AUB-WIKI: BENCHMARKING UNCERTAINTY DECOMPOSITION IN LANGUAGE MODELS UNDER CONTROLLED AMBIGUITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Uncertainty in Language Model (LM) answers is commonly decomposed into aleatoric uncertainty, arising from ambiguity in the question, and epistemic uncertainty, arising from missing model knowledge. Evaluating such decompositions requires a benchmark on which the two sources can be measured separately. Existing ambiguous-question datasets do not allow this: at best, they provide ground truth for ambiguity, and they lack an independent measure of knowledge difficulty for each clarification, so a method's aleatoric and epistemic estimates cannot be checked against separate references. We introduce Ambiguity and Uncertainty Benchmarking Wiki (AUB-Wiki), a benchmark of referentially ambiguous questions generated deterministically from Wikidata without any language model in the loop. Each question is paired with the complete set of clarifications implied by the knowledge graph, and each clarification has a verified answer and a model-independent coverage score. Most questions also have a matched unambiguous control. The number of clarifications and their coverage form two near-independent axes of ambiguity and knowledge difficulty. We use AUB-Wiki to benchmark uncertainty quantification and uncertainty decomposition methods across several open-weight language models, evaluating whether they attribute uncertainty to ambiguity or to incomplete knowledge. The benchmark and generation pipeline are released to support controlled evaluation of uncertainty decomposition.

1 INTRODUCTION

When large language models are deployed in knowledge-intensive and high-stakes settings (Aoki, 2024; Singhal et al., 2023), it matters not only whether a model is uncertain, but also where this uncertainty originates from. Uncertainty can reflect either missing knowledge or ambiguity in the input, and these sources may require different mitigation techniques, such as information retrieval or clarification. A recent line of work therefore proposes to decompose uncertainty into an aleatoric component associated with ambiguity and an epistemic component associated with missing knowledge (Hou et al., 2024; Walha et al., 2026), as an extension of uncertainty quantification methods such as semantic entropy (Kuhn et al., 2023) and $P(\mathrm{True})$ (Kadavath et al., 2022) that measure only total uncertainty. Evaluating these decompositions therefore requires testing uncertainty disentanglement: whether each component responds to the source it is intended to measure.

The distinction between aleatoric and epistemic uncertainty is well established in supervised learning (Hüllermeier & Waegeman, 2021), but a formal decomposition does not by itself establish that its estimated components separate the two sources in practice. In image classification, Mucsányi et al. (2024) found that commonly used decompositions produced strongly correlated components despite their different intended meanings. This motivates evaluating each component against both sources of uncertainty. In LLM question answering, such an evaluation requires reference information about both the interpretations a question permits and the model's knowledge of the corresponding answers.

Consider a question such as "In what year was the film directed by Kevin Tancharoen released?" Several films may satisfy this description, and a model may know some of them but not others. This simple example illustrates why evaluating the quality of uncertainty estimates requires a dataset that contains (i) the objectively quantified number of valid interpretations (as graded measure of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ambiguity) and (ii) objectively graded difficulties (quantifying how well the corresponding fact is known) that allow to test the model's knowledge of each interpretation. Existing benchmark datasets used to benchmark uncertainty quantification (UQ) and ambiguitiy detection methods do not allow for this. In AmbigQA (Min et al., 2020) and its descendants (Tomov et al., 2025; Stelmakh et al., 2022; Lee et al., 2023), the interpretations of a question are those subjectively identified by its annotators, so there is no objective rule against which these interpretations could be checked. These datasets also do not record how well known individual interpretations are, and their unambiguous questions are not constructed as matched controls for the ambiguous ones. A method can therefore appear to detect ambiguity by responding to differences in question difficulty or model knowledge, without the evaluation revealing which factor it has captured.

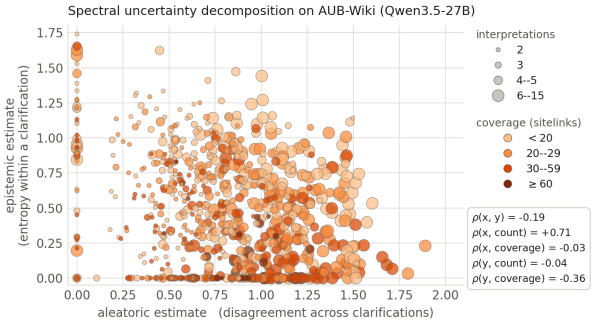


Figure 1: Evaluation of spectral uncertainty decomposition on the ambiguous records of AUB-Wiki (Qwen3.5-27B), enabled by the ambiguity and knowledge axes. The figure reports Spearman's ρ between the two estimates and with each axis. The dataset shows that the two estimates are fairly decorrelated. The aleatoric estimate is correlated with the ground-truth aleatoric signal (number of interpretations; larger circles to the right), and the epistemic estimate is inversely correlated with the coverage measure (darker circles to the bottom).

In this work, we introduce Ambiguity and Uncertainty Benchmarking Wiki (AUB-Wiki), a benchmark that records both the interpretations of a question and the coverage of the corresponding entities. Each interpretation set is defined by an explicit membership rule over Wikidata and audited for completeness within that rule. Each interpretation is paired with a clarification question, its answer, and its number of Wikimedia sitelinks (which we refer to as coverage). Interpretation count defines the *ambiguity axis*, while coverage provides a model-independent proxy for the *knowledge axis*. We test this proxy by measuring whether it predicts model accuracy on the clarifications. Where available, single-interpretation controls preserve the question template and approximately match coverage. This design allow each uncertainty component to be tested against both axes: whether it is specific to its intended source or if it also varies with the other (see Figure 1). The benchmark dataset is generated deterministically without language models, with provenance retained throughout.

Our contributions are summarized as follows:

- **AUB-Wiki, a reproducible and scalable benchmark dataset for referential ambiguity.** We introduce 2,158 ambiguous questions, complete relative to an explicit interpretation rule and audited for membership, with 1,425 matched unambiguous single-interpretation controls. Each interpretation is paired with a clarification, a ground-truth answer, and Wikipedia coverage.
- **A two-axis evaluation protocol for uncertainty.** We introduce the first benchmark design for uncertainty quantification methods that measures aleatoric and epistemic uncertainty on the same questions. This makes it possible to test not only whether a UQ method responds to ambiguity or missing knowledge, but also whether the estimated components are specific to their intended sources and to what extent they are disentangled from each other.
- **An empirical evaluation of uncertainty quantification and decomposition.** We evaluate five uncertainty quantification and decomposition methods across five open models from three model families and one frontier model, showing how their signals behave as ambiguity and knowledge coverage vary.

2 Related Work

Uncertainty quantification for LLMs. A large body of work estimates the total uncertainty of an LLM's answer from the model's own outputs. *Semantic entropy* (SE) (Kuhn et al., 2023; Farquhar et al., 2024) samples several answers, groups them into clusters of equivalent meaning via bidirectional entailment, and computes the entropy over these clusters. *Kernel language entropy* (KLE) (Nikitin et al., 2024) replaces this hard clustering with a graph kernel encoding pairwise semantic similarities between answers and quantifies uncertainty as its von Neumann entropy (Von Neumann, 2018), which generalizes SE. Gruber & Buettner (2024) derive the *predictive kernel entropy* from a bias–variance–covariance decomposition of kernel scores. In contrast to these sampling-based measures, $P(\mathrm{True})$ (Kadavath et al., 2022) presents the model with a proposed answer and uses the probability it assigns to the answer being true as a confidence estimate.

Ambiguity datasets. Several existing datasets make ambiguity explicit. AmbigQA (Min et al., 2020) annotates open-domain questions from Natural Questions (Kwiatkowski et al., 2019) with the interpretations identified by crowd workers, each paired with a disambiguated question and its answer. AmbigInst (Hou et al., 2024) is a small synthetic dataset of ambiguous task instructions whose disambiguations and input-output pairs were LLM-generated. Tomov et al. (2025) extend AmbigQA and MAQA (Yang et al., 2025) with ground-truth answer distributions estimated from factual co-occurrence and show that existing uncertainty estimators degrade to near-random performance on ambiguous questions. These datasets provide a reference for aleatoric uncertainty, i.e., whether a question is ambiguous or which answers it admits, but none provides a model-independent reference for epistemic uncertainty, i.e., for how well each interpretation is known.

Epistemic uncertainty evaluation. In classification, epistemic uncertainty is commonly evaluated through the proxy task of out-of-distribution (OOD) detection, in which inputs from a shifted distribution, such as corrupted images, should receive higher epistemic uncertainty than in-distribution inputs (Ovadia et al., 2019; Mukhoti et al., 2023; Mucsányi et al., 2024). This proxy is difficult to transfer to LLMs: the content of their web-scale pretraining corpora is only partially understood (Elazar et al., 2024), and even with full access to the training data, membership inference attacks barely outperform random guessing, partly because the boundary between members and non-members is inherently fuzzy (Duan et al., 2024). It is therefore unclear which questions are in or out of distribution for a given LLM. AUB-Wiki instead provides a model-independent knowledge proxy, interpretation coverage, whose validity we test directly on the evaluated models (Section 3.3).

Uncertainty decomposition. The standard decomposition, following Depeweg et al. (2018), relies on a posterior or an ensemble over models (Lakshminarayanan et al., 2017): the entropy of the averaged prediction is split into the expected entropy of the individual members (aleatoric) and the mutual information between the prediction and the model parameters, i.e., the disagreement between members (epistemic) (Hüllermeier & Waegeman, 2021). Mucsányi et al. (2024) benchmarked such decompositions on ImageNet, using soft human labels as ground-truth aleatoric uncertainty and OOD detection as the epistemic task, and found that the resulting aleatoric and epistemic estimates are strongly rank-correlated (at least 0.78 across methods) and hence not disentangled. They note, however, that part of this correlation is inevitable on ImageNet: ambiguous images are undersampled relative to unambiguous ones and are therefore also more epistemically uncertain for classifiers trained on this dataset. For LLMs, training model ensembles is impractical, and approximating them by varying in-context examples fails to separate ambiguous from unambiguous questions (Hou et al., 2024). *Input clarification ensembling* (ICE) (Hou et al., 2024) therefore reverses the roles: instead of ensembling models to isolate epistemic uncertainty, it ensembles clarifications of the input, generated by a clarification LLM, to isolate aleatoric uncertainty. The aleatoric term is the mutual information between the answer and the clarification, i.e., the disagreement of answers across clarifications, while the epistemic term is the average answer entropy within each clarified input, where the ambiguity has been resolved. *Spectral uncertainty decomposition* (SD) (Walha et al., 2026) follows the same structure but replaces Shannon entropies over discrete answer clusters with von Neumann entropies of kernel matrices over answer embeddings, thereby accounting for fine-grained semantic similarity between answers. Both methods were evaluated on ambiguity detection for their aleatoric term and on correctness prediction for total uncertainty, whereas their epistemic terms have not been tested against a dedicated reference.

3 THE AUB-WIKI BENCHMARK DATASET

3.1 DESIGN

The unit of the dataset is a *record*. Each record follows a specific *template*, which consists of an *entity class*, a *grouping property*, and an *answer property*, all defined over Wikidata (Figure 3). As an example, consider the following template: the entity class is *films*, the grouping property is the *director*, and the answer property is the *release year*. A *group* is defined by one value of the grouping property, such as a particular director. The record's ambiguous question implicitly refers to the members of this group: *In what year was the film directed by Kevin Tancharoen released?* The question's *interpretations* are the members of the entity class (e.g., films) that have the specified value of the grouping property (e.g., a specific director) under the membership rule derived from Wikidata. In our example, this corresponds to all films directed by *Kevin Tancharoen* that are listed in Wikidata. Each interpretation is paired with a *clarification* that explicitly names the entity (i.e., the specific film in our example) and with the answer obtained from its answer property. Thus, the ambiguous question has one valid answer for each clarification. An example record is shown in Figure 2.

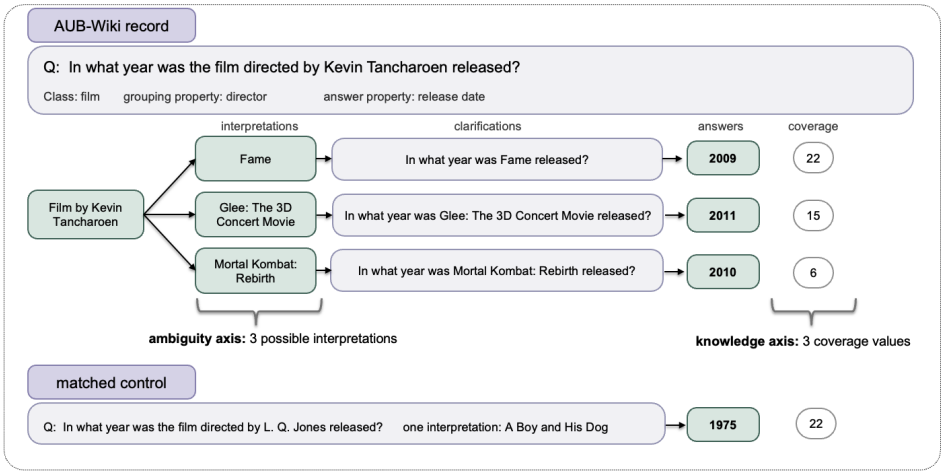


Figure 2: A record of AUB-Wiki and its matched unambiguous control. The group determines the interpretations; each carries a clarification of the question, its coverage in sitelinks, and the ground-truth answer. The number of interpretations is the ambiguity axis, and their coverage is the knowledge axis. The control shares the same template and a matched coverage but has only one interpretation (L. Q. Jones is credited as the director of a single film).

This design requires the grouping property to be *definitional*: it must describe an intrinsic relation to the entity, such as a film's director or a person's spouse, rather than an incidental relation such as an award or sponsor. It must also be sufficiently complete and approximately single-valued in Wikidata to support reliable group construction. The answer property must yield one exact value per interpretation. Candidate class–property pairs are screened against these requirements before being included in the templates, as described in Section 3.2.

Ambiguity and knowledge axes. The number of interpretations is the size of a record's group, making ambiguity quantitative rather than qualitative (unlike in AmbigQA (Min et al., 2020)). We include records with 2–15 interpretations; larger groups are excluded rather than truncated, so every interpretation set remains complete and small enough to enumerate, clarify, and answer in full. Interpretation count defines the benchmark's *ambiguity axis*. Each interpretation also has a *coverage* measure, given by its number of Wikimedia sitelinks.[1] We choose coverage to define the benchmark's *knowledge axis* as it reflects how widely the interpretation is documented, and therefore how often a model is likely to have encountered it in training. Wikipedia data is knowingly ubiquitous across documented LLM pretraining corpora (Liu et al., 2025; Elazar et al., 2024). This makes coverage a

[1] `wikibase:sitelinks` counts Wikimedia pages linked to the item, predominantly Wikipedia language editions.

model-independent proxy for how well the corresponding fact is known for LLMs. We later show that it predicts model correctness on clarified questions, and Appendix A.5 compares it with English page views and incoming links.

Matched controls. Where available, a record is paired with a *control*: a question of the same template whose description fits exactly one entity (and is therefore unambiguous), chosen so that its coverage matches the median coverage of the record's interpretations. The control therefore preserves the question form and knowledge coverage while reducing the interpretation count to one, providing a controlled comparison between ambiguous and unambiguous questions.

3.2 Construction process

We construct AUB-Wiki directly from Wikidata so that both interpretation count and coverage can be measured from the same source. Figure 3 summarizes the four-step process: template screening, generation, auditing, and release. No language model is used at any stage.

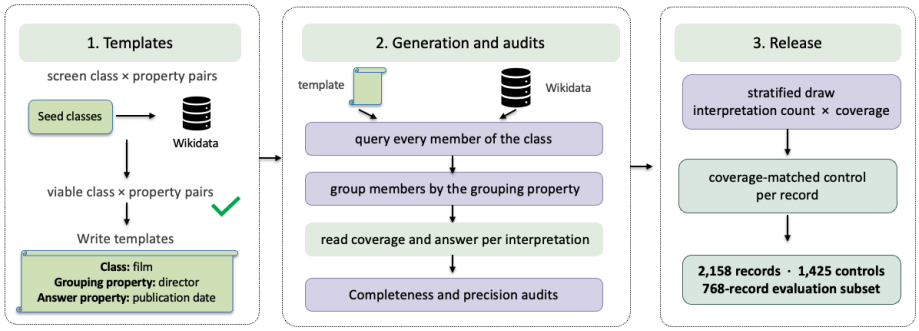


Figure 3: Construction of AUB-Wiki , generated from Wikidata (Vrandečić & Krötzsch, 2014) via the QLever endpoint (Bast & Buchhold, 2017): class and property pairs are screened into hand-written templates (left); each template is expanded against Wikidata into groups whose members are the interpretations, with their coverage and ground truth answers, and audited for completeness and precision (middle); records are then drawn stratified by interpretation count and coverage and matched to single-interpretation controls (right). No language model takes part.

Step 1: Template screening. For each candidate class we measure how completely its properties are populated and how many usable groups they induce. We retain class–property pairs that can define groups and supply exact answers. We then write the templates, fixing the class, properties, question and clarification wording, and answer type.

Step 2: Generation. Each template is expanded against Wikidata into groups whose members are the interpretations, together with their coverage and answers. Groups that satisfy the membership and quality rules become records if they contain 2–15 interpretations, while single-member groups become control candidates; groups are discarded rather than truncated when a required criterion fails.

Step 3: Audits. We audit the generated sets for missing interpretations, inexact answers, and ambiguous group names. Interpretation sets are checked against independently maintained English Wikipedia categories. Records with missing interpretations or ambiguous names are removed, whereas interpretations with an inexact answer (e.g. *around 1200 B.C.* instead of an exact year) are flagged.

Step 4: Release. From records whose interpretations all have an answer, we perform seeded stratified sampling across cells of interpretation count and coverage, while balancing topics and templates, preventing entity reuse and enforcing answer exactness. Each selected record is matched, where possible, to an available single-interpretation control with the closest coverage. This selection yielded a 2,158-record Tier-1 release. The audited pool of 6,734 records from which it is drawn is also released, unstratified, as Tier 2 (Appendix A.4). A representative, balanced 768-record evaluation subset is sampled from Tier 1 to run compute-intensive evaluations on fewer records.

Overall, AUB-Wiki contains 2,158 Tier-1 records, 8,223 clarifications, and 1,425 matched controls, with a 768-record evaluation subset. Further extensive details on the dataset construction process are included in Appendix A.

3.3 Dataset validation

The dataset is built around two measured axes: interpretation coverage (the knowledge axis) and interpretation count (the ambiguity axis). Before using these axes to evaluate uncertainty methods, we test whether they exhibit the intended behavioral properties in models. In brief, we test whether (i) coverage predicts clarification answer correctness and remains discriminative even for frontier models, (ii) interpretation count is largely independent of model knowledge, and (iii) interpretation count is reflected in model-generated interpretations across different models. The implementation of all three checks is described below. Together, these validity checks establish the behavioral properties required to use the two axes in the subsequent uncertainty experiments.

Notation. A record $r \in \mathcal{R}$ comprises an ambiguous question q_r, a set of interpretations $\mathcal{I}_r$ of size $|\mathcal{I}_r|$, a clarification $q_{r,i}$ with coverage $c_{r,i}$ for each interpretation i, a median coverage $c_r = \text{median}(\{c_{r,i} \mid i \in \mathcal{I}_r\})$, and, where available, a matched control q'_r. For a given model $\mathcal{M}$, the clarification answer correctness $C_{\mathcal{M}}(q_{r,i}) \in \{0, 1\}$ indicates whether the answer to $q_{r,i}$ generated by the model matches the ground-truth answer. We also ask the model to generate its own set of interpretations $\hat{\mathcal{I}}_r^{\mathcal{M}}$ for the ambiguous question q_r. The number of ground-truth interpretations recovered is $\left|\hat{\mathcal{I}}_r^{\mathcal{M}} \cap \mathcal{I}_r\right|$. All correlations are Spearman's ρ.

Validity checks. We use three complementary checks, evaluated separately for each model.

1. The *coverage check*
$$V_1 = \text{AUROC}\left(\{(c_{r,i}, C_{\mathcal{M}}(q_{r,i})) \mid i \in \mathcal{I}_r, r \in \mathcal{R}\}\right)$$
asks whether coverage predicts clarification answer correctness. A high V_1 indicates that coverage is a good predictor of answer correctness. Since correctness is measured on clarified questions, which by design carry no ambiguity, $C_{\mathcal{M}}(q_{r,i})$ provides a direct epistemic readout of whether the model knows the answer, as all aleatoric uncertainty has been eliminated. This check therefore validates whether coverage is a good signal of model knowledge. As an additional check, we compute the model's accuracy on clarified questions at different coverage levels; higher coverage levels should yield higher accuracy.
2. The *count check* asks whether interpretation count (our ground truth for aleatoric uncertainty and ambiguity) is also associated with the model's epistemic knowledge. This check thus validates the independence of the proposed knowledge and ambiguity axes. For each record, we compute the average model correctness across its clarifications. Because templates differ both in this average correctness and in their typical interpretation count, we compute the correlation within each template t and average over the set of templates $\mathcal{T}$ containing at least thirty released records, weighted by their record counts n_t:
$$V_2 = \sum_{t \in \mathcal{T}} \frac{n_t}{\sum_{t' \in \mathcal{T}} n_{t'}} \rho_t \left(|\mathcal{I}_r|, \frac{1}{|\mathcal{I}_r|} \sum_{i \in \mathcal{I}_r} C_{\mathcal{M}}(q_{r,i}) \right). \quad (1)$$
Thus, $V_2 \approx 0$ indicates that average correctness across clarifications does not systematically vary with interpretation count within templates.
3. The *enumeration check*
$$V_3 = \rho\left(|\mathcal{I}_r|, \left|\hat{\mathcal{I}}_r^{\mathcal{M}} \cap \mathcal{I}_r\right|\right)$$
asks whether interpretation count corresponds to intrinsic model beliefs. We measure this as the correlation between the number of ground-truth interpretations identified by the model (via a direct query) and the total ground-truth interpretation count.

4 Experiments

In this section, we evaluate the validity checks introduced in Section 3.3 and benchmark uncertainty quantification methods on AUB-wiki.

4.1 SETUP

Our models include Qwen3.5-4B, Qwen3.5-9B, Qwen3.5-27B, Gemma 4 12B, and Phi-4 14B locally, and Claude Opus 5 via API. We measure answer correctness and evaluate some of the uncertainty methods introduced in Section 2: $P(\text{True})$ (Kadavath et al., 2022), Semantic Entropy (SE) (Kuhn et al., 2023), Spectral Decomposition (SD) (Walha et al., 2026), and Input Clarification Ensembling (ICE) (Hou et al., 2024) on AUB-Wiki. We also add an LLM-based ambiguity probe, where we ask a model whether the given query is ambiguous, and measure the probability $P(\text{Yes})$ of the token *Yes* in the output. All correlations below are reported as Spearman's ρ. Additional implementation details are provided in Appendix C.

4.2 DO THE TWO AXES HOLD FOR THE MODELS?

We first check the three validity conditions from Section 3.3 for each model. Table 1 summarizes the results.

Coverage predicts knowledge. Across all models, clarification-level accuracy (i.e., accuracy across all unambiguous clarifications of all questions) increases monotonically with the coverage of the corresponding interpretation. Coverage alone predicts answer correctness (across all clarifications) with an AUROC of 0.75 to 0.81 for the open models. Accuracy levels shift upward with model size, while the AUROC decreases as larger models answer more low-coverage interpretations correctly. Opus 5 correctly answers 58% of the least-covered clarifications, compared with 5 to 26% for the open models, and has the lowest coverage AUROC at 0.65.

Interpretation count is only weakly associated with record average correctness. Interpretation count shows only a weak association with average correctness across clarifications. The weighted macro-average of $\rho(\text{count}, \text{record acc.})$ over templates ranges from -0.10 to $+0.06$ across the six models. We also evaluate the correlation between the two proposed axes (interpretation count vs. coverage). The resulting rank correlation, shown in Figure 4 (Appendix B), is small (-0.12) when computed over both records and controls. That means, our dataset does not induce a correlation between aleatoric and epistemic uncertainty and thus provides a suitable benchmark for assessing uncertainty disentanglement (section 4.3).

All models are able to track interpretation count. Models recover more ground-truth interpretations as the ground-truth interpretation count increases. Across the open models, $\rho(\text{count}, \text{identified})$ ranges from $+0.43$ to $+0.60$, and it reaches $+0.85$ for Opus 5. Thus, the ground-truth interpretation count is correlated with the number of valid interpretations that models recover from the question, and this correlation is stronger for more capable models.

Table 1: Validation of the two benchmark axes across models. We report per-clarification accuracy by coverage, the AUROC of coverage for predicting correctness, the within-template correlation $\rho(\text{count}, \text{record acc.})$, and the enumeration correlation $\rho(\text{count}, \text{identified})$. Opus 5 is evaluated on the 768-record subset. Bootstrap intervals are reported in Appendix D.

model	**accuracy by coverage**				**coverage**	ρ(**count**, **record acc.**)	ρ(**count**, **identified**)
	<20	**20–29**	**30–59**	**≥60**	**AUROC**		
Qwen3.5-4B	0.05	0.12	0.31	0.72	0.81	+0.06	+0.43
Qwen3.5-9B	0.11	0.26	0.51	0.83	0.78	−0.04	+0.44
Qwen3.5-27B	0.26	0.45	0.70	0.94	0.75	−0.09	+0.59
Gemma 4 12B	0.05	0.10	0.24	0.52	0.75	+0.02	+0.45
Phi-4 14B	0.20	0.38	0.62	0.84	0.75	−0.03	+0.60
Opus 5	0.58	0.67	0.78	0.94	0.65	−0.10	+0.85

4.3 WHAT DO UNCERTAINTY METHODS MEASURE?

We next evaluate (i) how uncertainty quantification methods identify ambiguity when controlling for knowledge, and (ii) whether uncertainty decomposition can disentangle the aleatoric and epistemic sources of uncertainty along the intended axes.

How do uncertainty quantification methods quantify ambiguity? To address this question, we consider the 215 records for which every interpretation has at least 30 sitelinks as a high-coverage

subset with little epistemic uncertainty and measure the correlation between estimated uncertainty and interpretation count (as proxy for ambiguity); we thus ask wether the uncertainty estimated by total UQ methods is able to capture the dominant ambiguity component when controlling for epistemic uncertainty.

Table 2 shows that total uncertainty scores do not correlate with interpretation count (i.e. severity of ambiguity): $\rho(SE, \text{count})$ is negative for every model, ranging from -0.28 to -0.21. Likewise, $1 - P(\text{True})$ shows no consistent relationship, with $\rho(1 - P(\text{True}), \text{count})$ ranging from -0.11 to $+0.15$. In contrast, the ambiguity probe raises monotonically with count, with $\rho(\text{probe}, \text{count}) = +0.31$ to $+0.45$ across models, indicating that models do have intrinsic knowledge on the severity of ambiguity that is not captured by total uncertainty quantification methods.

Table 2: Spearman correlations of total uncertainty scores with the interpretation count on the high-coverage stratum ($n = 215$ records per model), where every interpretation has at least 30 sitelinks. Bootstrap 95% intervals are shown below each estimate. SE and $1 - P(\text{True})$ are computed on the ambiguous question.

score	Qwen 4B	Qwen 9B	Qwen 27B	Gemma 12B	Phi-4 14B
SE	-0.21 [-0.35, -0.07]	-0.26 [-0.38, -0.13]	-0.21 [-0.33, -0.08]	-0.28 [-0.40, -0.16]	-0.21 [-0.33, -0.08]
$1 - P(\text{True})$	-0.11 [-0.26, +0.03]	-0.09 [-0.23, +0.04]	$+0.15$ [+0.01, +0.29]	$+0.03$ [-0.11, +0.16]	-0.10 [-0.24, +0.03]
probe $P(\text{Yes})$	$+0.36$ [+0.23, +0.47]	$+0.31$ [+0.18, +0.43]	$+0.41$ [+0.30, +0.52]	$+0.45$ [+0.33, +0.54]	$+0.37$ [+0.25, +0.48]

We additionally ask whether uncertainty estimates can distinguish ambiguous records from single-interpretation controls under the same high-coverage condition. Table 3 shows that uncertainty estimates provide little separation: both SE and $1 - P(\text{True})$ remain at or below chance with AUROC ranging from 0.41 to 0.51 across all models, except for Qwen3.5-27B (0.59 AUROC for $1 - P(\text{True})$). However, the ambiguity probe separates the two groups more consistently, with an AUROC of 0.59–0.70 , and its discrimination improves with ambiguity: AUROC rises from 0.54–0.64 for records with two interpretations to 0.72–0.84 for those with three or more. The same qualitative pattern holds on the full release (Appendix E.2). Together with the positive correlation between the probe and interpretation count, these results show that ambiguity is recoverable through direct elicitation, but is not reliably expressed in SE or $1 - P(\text{True})$ computed from the answers.

Table 3: On the high-coverage stratum; AUROC of each score for separating the ambiguous records (215) from the single-interpretation controls (473), and the probe by interpretation count. Bootstrap intervals are reported in Appendix E.2.

	answer-based		ambiguity probe		
model	SE	$1 - P(\text{True})$	all	2	$\geq$3
Qwen3.5-4B	0.46	0.46	0.62	0.56	0.75
Qwen3.5-9B	0.41	0.44	0.59	0.54	0.72
Qwen3.5-27B	0.43	0.59	0.70	0.64	0.84
Gemma 4 12B	0.45	0.48	0.64	0.56	0.83
Phi-4 14B	0.50	0.51	0.61	0.55	0.76

Do uncertainty decompositions separate ambiguity from knowledge? Uncertainty decomposition methods such as spectral uncertainty decomposition (SD) and input clarification ensembling (ICE) (Walha et al., 2026; Hou et al., 2024) aim to decompose total uncertainty into epistemic and aleatoric components by first generating clarifications using a frontier model and then measuring uncertainty within and across them. Table 4 shows that the aleatoric terms discriminate strongly between ambiguous records and coverage-matched controls: using the GPT-5.6 Sol clarifier, SD and ICE achieve AUROCs of 0.88–0.90. However, much of this discriminative signal is driven by the clarifier itself and the number of generated clarifications alone achieves an AUROC of 0.90 and correlates with interpretation count at $+0.79$.

Crucially, ambiguity detection alone does not establish that the decomposition separates the two sources of uncertainty. To assess disentanglement, we test each component against both interpretation count (as aleatoric ground truth) and the record's median coverage (epistemic ground truth) and thereby measure its association with its intended axis and its sensitivity to the other. Across models and both decomposition methods, the epistemic term decreases with interpretation coverage, with $\rho(\text{epistemic}, \text{coverage}) = -0.23$ to -0.37, while its correlation with interpretation count remains weak, at -0.04 to -0.13. Conversely, the aleatoric term increases with interpretation count, with $\rho(\text{aleatoric}, \text{count}) = +0.58$ to $+0.73$, while remaining nearly unrelated to coverage, at -0.06 to $+0.01$ (Table 4). Figure 1 further illustrates how the decomposition terms in SD align with the intended sources: epistemic uncertainty varies primarily with knowledge, whereas aleatoric uncertainty varies primarily with ambiguity, suggesting that the decomposition behaves as intended.

Table 4: Clarification-based decompositions on the 768-record subset for the spectral decomposition (SD) and ICE. Left: AUROC of the aleatoric term for separating ambiguous records from their 418 coverage-matched controls. Right: Spearman ρ of the epistemic term (mean entropy within a clarification) and the aleatoric term (disagreement across clarifications) with a record's median interpretation coverage and with its interpretation count. The last row scores the clarifier's number of clarifications alone, with no model in the loop. Bootstrap intervals are reported in Appendix E.3.

	AUROC	**epistemic (within clar.)**		**aleatoric (across clar.)**	
model	**amb. vs. control**	**ρ(coverage)**	**ρ(count)**	**ρ(count)**	**ρ(coverage)**
	SD / ICE	SD / ICE	SD / ICE	SD / ICE	SD / ICE
Qwen3.5-4B	0.89 / 0.89	-0.36 / -0.28	-0.10 / -0.12	$+0.70$ / $+0.71$	-0.02 / 0.00
Qwen3.5-9B	0.89 / 0.89	-0.34 / -0.29	-0.05 / -0.07	$+0.70$ / $+0.73$	-0.02 / -0.03
Qwen3.5-27B	0.90 / 0.89	-0.36 / -0.33	-0.04 / -0.05	$+0.71$ / $+0.73$	-0.03 / -0.03
Gemma 4 12B	0.88 / 0.89	-0.37 / -0.29	-0.11 / -0.13	$+0.59$ / $+0.71$	-0.00 / $+0.01$
Phi-4 14B	0.88 / 0.88	-0.32 / -0.23	-0.07 / -0.09	$+0.58$ / $+0.73$	-0.05 / -0.06
clarification count alone	0.90			$+0.79$	

5 Discussion and Conclusion

Uncertainty in language models can arise from limited model knowledge or from ambiguity in the input, yet evaluating methods that claim to separate these sources require a benchmark dataset where both can be measured in a controlled manner. While ambiguity datasets exist (Min et al., 2020; Ji et al., 2025; Hou et al., 2024), benchmark datasets that allow epistemic and aleatoric uncertainty to be evaluated jointly are still missing.

Here, we introduce AUB-wiki to allow for the principled and fact-grounded evaluation of uncertainty quantification and disentanglement methods: a Wikidata-based, LLM-free benchmark with reproducible interpretation sets and a scalable construction pipeline, using interpretation coverage as a proxy for knowledge and interpretation count as an ambiguity axis. This design allows uncertainty methods to be evaluated against the source they are intended to capture rather than against uncertainty as a single undifferentiated signal. Our experiments show that commonly used total uncertainty measures do not consistently reflect ambiguity, whereas uncertainty decomposition methods ICE and SD exhibit the intended separation: epistemic uncertainty aligns primarily with coverage, and aleatoric uncertainty with interpretation count. This disentanglement is in stark contrast to image classification, where commonly used uncertainty decompositions resulted in strongly correlated components despite their different intended meanings Mucsányi et al. (2024).

AUB-Wiki focuses on referential ambiguity in English, template-generated factual questions. Generalization to naturally occurring questions, other domains and languages remains to be established. Interpretation sets are complete relative to the specified membership rules and Wikidata version, which may exclude valid real-world interpretations outside this scope. Finally, the reference measurements also have limitations. Coverage predicts clarification answer accuracy, but measures how widely an entity is documented rather than directly observing a model's knowledge of a particular fact.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing, literature retrieval, and coding. For writing, they were used to rephrase and polish author-written text for clarity and academic style. For literature retrieval, they were used to identify candidate references, all of which were independently read and verified by the authors before citation. For coding, they assisted with code refinement, comments, debugging, and checks of coding practices; all AI-assisted code was reviewed and tested by the authors.

We did not use generative AI to generate synthetic data, formulate or prove mathematical claims, design the research methodology or experiments, or make decisions about benchmark construction and evaluation. No model-generated content is part of the benchmark or reported results beyond the outputs of the models explicitly evaluated in the experiments.

The authors reviewed all AI-assisted work and take responsibility for the final content of the paper, including its text, claims, code, and artifacts.

REPRODUCIBILITY STATEMENT

Benchmark construction is described in Section 3.2 and Appendix A, including template screening, membership rules, completeness audits, release selection, and the evaluation subset. Experimental settings, prompts, evaluation protocols, validity checks, and bootstrap procedures are documented in Section 4.1, Section 3.3, and Appendices C–E.

The supplementary material contains both benchmark tiers, the evaluation subset, datasheet, generation and evaluation code, templates and rules, and scripts to reproduce all tables and figures. The release draw is deterministic given the recorded seed, and regeneration from Wikidata is reproducible relative to the recorded data version.

REFERENCES

Goshi Aoki. Large language models in politics and democracy: A comprehensive survey. *arXiv preprint arXiv:2412.04498*, 2024.

Hannah Bast and Björn Buchhold. Qlever: A query engine for efficient sparql+ text search. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 647–656, 2017.

Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pp. 1184–1193. PMLR, 2018.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? In *International Conference on Learning Representations*, volume 2024, pp. 7735–7790, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Sebastian Gregor Gruber and Florian Buettner. A bias-variance-covariance decomposition of kernel scores for generative models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16460–16501, 2024.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=XPZIaotutsD`.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 19023–19042. PMLR, 2024.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Jiabao Ji, Min Li, Priyanshu Kumar, Shiyu Chang, and Saloni Potdar. DEEPAMBIGQA: Ambiguous multi-hop questions for benchmarking LLM answer completeness, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty quantification in natural language. In *Proceedings of the 2023 International Conference on Learning Representations*, 2023.

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

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Dongryeol Lee, Segwang Kim, Minwoo Lee, Hwanhee Lee, Joonsuk Park, Sang-Woo Lee, and Kyomin Jung. Asking clarification questions to handle ambiguity in open-domain qa. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11526–11544, 2023.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. Datasets for large language models: A comprehensive survey. *Artificial Intelligence Review*, 58(12):403, 2025.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 5783–5797. Association for Computational Linguistics, 2020.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Advances in neural information processing systems*, 37:50972–51038, 2024.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H. S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24384–24394, 2023.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929, 2024.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. Asqa: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8273–8288, 2022.

Tim Tomov, Dominik Fuchsgruber, Tom Wollschläger, and Stephan Günnemann. The illusion of certainty: Uncertainty quantification for LLMs fails under ambiguity, 2025.

John Von Neumann. *Mathematical foundations of quantum mechanics: New edition*. Princeton university press, 2018.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Nassim Walha, Sebastian G Gruber, Thomas Decker, Yinchong Yang, Alireza Javanmardi, Eyke Hüllermeier, and Florian Buettner. Fine-grained uncertainty decomposition in large language models: A spectral approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26090–26098, 2026.

Wikimedia Foundation. MediaWiki action API. `https://en.wikipedia.org/w/api.php`, 2026a. English Wikipedia article categories, accessed September 2026.

Wikimedia Foundation. Pageviews analysis. `https://pageviews.wmcloud.org/`, 2026b. Version 2026.09.15. Accessed: 2026-09-26.

Wikimedia Foundation. XTools: Page links. `https://xtools.wmcloud.org/`, 2026c. Number of English Wikipedia articles linking to an article, accessed September 2026.

Yongjin Yang, Haneul Yoo, and Hwaran Lee. Maqa: Evaluating uncertainty quantification in llms regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5861–5878, 2025.

A Benchmark construction

The pipeline operates on Wikidata and Wikipedia (Vrandečić & Krötzsch, 2014; Wikimedia Foundation, 2026a), uses no language model, and is deterministic.

A.1 Template screening and generation

We first identify class–property combinations that can support reliable question templates. Starting from 30 seed classes across 15 topics, we retrieve Wikidata members with at least 20 sitelinks and measure, for each property, its coverage, number of values per member, and the number of groups of 2–15 eligible members that it induces.

A grouping property must be item-valued, occur for at least 90% of the class, average at most 1.1 values per member, induce at least 20 usable groups, and be definitional rather than incidental, i.e., a relation Wikidata records whenever it records the entity, such as a film's director, and not one recorded only sometimes and open-endedly, such as ownership, awards, genre or cast membership, on which a group could be silently incomplete. An answer property must be time-, quantity-, or item-valued and provide a sufficiently complete and approximately single-valued answer. The 20-sitelink screening threshold is stricter than the threshold later used to construct the pool, making this stage conservative in which classes it admits.

The surviving class–property combinations are encoded as templates specifying the entity class, grouping property, answer property, question and clarification wording, and answer type. Fourteen templates were defined, of which eight remain in the released benchmark after the completeness audits and the pruning (Appendix A.3, Table 6). Expanding a template against Wikidata produces one group per grouping-property value. Groups with 2–15 eligible members become ambiguous-record candidates, groups with a single member become control candidates, and groups with more than 15 members are excluded rather than truncated.

A.2 MEMBERSHIP RULES

The membership rule is the criterion that decides which entities are the interpretations of a question. We write one rule per class by hand, following the typing conventions that Wikidata's editor projects publish for that kind of item. The rule is compiled into the generating query, so that the interpretation set of a record can be re-derived from the rule and a Wikidata version rather than resting on anyone's recall. The rule has two parts.

Which items belong to the class. Because Wikidata does not encode entity types uniformly, each rule specifies the admissible routes into a class. These may include direct *instance of* statements, selected subclasses, or a broader class combined with a secondary annotation. For example, novels may be represented as literary works whose form is *novel*; studio albums and operas use analogous conventions.

Subclass links are not followed when they would introduce unrelated entities, and we explicitly exclude items that should not form separate interpretations, such as editions, book series, or individual copies. Items that remain untyped in Wikidata are checked against corresponding English Wikipedia category information through the Wikimedia API (Wikimedia Foundation, 2026a) and groups containing unresolved items are excluded from the pool.

Which members count as distinct interpretations. A member must have at least one sitelink, and an English label, since the question is asked in English. Entities that are derived from, or duplicate, another member of the same group are removed using Wikidata relations such as *edition of*, *based on* and *part of the series*

Two conditions discard the whole record: (i) if two members share a label, as when a director has filmed the same title twice, their clarifications would be identical and could not name one thing each, (ii) if two groups share a label, the question is ambiguous in a way the interpretation count does not record. Labels are compared after normalization.

A.3 AUDITING INTERPRETATION SETS

The generated interpretation sets are audited for missing members before release selection. For Film, Music, and Literature templates, we compare the generated sets with independently maintained English Wikipedia categories corresponding to the grouping relation. For People templates, where equivalent categories are unavailable, we use complementary Wikidata queries that inspect the relation from the opposite direction and include statements of all ranks, followed by a manual check of a sample against the relevant Wikipedia pages.

Table 5 summarizes the category-based audit. The retained templates are complete for 97–100% of audited records, and incomplete records are removed from the pool. Two templates are dropped entirely because of substantial incompleteness.

Table 5: Completeness audit against English Wikipedia categories. A record is incomplete if the category lists an interpretation that Wikidata does not link to the group.

template	audited	complete	incomplete
film by director, release year	369	359	10
film by director, duration	269	261	8
studio album by performer	85	80	5
opera by composer, premiere year	11	11	0
novel by author, publication year	7	7	0
footballer's club (dropped)	414	225	189
film by composer (dropped)	33	18	15

Completeness is therefore claimed under the membership rule and relative to the Wikidata version used for construction. The design constraint of Section 3.1 is what makes the claim meaningful: because the grouping property is definitional and recorded whenever the entity is, a missing member

is a gap in Wikidata rather than a gap in the rule, and an entity without a Wikidata item is outside this universe rather than an omitted interpretation.

A.4 Release selection and controls

From the audited pool of records, we construct Tier 1 by drawing deterministically across strata of interpretation count (2, 3, 4–5, and 6–15) and median interpretation coverage, while balancing topics and templates and allowing each entity, whether the group or an interpretation, to appear in only one record. Each selected ambiguous record is paired with a single-interpretation control from the same template whose coverage is closest to the record's median interpretation coverage. This preserves the question form and approximately matches the knowledge proxy while reducing the interpretation count to one. Controls fall in the same coverage bin as their paired, with a mean coverage difference of 1.7 sitelinks.

The initial draw contained 2,959 records. Post-draw checks removed records that cannot be evaluated reliably, including cases with non-exact answers, shared clarification answers, ambiguous group names, or missing members identified by the People-specific audits (Table 6). After these exclusions, Tier 1 contains 2,158 records and 1,425 matched controls.

Table 6: Records and controls removed after the draw, by reason. The draw holds 2,959 records; the release 2,158.

reason	records	controls
All interpretations share one gold answer	101	—
Answer is not exact: unknown value (99 entities), circa (795), century (818), decade (150), millennium (44)	484	258
of which child / sibling / spouse of a person	246 / 144 / 94	154 / 62 / 41
Group's name also names another entity of the same type	201	116
People record missing a member the reverse or all-ranks query returns	13	2
Templates left with one record	2	0

Tier 2 contains the initial audited pool of 6,734 records and 5,792 single-interpretation controls, including all Tier-1 records under the same identifiers. It follows the same membership rules and completeness audits, but unlike Tier 1 it is not stratified, allows entities to recur across records, and provides controls as an unpaired pool. It also retains 1,256 records and 672 controls with inexact answers, which are flagged per interpretation and excluded from Tier 1. Both tiers are complete under the benchmark membership rules, and every interpretation is associated with an answer.

For the main experiments, greedy answers, $P(\mathrm{True})$, the ambiguity probe, and semantic entropy are computed on the full Tier-1 release. The clarification-based decompositions require repeated sampling for every clarification and are therefore evaluated on a fixed subset of 768 Tier-1 records, containing 3,358 clarifications and 418 matched controls. This *evaluation set* subset follows the same count–coverage cell structure as Tier 1, with balanced sampling across topics and templates (Table 8, right). Opus 5 is evaluated on the same subset.

Datasheet. The release ships with a datasheet describing every field and listing, for each tier, the records per count–coverage cell and per template, the Tier-1 controls per template, the draw command, and the Wikidata data version of each topic.

A.5 Coverage proxies

Sitelinks are one of several ways to measure how widely an entity is documented. To check that the choice does not drive the results, we fetched two alternatives for the interpretation and control entities of the release: English Wikipedia page views over twelve months (Wikimedia Foundation, 2026b), and the number of Wikipedia articles linking to the entity's article (Wikimedia Foundation, 2026c). The three measures agree only moderately with each other: sitelinks correlate with incoming links at $\rho = 0.69$ and with page views at 0.57.

What matters is how well each predicts whether a model knows the fact. We measure knowledge here by the teacher-forced probability of the ground-truth answers on clarifications, that is, the probability the model assigns to the gold value when it is appended to the answer prompt. On this measure the three proxies are close. Across the five open models, sitelinks correlate with it at $\rho = 0.30$, 0.41, 0.49, 0.18 and 0.42; page views at 0.37, 0.42, 0.41, 0.31 and 0.37; incoming links at 0.34, 0.43, 0.49, 0.23 and 0.43.

We keep sitelinks for three practical reasons: they come from the same Wikidata query as the records, they exist for every entity (some entities have no English article and hence no page views or incoming links), and they are not tied to a time window. The two alternative values are not part of the release.

B Dataset statistics

Table 7 gives the composition of the Tier-1 release by template, and Table 8 its distribution over the twelve cells of the draw. Controls exist for 66% of the records, so every comparison between ambiguous questions and controls is made on that subset.

Table 7: Composition of the Tier-1 release: records and matched controls per template.

topic	template (question)	records	controls
People	birth year of the child of {person}	534	399
People	birth year of the sibling of {person}	432	363
People	birth year of the person {person} was married to	409	359
Film	release year of the film directed by {director}	394	136
Film	duration of the film directed by {director}	295	129
Music	release year of the studio album by {performer}	76	30
Music	premiere year of the opera composed by {composer}	10	6
Literature	publication year of the novel written by {author}	8	3
	total	2,158	1,425

Table 8: Records per cell of the draw, by interpretation count and median interpretation coverage (sitelinks), for the Tier-1 release and for the 768-record evaluation subset.

	Tier 1 (2,158 records)				evaluation subset (768)			
interpretations	<20	20–29	≥ 30	total	<20	20–29	≥ 30	total
2	209	348	365	922	65	66	59	190
3	147	163	111	421	69	68	60	197
4–5	171	123	115	409	69	63	64	196
6–15	191	139	76	406	61	69	55	185

The two axes across the release. Figure 4 shows the coverage of every question of the Tier-1 release at each interpretation count, with the controls as the single-interpretation column. The median coverage stays between 17 and 27 sitelinks from one interpretation to ten and more, and the quartile ranges overlap at every count. Count and coverage are weakly related, at $\rho = -0.29$ over the records and -0.12 once the controls are included. The residual relation comes from the pool rule, which requires two interpretations with at least 15 sitelinks and therefore pins the median of two- and three-interpretation records at 15 or above; from four interpretations upward the correlation is -0.09.

C Experimental setup

Models and decoding. The five open models are run locally as their instruct variants with reasoning switched off, in bfloat16, with one prompt template per score (Appendix F). The scored answer is decoded greedily with at most 48 new tokens; Opus 5 is queried through an API at temperature 0. Semantic entropy uses ten samples at temperature 1, clustered by bidirectional entailment using DeBERTa-large-MNLI (He et al., 2021). Following Walha et al. (2026), we use ten samples per

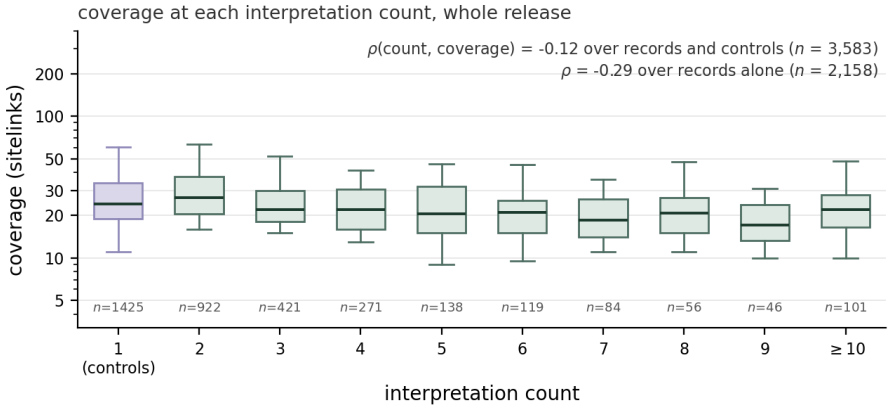


Figure 4: Coverage of the Tier-1 questions by interpretation count. A record's coverage is the median coverage of its interpretations; a control's is that of its single interpretation (count 1). Boxes mark the median and quartiles, whiskers the 5th and 95th percentiles; records with ten or more interpretations share the last box.

clarification at temperature 0.5 for SD and ICE. $P(\mathrm{True})$ shows the question and the greedy answer and reads the probability of the first token from the logits; the ambiguity probe reads $P(\mathrm{Yes})$ in the same way, and the enumeration that follows it is decoded greedily with at most 256 new tokens. Both SD and ICE decompose uncertainty by first inferring a set of clarifications for a given question using a frontier clarification model; since the frontier models used in the primary publications at the time of writing are outdated, we use the more recent GPT-5.6 Sol with reasoning disabled, temperature 0, and a fixed seed.

Evaluation data. Answer correctness, SE and $P(\mathrm{True})$ are evaluated on the Tier-1 release of 2,158 ambiguous records, together with 1,425 controls and 8,223 clarifications. Because of their higher computational cost, SD and ICE are only evaluated on the fixed 768-record evaluation subset with 418 controls. Opus 5 is also evaluated on this subset for the validation experiments reported in Table 1.

Decompositions. The spectral decomposition draws ten answers per clarification at temperature 0.5, embeds them with all-mpnet-base-v2 and takes the von Neumann entropy of the RBF kernel ($\gamma = 1$) of the embeddings; the epistemic term is the mean entropy within a clarification and the aleatoric term the entropy of the pooled answers minus that mean. ICE uses the same sampled answers, clustered by entailment using DeBERTa-v2-xlarge-MNLI (He et al., 2021), and the Shannon entropy of the cluster shares, decomposed in the same way. The number of clarifications used by the last row of Table 4 is the number the clarifier wrote for the question, with a single clarification for a question it judged unambiguous.

Correctness. An answer is scored by a deterministic string-matching-based rule for its answer type. A year is correct if the ground-truth year appears among the years in the answer, so that "8 June 1949" matches "1949" while "15th century" does not; BC dates are handled. A number is correct if the first number in the answer is within a tolerance of the gold value. For an ambiguous question, an answer counts as correct if it matches the answer of any interpretation; for a clarification or a control, it must match the single gold answer.

Statistics. All correlations are Spearman's ρ and all intervals are 95% bootstrap intervals over records, with 1,000 resamples.

D VALIDITY CHECKS

Table 9 gives the bootstrap intervals of the two release-wide quantities of Table 1. Both are narrow, since the coverage AUROC is computed over 8,223 clarifications and the enumeration correlation over 2,158 records (768 for Opus 5). The count check is a weighted mean of within-template correlations and is reported per template in Table 10.

Table 9: 95% bootstrap intervals for the AUROC of coverage for predicting a correct clarification answer and for the enumeration correlation $\rho(\text{named}, \text{count})$ of Table 1.

model	AUROC of coverage	ρ(named, count)
Qwen3.5-4B	0.81 [0.79, 0.82]	$+0.43$ [+0.38, +0.46]
Qwen3.5-9B	0.78 [0.77, 0.79]	$+0.44$ [+0.41, +0.48]
Qwen3.5-27B	0.75 [0.74, 0.76]	$+0.59$ [+0.56, +0.62]
Gemma 4 12B	0.75 [0.74, 0.77]	$+0.45$ [+0.41, +0.48]
Phi-4 14B	0.75 [0.73, 0.76]	$+0.60$ [+0.56, +0.63]
Opus 5	0.65 [0.62, 0.66]	$+0.85$ [+0.81, +0.89]

The count check per template. The check is restricted to templates with at least 30 released records, which leaves the six of Table 10 and excludes the opera (10 records) and novel (8) templates. In the five templates with more than 200 records the correlation lies between -0.17 and $+0.18$ for the open models and its sign is not consistent across models. The exception is the album template, the smallest kept, where accuracy falls with the count for all six models (-0.13 to -0.21); with 76 records it carries about 4% of the weight of the pooled figure.

Table 10: The count check per template and model: ρ, within template, between the number of interpretations of a record and the share of its clarifications the model answers correctly. The last row is the record-count weighted mean reported in Table 1. Opus 5 is scored on the 768-record subset, where the same templates hold 56 to 249 records.

template	n	Qwen 4B	Qwen 9B	Qwen 27B	Gemma 12B	Phi-4 14B	Opus 5
child of a person, birth year	534	$+0.01$	-0.06	-0.17	$+0.02$	-0.07	-0.23
sibling of a person, birth year	432	$+0.06$	$+0.01$	-0.07	$+0.06$	$+0.03$	-0.26
spouse of a person, birth year	409	$+0.06$	-0.01	-0.08	$+0.04$	-0.03	-0.06
film by director, release year	394	$+0.05$	-0.12	-0.10	-0.01	-0.09	$+0.00$
film by director, duration	295	$+0.18$	$+0.03$	$+0.02$	$+0.01$	$+0.11$	-0.09
album by performer, release year	76	-0.15	-0.21	-0.20	-0.16	-0.21	-0.13
weighted mean		$+0.06$	-0.04	-0.09	$+0.02$	-0.03	-0.10

E FURTHER EXPERIMENTS

E.1 ROBUSTNESS OF THE HIGH-COVERAGE STRATUM

The stratum holds knowledge fixed by coverage. An alternative holds it fixed by outcome: keeping, per model, the records whose every clarification it answers correctly. This excludes the effect of lack of knowledge as it means the models know the answer to all clarifications in the record. Table 11 shows the same picture as Table 2: semantic entropy falls with the count, $1 - P(\text{True})$ has no consistent sign, and the probe rises with the count for the three models with more than a hundred such records; for Qwen 4B and Gemma (50 and 29 records) its interval includes zero.

E.2 DETECTING AMBIGUITY: INTERVALS AND THE FULL RELEASE

Table 12 gives the bootstrap intervals of the AUROCs of Table 3. They are wider than the release-wide intervals of the paper, being computed on 215 records against 473 controls, but every conclusion drawn from the point estimates survives them: the intervals of semantic entropy lie at or below 0.5

Table 11: Records whose every clarification the model answers correctly: ρ of each score on the ambiguous question with the number of interpretations, with 95% bootstrap intervals.

model	n	ρ(count, SE)	ρ(count, $1 - P(\text{True})$)	ρ(count, probe)
Qwen3.5-4B	50	-0.25 [-0.46, -0.03]	-0.44 [-0.63, -0.25]	$+0.08$ [-0.18, +0.31]
Qwen3.5-9B	153	-0.29 [-0.42, -0.15]	-0.04 [-0.22, +0.14]	$+0.35$ [+0.20, +0.48]
Qwen3.5-27B	396	-0.18 [-0.28, -0.08]	$+0.19$ [+0.09, +0.28]	$+0.42$ [+0.34, +0.50]
Gemma 4 12B	29	-0.07 [-0.46, +0.30]	$+0.06$ [-0.31, +0.43]	$+0.24$ [-0.14, +0.54]
Phi-4 14B	263	-0.17 [-0.28, -0.04]	-0.27 [-0.38, -0.16]	$+0.49$ [+0.39, +0.58]

for every model, those of $1 - P(\text{True})$ include 0.5 for every model but Qwen3.5-27B, and those of the probe lie above 0.5 throughout.

Table 12: 95% bootstrap intervals for the AUROCs of Table 3: each score on the high-coverage stratum, and the probe by interpretation count.

model	SE	$1 - P(\text{True})$	probe, all	probe, 2	probe, $\geq$3
Qwen3.5-4B	0.46 [0.41, 0.50]	0.46 [0.41, 0.50]	0.62 [0.57, 0.66]	0.56 [0.51, 0.62]	0.75 [0.69, 0.81]
Qwen3.5-9B	0.41 [0.36, 0.46]	0.44 [0.40, 0.49]	0.59 [0.54, 0.63]	0.54 [0.48, 0.59]	0.72 [0.64, 0.79]
Qwen3.5-27B	0.43 [0.39, 0.48]	0.59 [0.55, 0.64]	0.70 [0.66, 0.74]	0.64 [0.59, 0.69]	0.84 [0.79, 0.90]
Gemma 4 12B	0.45 [0.40, 0.50]	0.48 [0.44, 0.53]	0.64 [0.59, 0.68]	0.56 [0.50, 0.61]	0.83 [0.77, 0.88]
Phi-4 14B	0.50 [0.45, 0.55]	0.51 [0.46, 0.55]	0.61 [0.57, 0.66]	0.55 [0.50, 0.60]	0.76 [0.70, 0.82]

Table 13 repeats the same detectors on the whole release, 2,158 ambiguous records against their 1,425 coverage-matched controls, with the probe also by interpretation count. The ordering is the one of the stratum. Semantic entropy and $1 - P(\text{True})$ sit at or below chance; the probe separates the two groups at 0.61 to 0.74, is near chance for records with two interpretations and reaches 0.75 to 0.92 for five or more, and its separation grows with model size within the Qwen family.

Table 13: AUROC of each score for separating ambiguous records from coverage-matched single-interpretation controls on the full release (2,158 against 1,425), and the probe by interpretation count, with 95% bootstrap intervals.

	answer-based		ambiguity probe, by count			
model	SE	$1 - P(\text{True})$	all	2	3–4	5–15
Qwen3.5-4B	0.47 [0.45, 0.49]	0.42 [0.40, 0.43]	0.61 [0.60, 0.63]	0.51 [0.48, 0.53]	0.65 [0.63, 0.67]	0.75 [0.73, 0.77]
Qwen3.5-9B	0.46 [0.44, 0.48]	0.42 [0.40, 0.44]	0.64 [0.62, 0.65]	0.49 [0.46, 0.51]	0.67 [0.65, 0.69]	0.84 [0.82, 0.86]
Qwen3.5-27B	0.47 [0.46, 0.49]	0.54 [0.52, 0.56]	0.74 [0.72, 0.75]	0.59 [0.57, 0.62]	0.79 [0.77, 0.81]	0.92 [0.91, 0.93]
Gemma 4 12B	0.48 [0.46, 0.49]	0.48 [0.46, 0.50]	0.65 [0.63, 0.67]	0.52 [0.50, 0.55]	0.69 [0.67, 0.71]	0.81 [0.79, 0.83]
Phi-4 14B	0.50 [0.48, 0.52]	0.44 [0.42, 0.46]	0.69 [0.67, 0.71]	0.56 [0.53, 0.58]	0.72 [0.70, 0.74]	0.87 [0.85, 0.89]

E.3 Decompositions: intervals and robustness

Tables 14 and 15 give the bootstrap intervals of the point estimates of Table 4. Two further checks concern the reading of that table.

E.4 How the clarifier behaves with controls

A control is a question whose description fits exactly one entity under the membership rule, but a clarifier working from its own knowledge may still propose more. On the 418 subset controls, Sol gave 95 (23%) two or more interpretations, naming 306 things in total: 75 (25%) are the control's own interpretation, 15 (5%) match a member that the generator dropped for lacking a Wikipedia article, a real entity outside the benchmark's universe, and 216 (71%) could not be matched to any Wikidata member of the group. The last group mixes invention with entities that have no Wikidata

Table 14: 95% bootstrap intervals for the AUROC of the aleatoric term of each decomposition with the Sol clarifier (Table 4, first column).

model	SD	ICE
Qwen3.5-4B	0.89 [0.87, 0.91]	0.89 [0.86, 0.91]
Qwen3.5-9B	0.89 [0.87, 0.91]	0.89 [0.87, 0.91]
Qwen3.5-27B	0.90 [0.88, 0.92]	0.89 [0.87, 0.91]
Gemma 4 12B	0.88 [0.86, 0.91]	0.89 [0.87, 0.91]
Phi-4 14B	0.88 [0.86, 0.90]	0.88 [0.86, 0.91]

Table 15: 95% bootstrap intervals for the epistemic term against coverage and the aleatoric term against the count, for both decompositions (Table 4).

model	method	epistemic, ρ(coverage)	aleatoric, ρ(count)
Qwen3.5-4B	SD	-0.36 [-0.42, -0.30]	$+0.70$ [+0.66, +0.74]
	ICE	-0.28 [-0.34, -0.21]	$+0.71$ [+0.66, +0.75]
Qwen3.5-9B	SD	-0.34 [-0.41, -0.28]	$+0.70$ [+0.66, +0.74]
	ICE	-0.29 [-0.36, -0.22]	$+0.73$ [+0.69, +0.77]
Qwen3.5-27B	SD	-0.36 [-0.42, -0.30]	$+0.71$ [+0.67, +0.75]
	ICE	-0.33 [-0.39, -0.26]	$+0.73$ [+0.68, +0.77]
Gemma 4 12B	SD	-0.37 [-0.43, -0.30]	$+0.59$ [+0.54, +0.64]
	ICE	-0.29 [-0.35, -0.22]	$+0.71$ [+0.67, +0.76]
Phi-4 14B	SD	-0.32 [-0.39, -0.26]	$+0.58$ [+0.53, +0.63]
	ICE	-0.23 [-0.29, -0.17]	$+0.73$ [+0.68, +0.77]

item and cannot be separated without a hand check; at most 4% of the controls therefore have a genuine second interpretation.

F PROMPTS

The prompts below are the ones used in every experiment, each sent through the model's chat template as a system message followed by a user message.

Prompt 1: Greedy answer, semantic-entropy samples and teacher-forced probability.

```
System: You are a precise question-answering assistant. Reply with the
    answer only, no explanation.

User: {format_hint} If you are not sure, give your best guess anyway.
    Do not say that you are unsure and do not ask questions.
Question: {question}
```

Prompt 2: $P(\text{True})$: the greedy answer is shown back and the probability of the first token is read from the logits.

```
System: You are a precise question-answering assistant. Reply with the
    answer only, no explanation.

User: Question: {question}
Proposed answer: {greedy answer}
Is the proposed answer correct? Reply with exactly one word: True or
    False.
```

Prompt 3: Ambiguity probe: $P(\text{Yes})$ is read from the first-token logits.

```
System: You are a precise question-answering assistant. Reply with the
    answer only, no explanation.

User: Question: {question}
Could this question refer to more than one thing, so that it has more
    than one correct answer depending on which one is meant? Reply with
    exactly one word: Yes or No.
```

Prompt 4: Enumeration of the interpretations, asked after the probe.

```
System: You are a precise question-answering assistant. Output only the
    list, no explanation.

User: Question: {question}
This question may refer to more than one thing. List every thing it
    could refer to that you know of, one per line, in the form '<name>:
    <answer>', where <answer> is the answer to the question for that
    thing. For the <answer> part: {format_hint} If the question can
    refer to only one thing, write just that one line.
```

Prompt 5: Clarifier prompt (GPT-5.6 Sol, reasoning off, temperature 0); the question is appended after the last line.

```
System: You are a helpful assistant.

User:
**Objective**
The question below describes a thing (a film, a book, an album, an
    opera, a person, a club, ...) by one of its properties instead of
    naming it, for example "the novel written by George Orwell" or "the
    child of Charles Darwin". Your task is to decide which real things
    fit that description, using your factual knowledge of the world, and
    to write one clarified question per thing.

**Rules**
1. First list every real thing that fits the description. For "the
    novel written by George Orwell" that is Animal Farm, Nineteen
    Eighty-Four, Burmese Days, and so on; for "the child of Charles
    Darwin" that is William, Anne, Henrietta, George, Francis, and so
    on. Do not invent things you are not sure exist.
2. Then write one clarified question per thing: the original question
    with the description replaced by the thing's name (for example "In
    what year was Animal Farm published?"). If two things share a name,
    add a distinguishing detail in parentheses, such as the year.
3. Do not add clarifications that merely rephrase the question, split
    it by time period, format, region, or interpretation, or ask about
    something the question did not ask. Only distinct real things count.
4. If exactly one real thing fits the description, output exactly one
    clarification: the original question unchanged.
5. Do not give more than 15 clarifications.

**Output Format**
### Things:
[one real thing per line, name only, with a year in parentheses if it
    helps to tell it apart]
Count: [the number of things]

### Clarifications:
#1 [clarified question for the first thing]
#2 [clarified question for the second thing]
...

If exactly one thing fits:
### Things:
[the one thing]
Count: 1

### Clarifications:
#1 [the original question, unchanged]

**Task**
Question: {question}
```

Prompt 6: Answer prompt for SD and ICE, from Walha et al. (2026); each clarification is appended after the last line and sampled ten times at temperature 0.5.

```
System: You are a helpful assistant.

User:
**Objective**
In the following, I will provide a question and you need to provide a
    corresponding answer. Your answer has to be short and precise. Do
    not write extra text or explanation, just give the answer directly.
    If the question is unclear or you do not know the answer, do not
    answer with phrases like "I'm sorry.." or "The question is unclear".
    Instead, you need to give a random guess for the answer. Do not ask
    follow-up questions or indicate that you do not know the answer. You
    should always provide a short and precise answer; either the true
    answer if you know it or your random guess if you are unsure. It
    should not be recognizable in your output whether your answer is the
    true answer or the random guess.
Your output should follow the format specified below in the Output
    Format and Example sections.

**Output Format**
Answer: [Your short and precise answer or random guess. Do not include
    any additional information.]

**Examples**
Question: When did the british army got final defeat against the united
    state of america?
Answer: February, 1815

Question: What kind of dog in little rascals movie?
Answer: doberman pinscher

Question: Where does the last name carson come from?
Answer: Scottish and Irish origin

Question: Who wrote the music for game of thrones?
Answer: Ramin Djawadi


**Task**
Question: {clarification}
```