a uniform prior $\alpha_i = 1$ for all classes i. After observing co-occurrence counts n_i, the posterior parameters become $\alpha_i = 1 + n_i$. To prevent low-count posteriors from remaining too uniform—which would erroneously decouple the model prediction p from p^*—we introduce a scaling factor $\gamma \geq 1$, defining

$$\alpha_i = 1 + \gamma\, n_i.$$

Then, under the Dirichlet posterior, the *aleatoric uncertainty* is given by:

$$\begin{aligned}
\mathbb{E}_{p^* \sim \mathrm{Dir}(\alpha)}\big[H(p^*)\big] &= \mathbb{E}_{p^* \sim \mathrm{Dir}(\alpha)}\big[-\sum_{i=1}^{C} p_i^* log(p_i^*)\big] \\
&= -\sum_{i=1}^{C} \mathbb{E}_{p^* \sim \mathrm{Dir}(\alpha)}[p_i^* log(p_i^*)] \\
&= -\sum_{i=1}^{C} [\frac{\alpha_i}{\alpha_0}(\psi(\alpha_i + 1) - \psi(\alpha_0 + 1)]
\end{aligned}$$

where ψ is the digamma function, and we leverage the fact that each $p_i^* \sim Beta(\alpha_i, \alpha_0 - \alpha_i)$. Likewise, the *epistemic uncertainty* is defined as

$$\begin{aligned}
\mathbb{E}_{p^* \sim \mathrm{Dir}(\alpha)}\big[KL(p^* \,\|\, p)\big] &= \mathbb{E}_{p^* \sim \mathrm{Dir}(\alpha)}\big[CE(p^* || p)\big] - \mathbb{E}_{p^* \sim \mathrm{Dir}(\alpha)}\big[H(p^*)\big] \\
&= -\sum_{i=1}^{C} \mathbb{E}_{p^* \sim \mathrm{Dir}(\alpha)}[p_i^*] log(p_i) - \mathbb{E}_{p^* \sim \mathrm{Dir}(\alpha)}\big[H(p^*)\big] \\
&= -\sum_{i=1}^{C} \frac{\alpha_i}{\alpha_0} log(p_i) + \sum_{i=1}^{C} [\frac{\alpha_i}{\alpha_0}(\psi(\alpha_i + 1) - \psi(\alpha_0 + 1) \\
&= \sum_{i=1}^{C} \frac{\alpha_i}{\alpha_0}\big[(\psi(\alpha_i + 1) - \psi(\alpha_0 + 1) - log(p_i)\big]
\end{aligned}$$

We perform ablation studies over different values of γ (see Table 3). Increasing γ corresponds to making a stronger assumption that the retrieved p^* is exact, which causes the concordance score to approach the values reported in our main results. For smaller γ, p^* becomes more independent of p, especially given the relatively low counts noted earlier. Interestingly, estimator performance degrades further when we relax the assumption that p^* is exact, corroborating our main findings.

Table 3: Concordance scores AUC_c for Gemma 3-12B for different likelihood multipliers (γ) across uncertainty estimators.

Likelihood Multiplier (γ)	**MAQA***				**AmbigQA***			
	SE	**MSP**	**SAR**	**IP**	**SE**	**MSP**	**SAR**	**IP**
$\gamma = 1$	0.50	0.50	0.53	0.54	0.56	0.57	0.57	0.56
$\gamma = 2$	0.51	0.51	0.54	0.56	0.58	0.58	0.59	0.58
$\gamma = 5$	0.53	0.52	0.55	0.57	0.61	0.60	0.61	0.61
$\gamma = 10$	0.54	0.52	0.56	0.58	0.62	0.61	0.63	0.62
$\gamma = 100$	0.55	0.53	0.57	0.59	0.65	0.63	0.65	0.65
Main KL	0.55	0.53	0.58	0.60	0.66	0.64	0.66	0.66

A.2 Different p^* estimation methods

We assess the robustness of our results by evaluating different strategies for estimating the ground-truth p^*, as outlined in Section C. Across all estimators, the three methods yield highly similar results (Table 4), consistent with our observation that their estimated ground truths are strongly aligned. Note that, since we discard samples where at least one class has zero counts, different estimation strategies result in slightly different final datasets.

Table 4: Concordance scores AUC_c for Gemma 3-12B for different estimation methods for ground truth p^*

p^* Estimation Method	MAQA*				AmbigQA*			
	SE	MSP	SAR	IP	SE	MSP	SAR	IP
Wikipedia English	0.55	0.53	0.58	0.60	0.66	0.64	0.66	0.66
RedPajama-V1	0.55	0.53	0.58	0.59	0.65	0.63	0.65	0.65
The Pile	0.53	0.53	0.58	0.58	0.60	0.58	0.60	0.61

A.3 Instruct Models

For instruct models, we observe the same qualitative patterns as for base models (Table 5). Note that for GPT 4.1-Mini, the estimators SAR and MI were not computed since the model is only accessible via an API.

An additional insight is that the entropy for instruct models collapses to zero for most samples, even in cases with non-trivial aleatoric uncertainty. This behavior is undesirable, as it indicates that the models fail to represent any meaningful predictive distribution. Compared to base models, this collapse results in substantially worse model performance (average EU) (Figure 4). Moreover, the entropy collapse also degrades estimator performance on TriviaQA, since a model that always outputs a single answer provides no variability and thus no basis to distinguish certain from uncertain cases.

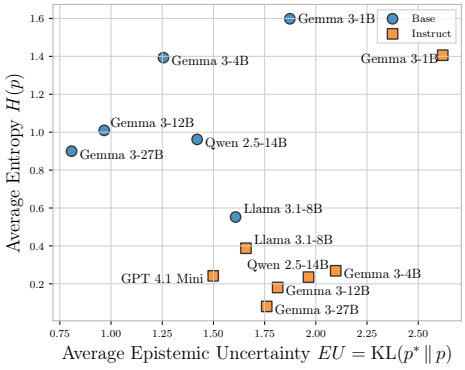

Figure 4: Entropy collapse of Instruct models on MAQA* leads to higher EU

Table 5: Concordance scores AUC_c for all estimators of instruct models on TriviaQA (AU=0) and on AmbigQA & MAQA (AU≥0). An $AUC_c = 0.50$ corresponds to random chance.

Model	AU = 0				AU ≥ 0							
	TriviaQA				MAQA*				AmbigQA*			
	SE	MSP	SAR	IP	SE	MSP	SAR	IP	SE	MSP	SAR	IP
Llama 3.1-8B-Instruct	0.84	0.79	0.83	0.81	0.54	0.52	0.53	0.54	0.60	0.59	0.60	0.60
Gemma 3-12B-Instruct	0.76	0.76	0.77	0.74	0.53	0.54	0.55	0.54	0.57	0.57	0.58	0.57
Qwen 2.5-14B-Instruct	0.73	0.69	0.73	0.69	0.55	0.54	0.56	0.55	0.57	0.56	0.57	0.56
GPT 4.1 Mini	0.87	0.85	-	-	0.50	0.52	-	-	0.59	0.61	-	-

A.4 Effect of Model Size

We evaluate different versions of Gemma 3—1B, 4B, 12B, and 27B—and observe that smaller models yield better performance for UQ estimation methods (Table 6) on MAQA*. This effect appears to stem from the fact that smaller models often do not know the correct answers, and thus produce arbitrary outputs that form a high-entropy distribution. Such cases naturally coincide with high epistemic uncertainty, as the model lacks knowledge of the answers. Conversely, when a smaller model does know the answers, the resulting distribution has lower entropy and correspondingly lower epistemic uncertainty. As shown in Figure 4, the average entropy decreases substantially with model size. Crucially, this reduction is accompanied by improved performance, indicating that larger

models more accurately capture the underlying ground-truth distributions. The reduced estimator performance of smaller models on TriviaQA is consistent with prior observations in the literature (Kuhn et al., 2023).

Table 6: Concordance scores AUC_c for all estimators of different model sizes on TriviaQA (AU=0) and on AmbigQA* & MAQA* (AU≥0). An $AUC_c = 0.50$ corresponds to random chance.

Model	AU = 0				AU ≥ 0							
	TriviaQA				MAQA*				AmbigQA*			
	SE	MSP	SAR	IP	SE	MSP	SAR	IP	SE	MSP	SAR	IP
Gemma 3-1B	0.78	0.72	0.77	0.78	0.69	0.63	0.71	0.69	0.67	0.64	0.64	0.64
Gemma 3-4B	0.85	0.76	0.82	0.85	0.65	0.57	0.65	0.67	0.69	0.65	0.69	0.67
Gemma 3-12B	0.91	0.79	0.86	0.90	0.55	0.53	0.58	0.60	0.66	0.64	0.66	0.66
Gemma 3-27B	0.93	0.80	0.87	0.91	0.52	0.52	0.56	0.57	0.65	0.62	0.65	0.64

A.5 AUCROC for different uncertainty thresholds δ

For completeness, we also report AUCROC scores for thresholds δ other than $\log(2)$ across all datasets (Table 7). As discussed in Section 4.1, the higher values observed on AmbigQA* are largely explained by its considerable proportion of near-zero entropy ground-truth samples.

Table 7: AUCROC scores for Gemma 3-12B for different uncertainty thresholds δ across all estimators

Uncertainty Threshold δ	AU = 0				AU ≥ 0							
	TriviaQA				MAQA*				AmbigQA*			
	SE	MSP	SAR	IP	SE	MSP	SAR	IP	SE	MSP	SAR	IP
$\delta = \log(1.5)$	0.95	0.89	0.92	0.94	0.54	0.52	0.58	0.61	0.78	0.73	0.77	0.78
$\delta = \log(2)$	0.93	0.87	0.90	0.92	0.56	0.53	0.60	0.63	0.75	0.71	0.74	0.75
$\delta = \log(3)$	0.90	0.85	0.89	0.89	0.58	0.56	0.62	0.65	0.73	0.70	0.73	0.73

B Implementation Details

B.1 Approximations

Approximation of p To estimate the probability $p(y)$ of a semantic class $y \in \mathcal{C}$, we sample K answers $a_1, \ldots, a_K$ from the model and then cluster them into semantic classes using an auxiliary entailment model. The probabilities of each semantic class are then obtained by aggregating and normalizing the answer probabilities within each class:

$$p(y) \approx \frac{\tilde{p}(y)}{\sum_{j=1}^{|\mathcal{C}|} \tilde{p}(y_j)}, \quad \text{where} \quad \tilde{p}(y) = \frac{1}{K} \sum_{i=1}^{K} \mathbb{I}(a_i \in y) p(a_i), \quad a_i \sim p(a).$$

As $K \to \infty$, the approximation converges to the model's true semantic answer distribution. We use a higher number of samples $K = 30$ to ensure a reasonable approximation. Semantic clustering follows the procedure of Kuhn et al. (2023), employing a bi-directional entailment check with the *deberta-v2-xlarge-mnli* model He et al. (2021). Samples are drawn via multinomial sampling with the default temperature, top-p, and top-k settings of each model. This choice is deliberate, as different model families and versions (e.g., base vs. instruct) provide different defaults, and we aim to evaluate them under their most realistic production settings.

Calculation of Epistemic Uncertainty $KL(p^* \,\|\, p)$ The distribution p^* defines probabilities over the set of semantically distinct correct answers. Since the model distribution p is sampled and may be arbitrary, their supports need not coincide. Moreover, matching classes may also differ in surface form. As such, they need to be *aligned* to be able to calculate the epistemic uncertainty. As an example, consider:

$$p^* = \{\text{Heat} : 0.3, \text{Fuel} : 0.34, \text{Oxygen} : 0.36\}$$
$$p = \{\text{It's Heat} : 0.4, \text{Carbon} : 0.2, \text{Oxygen} : 0.4\}.$$

We construct a joint support set $\{\text{Heat}, \text{Fuel}, \text{Oxygen}, \text{Carbon}\}$, imputing missing values with 0 in p^* and with $\epsilon = 0.01$ in p to avoid undefined terms in the KL-divergence due to $\log(0)$. Using ϵ for the model distribution is justified, since in principle the model assigns non-zero probability to any possible sequence, making the support of p^* always a subset of the support of p. To determine the common support set, we apply the same semantic clustering procedure used for estimating p, based on bidirectional entailment with *deberta-v2-xlarge-mnli* He et al. (2021).

B.2 UQ Estimators

Semantic Entropy (SE) For semantic entropy, we follow Kuhn et al. (2023) The method first estimates the semantic distribution p as outlined in Section B.1 using K samples, and then computes the entropy:

$$H(p) = -\sum_{i=1}^{|C|} p_i \log p_i.$$

Maximum Sentence Probability (MSP) A simple yet effective estimator is the maximum sentence probability (MSP), defined as:

$$\text{MSP} = 1 - \max_a \; p(a \mid x),$$

where $p(a \mid x)$ is the probability assigned to answer a. Importantly, we do not compute $\max_y p(y \mid x)$ from the semantic distribution p estimated above; instead, we directly perform beam search with 5 beams to identify the highest-probability answer. This approach is similar to a recent proposal by Aichberger et al. (2024)

Shifting Attention to Relevance (SAR) Instead of having hard clusters, SAR computes continuous semantic similarity scores to determine the importance of samples. Additionally, SAR mitigates the influence of irrelevant tokens by calculating the importance of each token on the semantics of the answer (Duan et al., 2024). We use the implementation of (Vashurin et al., 2025) using *cross-encoder/stsb-roberta-large* as the semantic similarity model and $K = 30$ samples.

Iterative Prompting (IP) The proposed estimator (Yadkori et al., 2024) should not be confused with the traditional MI estimator (Depeweg et al., 2018). The core idea behind the method is based on the idea that if a model is epistemically certain, it is less likely to change its answer by the inclusion of a wrong answer in the input context. For a detailed explanation of this method, we refer to Yadkori et al. (2024). In our implementation, we limit the number of samples to $K = 10$. Conditional probabilities are obtained via teacher forcing and extracted explicitly from the model output. We use hyperparameters $\gamma_1 = \gamma_2 = 10^{-9}$ and employ the prompt schema shown in Prompt 3 to obtain the conditional probabilities.

B.3 Inference Prompts

For base models, we employ few-shot prompts to guide the model toward producing answers in the desired format (Prompt 1). In contrast, instruct models are queried with a single instruction that specifies the expected answer style (Prompt 2).

Prompt 1: Prompt for base models.

```
Q: What is one planet in our solar system that has rings?
A: Saturn

Q: Name one programming language you know.
A: Python

Q: Who is one of the singers in the band ABBA?
A: Agnetha Faeltskog

Q: What is one color in the German flag?
A: Black

Q: {question}?
```

```
A:
```

Prompt 2: Prompt for instruct models.

```
Answer the following question with one word or phrase:
{question}?
```

Prompt 3: Prompt for MI estimator

```
A possible answer to the question {question} is {answer}.
Q: {question}?
A:
```

C Dataset Creation

Our dataset construction process consists of the following steps:

Question Rephrasing: Each original question is reformulated to explicitly request exactly one specific answer. E.g.: *"What are the essential components of the fire triangle?"* →*"What is one essential component of the fire triangle?"*. This prevents the model from producing multiple answers in a single generation. The rephrasing is done with `gpt-4.1-mini`.

Keyword Extraction: To enable the co-occurrence search, we extract a main keyword for the co-occurrence search. The keyword can either be a single word, like the subject, or a phrase. Critically, the co-occurrence of the keyword and the answer should reliably indicate the presence of the fact in the retrieved document. This is a valid assumption in most cases, as Elsahar et al. (2018) shows that when only the subject and object of a subject-object-relation triple co-occur in text, the resulting triple is often also present. However, for our main dataset *Wikipedia English*, we take additional measures to enhance quality as explained Section C.1. The keyword extraction is done using `gpt-4.1-mini` with Prompt 5 - except for the proxy using The Pile, which employs entity linking.

Co-occurrence Search: For each question, we perform a co-occurrence search for each answer on the proxy corpora. The final ground-truth distribution $p^*(\cdot|q)$ for a given question q is then obtained by the relative frequency of the individual answer counts to all answer counts. To reduce potential biases, we discard samples in which at least one candidate answer has zero counts. Due to this fact, using the different proxies *English Wikipedia*, *RedPajama-V1*, and *The Pile* can result in different samples in the final datasets.

C.1 Wikipedia English

Dataset curation We use the structured Wikipedia Wikimedia Enterprise (2024) dataset, and specifically the English version, which consists of all English article pages in a structured way. For each article, we are leveraging all data in the *sections* tag. For the co-occurrence search, we use Pyserini and build the search index locally Lin et al. (2021). To define what constitutes a document—i.e., how articles are chunked for indexing—we leverage the dataset's hierarchical structure: articles are organized into sections and subsections down to the level of individual paragraphs or lists. We assume that relevant facts are contained at this lowest level, which represents a coherent unit of text. The average length of the resulting chunks is around 65 words, with the distribution following a power-law: fewer than 1% of the chunks exceed 300 words, while only a small number of outliers contain more than 2000 characters (≈400 words). For such extreme outliers, we apply additional splitting at sentence boundaries. Importantly, apart from these rare cases, we keep the chunks intact and do not split them further, ensuring high recall of facts. Importantly, we also apply stemming to reduce words to their base forms, avoiding reliance on overly specific surface forms. The final index contains 65,069,586 documents.

Co-occurrence counting In the retrieval step, we return all documents containing both the keyword and the candidate answer for a given question. Because the relationship between a question and its answer can be complex, relying on a single keyword often yields high recall but only moderate

precision. For instance, consider the question *"Who is the founder of Apple?"*—one valid answer is *Steve Jobs*. If we extract *Apple* as the main keyword, then any fact expressing *"Steve Jobs founded Apple"* will naturally contain both *Steve Jobs* and *Apple*, which ensures high recall. However, the mere co-occurrence of *Steve Jobs* and *Apple* does not always capture the intended fact—for example, *"Steve Jobs was the CEO of Apple"*. Such cases reduce precision. Hence, to ensure high precision, we apply an entailment procedure. Given a retrieved document through the co-occurrence search, we pass it to an LLM to verify that the fact is indeed present. For this step, we use *Gemma-3 12B Instruct* with the prompt shown in Prompt 4 and examples in Table 9. To keep the entailment step computationally feasible, we cap the number of retrieved documents per candidate answer at 1000—a threshold that we observe is rarely exceeded. The final number of samples for MAQA* is 468 and for AmbigQA* 2553 (Table 8).

Prompt 4: Prompt for entailment check.

```
You are an expert at verifying factual entailment. I.e., is the fact
present in the text?
Given the following TEXT and FACT, answer with "yes" if the FACT
follows from the TEXT, or "no" if it does not.

TEXT: {text}
FACT: {fact}
Answer:
```

C.2 RedPajama-V1

Dataset curation The Infini-Gram API provides access to co-occurrence counts across a range of large-scale pre-training datasets Liu et al. (2024). We use *RedPajama-v1* Weber et al. (2024), which closely replicates the LLaMA pre-training corpus and includes a diverse set of data sources.

Co-occurrence counting Similarly, as for Wikipedia English, we query for co-occurrences of the keyword with each candidate answer. For the Infini-Gram API we use the parameters *max_diff_tokens*$= 100$ and *max_clause_freq*$= 50000$. Since the underlying tokenizer (LLaMA 2) is sensitive to whitespaces for a keyword answer pair, we test all four combinations of including or removing a whitespace at the beginning of the keyword or answer. To obtain the final counts, we sum up the retrieved counts of the four different possibilities. Due to limited document access in Infini-Gram, we do not perform an entailment-checking phase. The final number of samples for MAQA* is 470 and for AmbigQA* 2331 (Table 8).

C.3 The Pile

Dataset Curation In contrast to the previous two approaches, this method follows a different strategy for obtaining keywords and answers. It relies on entity linking, which identifies entities such as people, cities, or songs in both the question and the answer. The co-occurrence of a question entity with an answer entity is then retrieved from the Pile corpus Gao et al. (2020). Following the approach of Kandpal et al. (2023), we use the DBpedia Spotlight entity linker Daiber et al. (2013) to extract entities from questions and answers. To improve accuracy, each answer is appended to its corresponding question before entity linking. When multiple candidate entities are returned for a question, we employ *Gemma-3 12B Instruct* to filter for the most relevant one. The linker's parameters are set to *confidence* $= 0.4$ and *support* $= 1$.

Co-occurrence counting After obtaining the entity sets, we match them with pre-extracted entities from The Pile provided by Kandpal et al. (2023) to compute co-occurrence statistics. Similarly, as for RedPajama-V1, we do not perform an entailment-checking phase as we do not have access to the underlying documents.The final number of samples for MAQA* is 120 and for AmbigQA* 861 (Table 8).

C.4 Characteristics

Summary statistics are reported in Table 8. Compared to Wikipedia English, the other two strategies have access to a substantially larger pre-training corpus and therefore yield considerably higher

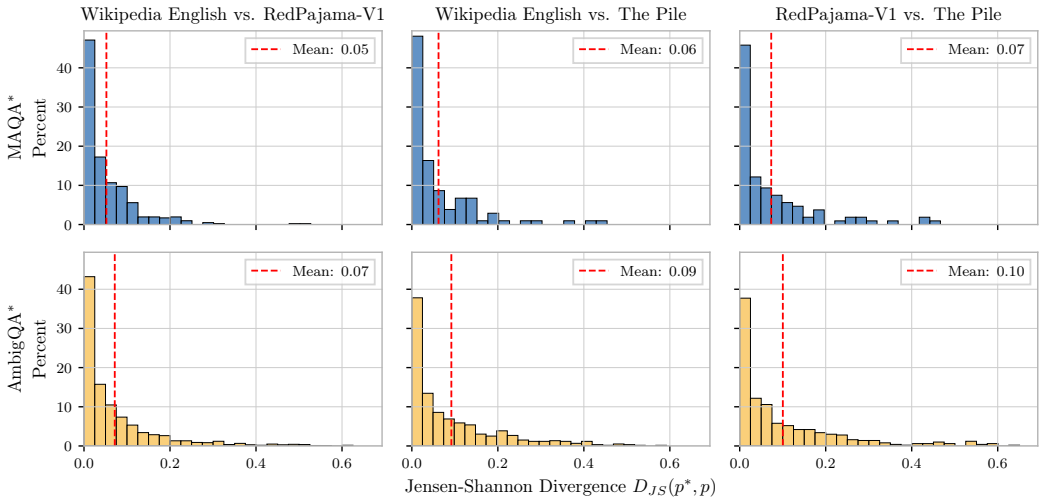

Figure 5: Comparison of retrieved ground-truth distribution p^* using different strategies

counts. Nevertheless, the average entropies and their standard deviations remain in a similar range. As mentioned previously, we use *English Wikipedia* as our principal strategy since it is the most controlled method with entailment checking, ensuring high precision and high recall. As can be seen in Table 8, it also provides most samples on AmbigQA* and similarly many on MAQA* as RedPajama-V1. Using The Pile, in contrast, produces significantly fewer samples compared to the other two methods, as entity linking often can't find an entity in either question or answer, and hence such samples have to be discarded. To assess how well the estimated ground truths p^* align across datasets, we compute the Jensen-Shannon divergence for all pairwise couplings on MAQA* and AmbigQA* (Figure 5). The Jensen-Shannon divergence is given by: $\mathrm{JS}(p \,\|\, q) = \frac{1}{2}\left[\mathrm{KL}(p \,\|\, m) + \mathrm{KL}(q \,\|\, m)\right]$,where $m = \frac{1}{2}(p + q)$. It has the useful property of being symmetric, as we do not consider one strategy over the other as the truth. Overall, all strategies produce largely consistent ground truths, as reflected in the low average JS divergence and the characteristic power-law distribution (Figure 5).

Table 8: Summary statistics for p^* estimation strategies: samples n, mean answer-counts, and mean entropies (mean $\pm$ std).

p^* **Estimation Method**	**MAQA***			**AmbigQA***		
	n	**Count** $\pm$ **SD**	$\overline{H}$ $\pm$ **SD**	**n**	**Count** $\pm$ **SD**	$\overline{H}$ $\pm$ **SD**
Wikipedia English	468	115.49 $\pm$ 178.63	1.11 $\pm$ 0.44	2553	55.81 $\pm$ 104.77	0.64 $\pm$ 0.33
RedPajama-V1	470	143066.88 $\pm$ 1234461.86	0.99 $\pm$ 0.44	2331	1220812.64 $\pm$ 29688717.59	0.52 $\pm$ 0.35
The Pile	120	46281.74 $\pm$ 138909.80	0.97 $\pm$ 0.44	861	19881.87 $\pm$ 55441.02	0.60 $\pm$ 0.33

D Related Work

UQ for LLMs A wide range of methods for uncertainty quantification in LLMs have been proposed (Vashurin et al., 2025; Liu et al., 2025). Many methods rely on the predictive distribution p. The most prominent approaches here quantify the variation in p, with Semantic Entropy (Kuhn et al., 2023) being the most widely adopted, alongside variants such as Duan et al. (2024); Nikitin et al. (2024).

Ambiguity in QA Tasks Previous work are benchmarked on QA datasets like TriviaQA (Joshi et al., 2017) which only contain a single correct answer per question (Devic et al., 2025). Few works consider the presence of aleatoric uncertainty. Hou et al. (2024) introduce aleatoric uncertainty through ambiguity in the question's phrasing. Crucially, this does not cover the case where the ambiguity is inherent to the answer. Yadkori et al. (2024) proposes a method based on the idea that an epistemically confident model should be less likely to be misled by the inclusion of a wrong answer in the input context. While their estimator is theoretically well argued for by analyzing the joint distribution of responses, their assumption on the LLM behavior (*Assumption 5.3*) seems not to hold

in reality, which is supported by the emerging research field of knowledge conflicts (Xie et al., 2024; Xu et al., 2024). Correspondingly, our results show that the method is ineffective under ambiguity as well.

The absence of evaluations under ambiguity is a consequence of the lack of suitable benchmarks. Only few datasets explicitly consider ambiguous questions, namely AmbigQA (Min et al., 2020) and MAQA Yang et al. (2025). To our knowledge, MAQA is the only dataset with questions for which we ambiguity is inherent to the task and can not be resolved with a more precise phrasing. It, however, does not quantify the true distribution p^* and, therefore, can not be used for a quantitative study on UQ under ambiguity.

E Proofs

Proposition 1 (Non-Identifiability of Epistemic Uncertainty)**.** *Let $K \geq 2$ and Δ^{K-1} be the probability simplex over K classes. For any function $f : \Delta^{K-1} \to \mathbb{R}$ and any $p \in \Delta^{K-1}$, there exist $p_1^*, p_2^* \in \Delta^{K-1}$ such that*

$$\mathrm{KL}(p_1^* \,\|\, p) = 0 \quad \textit{and} \quad \mathrm{KL}(p_2^* \,\|\, p) = -\log \min_i p_i \geq \log K,$$

Thus, any function $f(p)$ can both indicate zero or high ($\geq \log(K)$) epistemic uncertainty.

Proof. Fix $p \in \Delta^{K-1}$. Set $p_1^* := p$. Then $\mathrm{KL}(p_1^* \parallel p) = 0$. Let $j \in \arg\min_i \; p_i$ and define $p_2^* := \mathbf{1}[y = j]$. Then

$$\mathrm{KL}(p_2^* \,\|\, p) = -\log p_{\min}.$$

Thus, for the same p, EU can be 0 or large, while $f(p)$ is fixed. □

Proposition 2 (Zero aleatoric uncertainty implies EU is NLL)**.**

$$H(p^*) = 0 \implies EU = -\log(p(y = y^*))$$

Proof. If $H(p^*) = 0$, then $p^*(y) = \mathbf{1}[y = y^*]$. From this it follows:

$$EU = KL(p^* \,||\, p) = -\sum_{y \neq y^*} 0 \log(p(y)) - \log(p(y = y^*)) = -\log(p(y = y^*))$$

□

Theorem 1 (High Entropy ⇒ High Epistemic Uncertainty)**.** *Let there be $K \geq 2$ classes and $\delta \in [0, \log K]$ be a threshold on the entropy. Furthermore, let α_δ be the maximal possible probability on some class s.t. $H(p) \geq \delta$. Then the epistemic uncertainty with $H(p) \geq \delta$ is at least:*

$$EU \;\geq\; -\log \alpha_\delta.$$

Proof. We first define α_δ mathematically and how to obtain it.

$$\alpha_\delta \;=\; \max\Big\{ \max_j p_j \;:\; H(p) \geq \delta \Big\}, \qquad \delta \in [0, \log K].$$

Let $H_{max}(\alpha) = -\alpha \log \alpha - (1 - \alpha) \log \frac{1-\alpha}{K-1}$. This is the maximum entropy achievable by a distribution whose largest class probability is $\alpha \in [1/K, 1]$. Then α_δ is the solution of $H_{max}(\alpha) = \delta$. Now we seek the lowest possible $EU = -log(p(y = y^*))$ under the constraint $H(p) \geq \delta$. This exactly occurs if the maximal possible probability α_δ is on the correct class and hence $EU = -\log(p(y = y^*)) \geq -\log(\alpha_\delta)$ □

Theorem 2 (Low Entropy ⇒ Low Epistemic Uncertainty with High Probability)**.** *Let there be $K \geq 2$ classes and $\delta \in [0, \log 2]$ be a threshold on the entropy. Furthermore let $\mathcal{L} = \mathbb{E}_{(x,y)}[-\log p_y]$ the average loss and γ_δ be the minimal possible maximum probability on some class s.t $H(p) \leq \delta$. Then the probability that the epistemic uncertainty with $H(p) \leq \delta$ will be less than $-\log(\gamma_\delta)$ satisfies:*

$$\mathbb{P}(EU \leq -\log(\gamma_\delta) \mid H(p) \leq \delta) \geq 1 - \frac{\mathcal{L}}{-\log(1 - \gamma_\delta) * \mathbb{P}(H(p) \leq \delta)}$$

Proof. We first define γ_δ mathematically and how to obtain it:

$$\gamma_\delta \;=\; \min\Big\{ \max_j p_j \;:\; H(p) \leq \delta \Big\}, \qquad \delta \in [0, \log 2],$$

Denote $H_B(\gamma) = -\gamma \log \gamma - (1-\gamma)\log(1-\gamma)$ as the binary entropy function. Then γ_δ is the solution of $H_B(\gamma) = \delta$ for $\gamma \in [1/2, 1]$ and we can now proceed:

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_{(x,y^*)}\big[-\log p_{y^*}\big] && (1)\\
&= \mathbb{E}_{(x,y^*)}\big[-\log p_{y^*} \mid H(p) \leq \delta\big]\mathbb{P}(H(p) \leq \delta) && (2)\\
&\quad + \mathbb{E}_{(x,y^*)}\big[-\log p_{y^*} \mid H(p) > \delta\big]\mathbb{P}(H(p) > \delta)\\
&= \mathbb{E}_{(x,y^*)}\big[-\log p_{y^*} \mid H(p) \leq \delta \cap \arg\max p \neq y^*\big]\mathbb{P}(H(p) \leq \delta \cap \arg\max p \neq y^*) && (3)\\
&\quad + \mathbb{E}_{(x,y^*)}\big[-\log p_{y^*} \mid H(p) \leq \delta \cap \arg\max p = y^*\big]\mathbb{P}(H(p) \leq \delta \cap \arg\max p = y^*)\\
&\quad + \mathbb{E}_{(x,y^*)}\big[-\log p_{y^*} \mid H(p) > \delta\big]\mathbb{P}(H(p) > \delta)\\
&\geq -\log(1-\gamma_\delta)\mathbb{P}(H(p) \leq \delta \cap \arg\max p \neq y^*) - \log(\alpha_\delta)\mathbb{P}(H(p) > \delta) && (4)
\end{aligned}$$

Where we use in 2 the law of total expectation to separate into high and low entropy predictions. In 3, we further partition the space of low entropy predictions into correct and incorrect ones. Lastly, in 4, we bound the expectation values. High entropy predictions occur at least loss $-\log(\alpha_\delta)$ according to theorem 1. Low entropy predictions that are incorrect will have maximally $1-\gamma_\delta$ mass on an *correct* class and as such occur at least $-\log(1-\gamma_\delta)$ loss. Rearranging terms and substituting $\mathbb{P}(H(p) > \delta) = 1 - \mathbb{P}(H(p) \leq \delta)$yields

$$\mathbb{P}(H(p) \leq \delta \cap \arg\max p \neq y^*) \leq \frac{\mathcal{L} + (1 - \mathbb{P}(H(p) \leq \delta))\log(\alpha_\delta)}{-\log(1-\gamma_\delta)}$$

Dividing by $\mathbb{P}(H(p) \leq \delta)$ we finally get the conditional bound:

$$\mathbb{P}(\arg\max p \neq y^* \mid H(p) \leq \delta) \leq \frac{\mathcal{L} + (1 - \mathbb{P}(H(p) \leq \delta))\log(\alpha_\delta)}{-\log(1-\gamma_\delta)\mathbb{P}(H(p) \leq \delta)}$$

which can be rewritten to obtain an upper bound as:

$$\begin{aligned}
\mathbb{P}(\arg\max p = y^* \mid H(p) \leq \delta) &\geq 1 - \frac{\mathcal{L} + (1 - \mathbb{P}(H(p) \leq \delta))\log(\alpha_\delta)}{-\log(1-\gamma_\delta)\mathbb{P}(H(p) \leq \delta)} && (5)\\
&= 1 - \frac{\mathcal{L}}{-\log(1-\gamma_\delta)\mathbb{P}(H(p) \leq \delta)} && (6)\\
&\quad + \frac{-\log(\alpha_\delta)(1 - \mathbb{P}(H(p) \leq \delta))}{-\log(1-\gamma_\delta)\mathbb{P}(H(p) \leq \delta)} && (7)
\end{aligned}$$

Realizing that $-\log(p_{y^*}) \leq -\log(\gamma_\delta) \iff \arg\max p = y^*$ - since γ_δ is the minimum possible maximum probability - we get:

$$\begin{aligned}
\mathbb{P}(\log p_{y^*} \leq -\log(\gamma_\delta) \mid H(p) \leq \delta) &\geq 1 - \frac{\mathcal{L}}{-\log(1-\gamma_\delta)\mathbb{P}(H(p) \leq \delta)} && (8)\\
&\quad + \frac{-\log(\alpha_\delta)(1 - \mathbb{P}(H(p) \leq \delta))}{-\log(1-\gamma_\delta)\mathbb{P}(H(p) \leq \delta)} && (9)
\end{aligned}$$

Abbreviating $-\log p_{y^*}$ as *epistemic uncertainty* EU and simplifying by leaving out the second term, we obtain the bound stated in the theorem.

$$\mathbb{P}(EU \leq -\log(\gamma_\delta) \mid H(p) \leq \delta) \geq 1 - \frac{\mathcal{L}}{-\log(1-\gamma_\delta) * \mathbb{P}(H(p) \leq \delta)} \tag{10}$$

□

E.1 Non-trivial aleatoric uncertainty

When constraining $H(p^*) = 0$, we implicitly restrict p^* to be a an indicator vector over one of the K classes. As shown in Theorems 1 and 2, this setting allows for informative bounds on epistemic uncertainty. However, this is only one case. Consider instead the situation where p^* is known exactly. While this assumption is unrealistic (since complete knowledge of p^* makes estimation redundant), it helps to illustrate non-trivial aleatoric uncertainty. For example, if p^* is uniform, we obtain maximal aleatoric uncertainty with $H(p^*) = \log K$. However, we can, in fact, exactly determine the epistemic uncertainty:

$$EU = KL(p^* \mid\mid p) = \sum_y \frac{1}{K} \log(\frac{1}{Kp(y)}) = -\log(K) - \frac{1}{K} \sum_y \log(p(y))$$

Similarly when relaxing the constraint slightly to allow p^* be a high entropy distribution (e.g., $H(p^*) \in [\log K - \epsilon, \log K]$) estimation of epistemic uncertainty using $H(p)$ should work reasonably: low predictive entropy necessarily implies high epistemic uncertainty, whereas high predictive entropy indicates lower epistemic uncertainty.

These illustrations clarify what we mean by *non-trivial* aleatoric uncertainty: cases where no strong restrictions on $H(p^*)$ are imposed. This is the typical regime in realistic applications, since constraining $H(p^*)$ would require prior knowledge about the ambiguity structure of the task itself. This is especially the case in many linguistic problems, as a specific language task can have an arbitrary structure.

Table 9: Examples of entailment check in the co-occurrence pipeline for Wikipedia English

Idx	Question	Keyword	Answer	Positive Example	Negative Example
72	Who was the recipient of the Bharat Ratna award when it was first awarded?	Bharat Ratna	['C. Rajagopalachari']	article rajaji national park, section abstract: rajaji national park was named after c. rajagopalachari (rajaji), a prominent leader of the freedom struggle, the first and last governor-general of independent india and one of the first recipients of india's highest civilian award, bharat ratna (in 1954).	article central college, bangalore, section notable students: bharat ratna sir m. visvesvaraya, harshavardhan mudaliar, prof in english, bharat ratna c. rajagopalachari, bharat ratna c. n. r. rao, indian chemist, shivakumara swami, pusapati vijayarama gajapati raju, maharaja of vizianagaram, h. narasimhaiah, guruswami mudaliar, hospet sumitra, n. santosh hegde, justice, navaratna rama rao, leading administrator, author and founder of the sericulture department, n. s. subba rao, maya rao (1928-20...
186	What is one specific type of agricultural product the Wachau Valley is known for?	Wachau Valley	['grapes']	article wachau, section wine: the wachau valley is well known for its production of apricots and grapes, both of which are used to produce specialty liquors and wines. the wine district's rolling vineyards produce complex white wines. wachau is a source of austria's most prized dry rieslings and grüner veltliners, some of the best from the steep stony slopes next to the danube on which the vines are planted. the temperature variation in the valley between day and cold nights has a significant ro...	No negative example found
198	What is the name of one Unforgivable Curse from the Harry Potter books?	Unforgivable Curses	['Imperio']	article imperio, section abstract: imperio, a curse in the harry potter series (see magic in harry potter#unforgivable curses), imperio (band), austrian band	No negative example found
205	Who was one of the main cast members of 'The Big Valley' TV show?	The Big Valley	['Barbara Stanwyck']	article the big valley, section reception : popularity: in the comedy film airplane! (1980), the wacky air traffic controller johnny, played by stephen stucker, paid homage to big valley ' s penchant for big drama in one of his many asides. after lloyd bridges ' character frets about a pilot who cracked under pressure, johnny says: "it happened to barbara stanwyck!" and "nick, heath, jarrod – there's a fire in the barn!" the big valley also has seeped into the darker cinematic subconscious. in b...	article peter breck, section career : after the big valley: on january 20, 1990, while teaching at the drama school, breck was notified of barbara stanwyck's death. she requested no funeral nor memorial.
297	Which stadium did the New Orleans Saints use for their home games in the seasons following Hurricane Katrina in 2005?	New Orleans Saints	['Alamodome']	article tom benson, section biography : new orleans saints : saints relocation controversy: when it became clear that hurricane katrina 's extensive damage to new orleans and the superdome would make it impossible for the saints to play there in 2005, the team temporarily relocated its operations to san antonio and began negotiations to play home games at the alamodome. (the saints, after discussions with the nfl and louisiana state university, eventually agreed to play one "home" game at giants...	article 2001 minnesota vikings season, section preseason : game summaries : week 1: at new orleans saints: at alamodome, san antonio, texas

Prompt 5: Prompt for keyword extraction.

```
You are a keyword extraction assistant helping to identify the
keywords in a question for a co-occurrence search.
The goal is to check how often the answer to a specific question
(fact) appears in a text corpus.
To do this, you must identify the keywords in the question that are
needed to find the fact in the text corpus.

Your job is to analyze a question/answer pair and pull out:
- The minimal term(s) that, when paired with the known answer entities
    , reliably locate the same fact in a text corpus.
- The goal is to have as few terms as possible while still being able
    to find the fact.

Guidelines:
- Extract the main keyword from the question that shrinks the search
    space.
  E.g., for a song title question, the main keyword is the title of
      the song.
- Extract additional keywords needed to find the fact in a text corpus
    .
  E.g., for a song title, additional keywords are the artist and album
      .
- The main keyword should be a single term or short phrase that
    captures the essence of the question.
- Additional keywords should be a short list of terms (not too long).

Return exactly this JSON (no extra fields or explanation):

{
  "main_keyword":  [string],
  "additional_keywords":  [ string, .. ]
}


Example 1
Input:
Question: "Who were the writers of the song 'Tell Your Heart to Beat
    Again'?"
Answer:   "Bernie Herms, Mathew West, Randy Phillips"
Output:
{
    "main_keyword":  ["Tell Your Heart to Beat Again"],
    "additional_keywords":  ["writers"]
}

Example 2
Input:
Question: "What are the names of recognized dwarf planets in the solar
     system as of 2024?"
Answer:   "Ceres, Eris, Pluto, Makemake, Haumea"
Output:
{
    "main_keyword":  ["dwarf planet"],
    "additional_keywords":  ["solar system"]
}

Example 3
Input:
Question: "What is the legal age of marriage in the United States?"
Answer:   "18, 19, 21"
Output:
{
    "main_keyword":  ["marriage"],
```

```
    "additional_keywords":  ["legal age", "United States"]
}


Now process the following and produce **only** the JSON:

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