DEMYSTIFYING AND RESOLVING THE PATHOLOGIES OF LINGUISTIC UNCERTAINTY IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Across diverse languages, speakers naturally convey uncertainty through epistemic hedges before asserting claims; however, uncertainty quantification (UQ) in large language models (LLMs) relies almost exclusively on post-hoc numerical scores. Can LLMs learn human-like, verbalized uncertainty? To investigate this, we systematically analyze model behavior along two primary axes: confidence placement (pre- vs. post-assertion) and modality (numeric scores vs. linguistic hedges). Our analysis uncovers two critical pathologies: (i) pre-assertion confidence acts as an ungrounded prior that is statistically decoupled from the downstream assertion; and (ii) linguistic hedging is acutely brittle under prompt paraphrasing. Consequently, naive epistemic hedging severely degrades model reliability. To remedy these deficiencies, we propose STRATIFIED, a framework that decomposes generation into distinct epistemic and linguistic tiers by first committing to a statement and its numeric confidence before verbalizing the hedged assertion. Furthermore, to elicit fine-grained hedging across long-form responses, we construct WIKIEU, a benchmark spanning diverse, factually verifiable queries, alongside an optimization recipe combining confidence-balanced supervised fine-tuning and reinforcement learning with a gated reward mechanism. Across three benchmarks and two models, our approach substantially improves verbalized uncertainty expression without sacrificing factual fidelity.

1 INTRODUCTION

In languages such as English and Chinese, speakers frequently front uncertainty before introducing a claim, using epistemic adverbs (*perhaps*, *probably*) or matrix clauses ("*I suspect that...*") (Hyland, 1998; Kärkkäinen, 2003). This fronting mechanism establishes an epistemic frame, calibrating the interlocutor's expectations before the core proposition unfolds (Lakoff, 1973). In contrast, uncertainty quantification (UQ) in large language models (LLMs) typically operates as a retrospective step, requiring models to produce the complete answer before emitting a numeric confidence score (Lin et al., 2022; Kadavath et al., 2022). This formulation departs fundamentally from human communication in both temporal ordering and communicative function, raising a central question: **Can LLMs learn to convey uncertainty through natural, human-like epistemic hedging?**

Prior studies have explored epistemic hedging via prompting (Yona et al., 2024; Liu et al., 2025), supervised fine-tuning (SFT, Mielke et al. 2022; Band et al. 2024; Chaudhry et al. 2025; Eikema et al. 2025), and reinforcement learning (RL, Liu et al. 2026). However, existing efforts largely treat verbal hedging as a surface formatting objective, leaving its underlying behavioral dynamics and epistemic fidelity unexamined.

To systematically investigate these mechanisms, we conduct a factorial study on TriviaQA (Joshi et al., 2017) spanning two axes: **position** (pre-assertion vs. post-assertion) $\times$ **modality** (numeric percentage vs. verbal hedge). Each format is optimized using Group Relative Policy Optimization (GRPO, Shao et al. 2024) under a reward objective that couples answer correctness with confidence calibration (Damani et al., 2026). Two key findings emerge:

- **Pre-assertion confidence acts as an ungrounded prior.** Sweeping fronted confidence from 10% to 100% leaves downstream answer probabilities essentially unchanged (Pearson's $|r| <$

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.

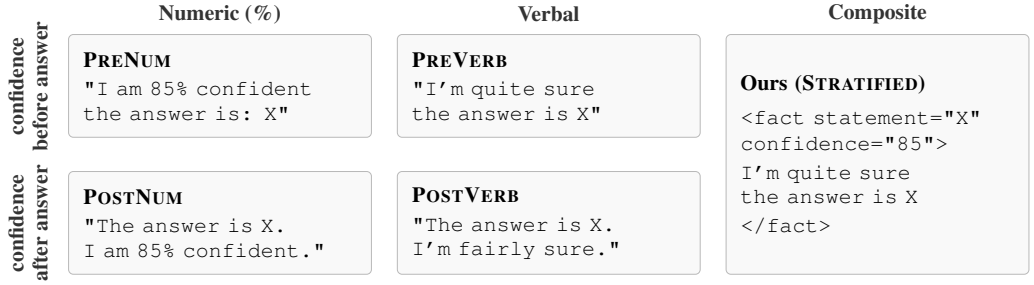


Figure 1: Factorial formulation of uncertainty elicitation. The left 2×2 grid spans **position** (pre-assertion vs. post-assertion) and **modality** (numeric vs. verbal). The right panel illustrates the proposed STRATIFIED format, which first generates statements with numeric confidence scores before rendering natural verbal hedges.

0.1). Unable to condition on the ungenerated assertion, it fails to reflect answer correctness and lags post-assertion confidence by 6 AUROC points.

- **Verbal hedging is acutely brittle to surface phrasing.** Under paraphrased system prompts that strictly preserve the formatting contract, verbal confidence degrades by 0.143 AUROC on average, whereas numeric confidence shifts by ≤ 0.021, exposing the severe vulnerability of verbal hedging across differing prompt contexts.

These findings demonstrate that directly fronting verbal hedges severely undermines reliability. To address this, we propose STRATIFIED, a two-tier framework that first commits to an assertion and its confidence before verbalizing the hedged response, thereby grounding verbal hedges in explicit, numeric posterior estimates.

We next extend this formulation to long-form generation involving multiple atomic statements, where open-ended verification is plagued by retrieval noise and invites reward hacking via the selective reporting of safe, trivial facts (Yang et al., 2025). To enable both controlled training and rigorous evaluation, we construct WIKIEU (**Wiki**data **E**pistemic **U**ncertainty Benchmark), conditioning entity profiles on specified properties grounded in verified relational triples. On WIKIEU, we train models via a two-stage recipe: Confidence-Balanced SFT (CB-SFT) to upweight moderate-uncertainty statements against binary collapse, followed by Hierarchical Gated RL (HG-RL), which pairs span-level credit assignment with a response-level quality gate to calibrate individual assertions without compromising global fidelity.

Experiments on WIKIEU, UNCLE (Yang et al., 2025), and HotpotQA (Yang et al., 2018) across Qwen2.5-7B-Instruct (Qwen Team, 2025) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) show that our training recipe substantially improves the reliability of verbal hedging while maintaining factual fidelity, and that STRATIFIED consistently outperforms direct hedging across all settings.

2 PROBLEM FORMULATION

2.1 NUMERIC VS. LINGUISTIC CONFIDENCE

Uncertainty quantification (UQ) is conventionally formulated as estimating a scalar confidence score $c \in [0, 1]$ via an estimator $c = f(x, y)$ (Kadavath et al., 2022; Kuhn et al., 2023), where x denotes the input prompt and y is the model's generated answer. The estimated confidence is evaluated against ground-truth binary correctness $z = \mathbb{I}(y \equiv y^*) \in \{0, 1\}$, with y^* denoting the reference answer, using standard metrics such as AUROC, ECE, and Brier score.

Unlike post-hoc numeric estimation, natural linguistic hedging spans diverse syntactic structures. In head-initial constructions, speakers canonically frame their epistemic commitment before articulating the claim itself. We formalize this fronted generation process as:

$$u \sim p_\theta(\cdot \mid x), \quad y \sim p_\theta(\cdot \mid x, u), \tag{1}$$

where u denotes the emitted linguistic hedge, and y is the downstream assertion conditioned on both the prompt x and prefix u.

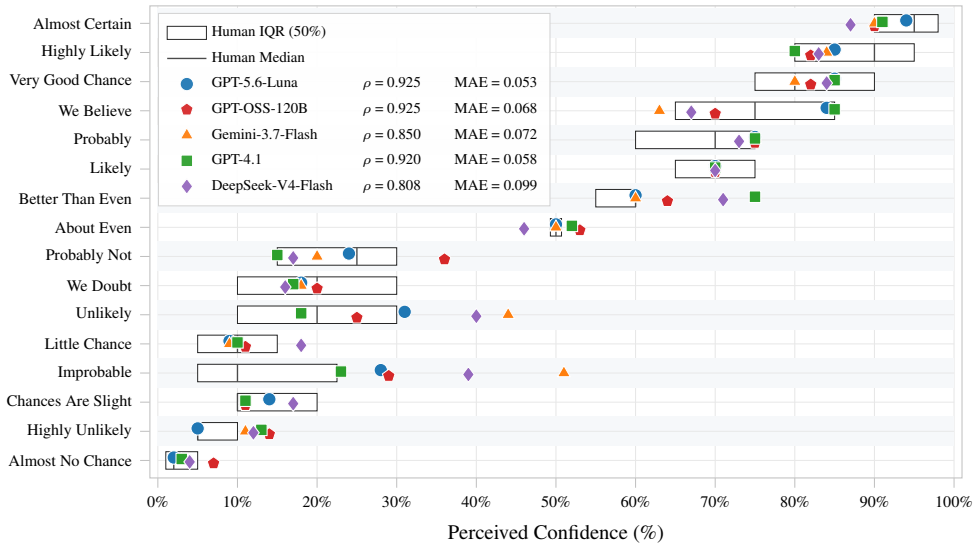


Figure 2: Alignment between LLM-extracted confidence and human perception across 16 linguistic hedging expressions. White boxes represent the human interquartile range (IQR), and vertical lines mark human medians. Markers show perceived confidence predictions from evaluated extractors. The legend summarizes Spearman rank correlation (ρ) and Mean Absolute Error (MAE) against human medians.

2.2 Measuring Linguistic Confidence

Evaluating linguistic confidence introduces a modality mismatch: evaluation metrics inherently require continuous values in $[0, 1]$, whereas linguistic expressions u are discrete text tokens. To bridge this gap, we adopt an LLM-based confidence extractor, $f_\phi : \mathcal{V}^* \to [0, 1]$, which maps an emitted hedge u to an implied scalar confidence $\hat{c} = f_\phi(u)$.

To ensure the reliability of f_ϕ, we benchmark its alignment with human perception following the validation protocol of Yona et al. (2024). Specifically, we sample 500 questions from TriviaQA (Joshi et al., 2017), prepend 16 hedging expressions annotated with human-perceived probabilities from Fagen-Ulmschneider (2023), and evaluate candidate extractors across multiple LLMs.

As shown in Figure 2, all evaluated models exhibit strong rank correlation with human annotations (Spearman's $\rho > 0.8$). While GPT-5.6-Luna achieves the highest overall alignment, GPT-OSS-120B demonstrates competitive correlation while offering open-weight accessibility and deterministic reproducibility. We therefore adopt GPT-OSS-120B as the default confidence extractor f_ϕ throughout this work. Further validation results using GPT-5.6-Luna are provided in §G.

2.3 Reinforcement Learning for Linguistic Confidence

Expressing confidence verbally treats confidence estimation as an alignment objective over text generation. However, standard SFT lacks ground-truth confidence annotations, often necessitating noisy heuristics such as empirical self-consistency (Wang et al., 2023b; Manakul et al., 2023). RL offers a principled alternative by enabling the policy to actively explore verbal expressions under direct calibration feedback.

To this end, we adopt the Reinforcement Learning with Calibration Rewards (RLCR) objective (Damani et al., 2026). Given the extracted confidence $\hat{c} = f_\phi(u)$ and task correctness $z = \mathbb{I}(y \equiv y^*)$, the reward function couples correctness with a Brier-score calibration penalty:

$$R_{\text{RLCR}}(x, u, y; y^*) = z - (\hat{c} - z)^2. \tag{2}$$

This formulation penalizes miscalibrated verbal expressions, incentivizing the policy θ to align its expressed hedging with its empirical probability of success.

Table 1: Performance comparison on TriviaQA with Qwen2.5-7B across position and modality. The best and second-best results within each format group are in **bold** and underlined, respectively.

	Numeric				Linguistic			
	Acc ↑	AUROC ↑	Brier ↓	ECE ↓	Acc ↑	AUROC ↑	Brier ↓	ECE ↓
PRE	.700	.832	.132	.011	.709	.844	.120	.063
POST	.698	**.912**	**.094**	**.010**	**.713**	**.904**	**.104**	.055
STRATIFIED	**.701**	.905	.101	.034	.701	.887	.108	**.030**

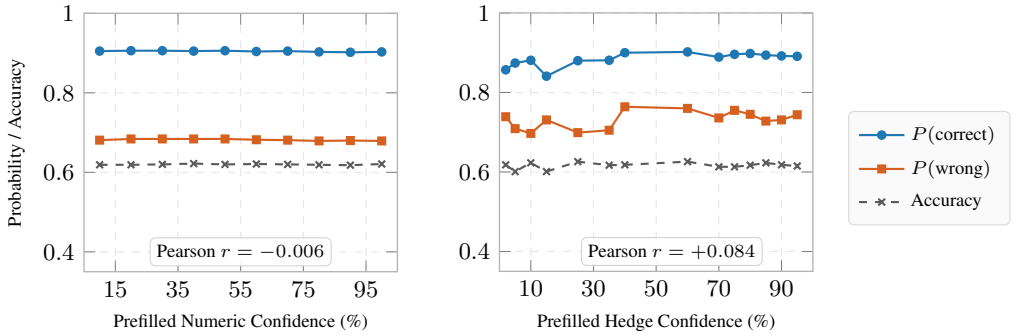


Figure 3: Prefill intervention experiments under varying confidence prefixes: (left) numeric openers sweeping $X \in [10, 100]\%$, and (right) linguistic hedge openers mapped to corresponding confidence values. We separate the correct and incorrect answers for analysis, computing the probability of each answer as the product of token probabilities within the answer span: $P(y) = \prod_t P(y_t \mid y_{<t})$.

3 PART I: CAN LLMS LEARN LINGUISTIC UNCERTAINTY?

3.1 SETUP

Conventional numeric uncertainty quantification and natural linguistic hedging differ across two orthogonal dimensions: temporal placement and representation modality. To disentangle their individual contributions to the reliability gap, as shown in Figure 1, we establish a 2×2 factorial design across position (pre- vs. post-assertion) and modality (numeric scores vs. linguistic hedges), yielding four format variants: PRENUM, PREVERB, POSTNUM, and POSTVERB. In addition, we evaluate our proposed STRATIFIED framework, which bridges both paradigms by grounding fronted linguistic hedges in post-hoc numeric estimates.

We fine-tune Qwen2.5-7B-Instruct (Qwen Team, 2025) on TriviaQA (Joshi et al., 2017), sampling 50,000 questions for training and holding out 7,785 validation questions for evaluation. All models are trained using the RLCR objective, with format violations penalized by -1.

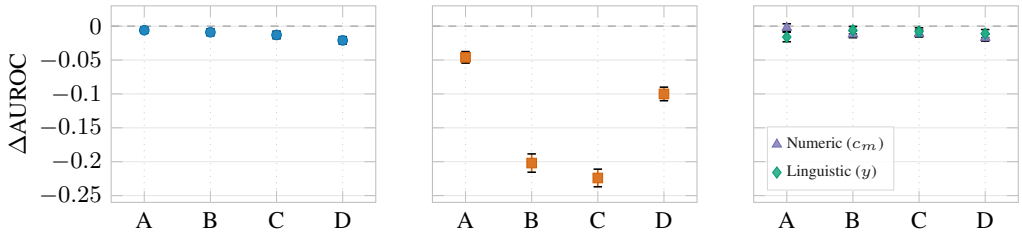
For STRATIFIED, generation factorizes into an epistemic state $m = (a_m, c_m)$ and a downstream hedged response y: $p_\theta(m, y \mid x) = p_\theta(m \mid x)p_\theta(y \mid x, m)$, where a_m is the assertion and $c_m \in [0, 1]$ is its scalar confidence score. The policy is optimized via:

$$R_{\text{STRATIFIED}} = R_{\text{RLCR}}(a_m, c_m; y^*) - \lambda \big| f_\phi(u) - c_m \big|, \tag{3}$$

which rewards correctness and calibration via Equation 2, while penalizing epistemic divergence between the emitted hedge $u \subset y$ and the numeric confidence c_m scaled by hyperparameter λ.

3.2 POSITION: PRE-ASSERTION CONFIDENCE ACTS AS AN UNGROUNDED PRIOR

When emitted prior to an assertion, confidence operates mathematically as a prior $P(c \mid x)$, reflecting general query difficulty rather than the veracity of a specific claim. For this prior to be reliable, emitting confidence c must actively condition downstream generation, steering probability mass toward candidate answers whose correctness aligns with the expressed certainty.



(a) POSTNUM (b) POSTVERB (c) STRATIFIED

Figure 4: Prompt-perturbation sensitivity across formats. Each point denotes the change in AUROC under four prompt perturbations (A–D); whiskers indicate 95% confidence intervals. The horizontal dashed line marks zero performance shift relative to the unperturbed baseline (ΔAUROC $= 0$).

Intervention experiments (Figure 3) reveal that this directional conditioning fails in practice. Downstream answer probabilities remain invariant across prefilled confidence levels, exhibiting negligible Pearson correlations between fronted confidence and sequence probabilities ($r = -0.006$ for PRENUM; $r = 0.084$ for PREVERB). The conditional distribution thus collapses onto the unconditioned marginal: $P(y \mid x, c) \approx P(y \mid x)$. This decoupling directly explains the persistent degradation of pre-assertion formats across Table 1.

In contrast, post-assertion models evaluate the true posterior $P(c \mid x, y)$, grounding confidence directly in the realized answer y. As shown in Figure 8, both generated and perceived confidence strongly correlate with the generation probability of the realized answer. STRATIFIED retains this posterior grounding by first committing to an epistemic state $m = (a_m, c_m)$ before verbalizing the hedged claim, reconciling post-hoc reliability with natural linguistic hedging.

3.3 MODALITY: LINGUISTIC CONFIDENCE IS BRITTLE TO PROMPT PHRASING

Despite their strong capabilities, LLMs remain sensitive to subtle variations in prompt formatting (Wang et al., 2023a; Sclar et al., 2024). While numeric confidence follows a uniform format, linguistic hedges are embedded within variable syntax. This entangles epistemic signaling with surface syntax, destabilizing the learned uncertainty mapping under distributional shifts despite competitive in-distribution performance. To quantify this sensitivity, we evaluate POSTVERB, POSTNUM, and STRATIFIED under four format-preserving prompt perturbations (A–D).

As shown in Figure 4, linguistic confidence degrades sharply under prompt perturbations: POSTVERB suffers an average AUROC drop of 0.143, whereas numeric readouts shift by at most 0.021. STRATIFIED resolves this fragility by grounding verbalization in the epistemic state $m = (a_m, c_m)$, matching numeric robustness while delivering natural linguistic hedging.

4 PART II: FROM SINGLE CLAIMS TO LONG-FORM RESPONSES

The empirical findings in Part I demonstrate that direct linguistic hedging induces severe reliability issues, whereas STRATIFIED resolves this pathology by decoupling epistemic estimation from surface verbalization. However, these observations were established on short-form question answering. Scaling STRATIFIED to extended, multi-statement generations introduces distinct challenges in verifiable evaluation and granular credit assignment. In this part, we extend our framework to address both challenges.

4.1 DATA PREPARATION: THE WIKIEU BENCHMARK

In open-domain long-form generation, validating claims typically relies on web retrieval (Min et al., 2023; Wei et al., 2024), which is computationally prohibitive and influenced by search noise. To construct a controlled and verifiable testbed, we introduce WIKIEU (**Wiki**data **E**pistemic **U**ncertainty Benchmark), derived from Wikidata (Vrandečić & Krötzsch, 2014), spanning diverse entity popularity levels based on Wikipedia sitelink counts.

For each entity, we curate relational triplets (s, p, o), filtering out ambiguous predicates. We then construct prompts requiring responses that cover all curated triplets for the given entity. While inspired by Yang et al. (2025), our dataset encompasses broader domains and relation types. In addition, we implement quality control to prevent trivial or excessively difficult knowledge from dominating the corpus, thereby improving both training and evaluation. Specifically, across 16 independent rollouts at temperature T=0.8, we downsample triplets that Qwen2.5-7B-Instruct either answers correctly in all rollouts or fails to answer correctly even once. Finally, we split the dataset into 8,980 training and 500 evaluation entities. Full details are provided in §A.

4.2 CB-SFT: Mitigating Mid-Band Collapse

Because off-the-shelf LLMs struggle to express graded uncertainty, we construct a cold-start SFT corpus with diverse hedging expressions prior to RL. Specifically, we treat the empirical success frequency of each triplet across rollouts as its confidence score and map it to a verbal hedge sampled from the corresponding tier. We then employ DeepSeek-V4-Flash (Xu et al., 2026) to synthesize these annotated atomic statements into fluent, coherent paragraphs.

However, mid-confidence statements are relatively less frequent than tail cases in the synthesized data. Standard SFT disproportionately amplifies this mild skew, collapsing the policy into bimodal extremes (expressing either high certainty or overt doubt). Crucially, this collapse deprives subsequent RL of generative support to explore intermediate uncertainty.

To resolve this issue, we propose **Confidence-Balanced SFT (CB-SFT)**, which applies inverse-frequency loss reweighting at the statement level during training:

$$\mathcal{L}_{\text{CB-SFT}}(\theta) = -\sum_{t=1}^{T} w(y_t) \log p_\theta(y_t \mid x, y_{<t}), \quad \text{where} \quad w(y_t) = \frac{1/K}{p(c(y_t))}. \tag{4}$$

Here, confidence is partitioned into K=10 uniform tiers ($c \in \{0.1, 0.2, \ldots, 1.0\}$), $c(y_t)$ denotes the confidence tier of the statement encompassing token y_t, and $p(c)$ is its empirical frequency in the corpus. By equalizing gradient contributions across all confidence tiers, CB-SFT prevents mid-band collapse and provides balanced behavioral priors for policy optimization.

4.3 HG-RL: Statement-Level Uncertainty with Gated Rewards

While WikiEU provides verifiable ground truth for individual assertions, scaling RL to multi-statement generation introduces a fundamental credit assignment dilemma:

- **Global Credit Blurring**: Aggregating claim rewards into a single sequence-level scalar ($\sum_i R_i$) entangles heterogeneous feedback. For instance, a response pairing one accurate statement ($+1$) with one hallucination (-1) nets to zero, extinguishing the learning signal for both.
- **Local Reward Exploitation**: Conversely, assigning localized rewards in isolation invites degenerate shortcuts: policies readily game the objective by cherry-picking trivial facts, truncating generations, or abandoning discourse coherence to evade uncertainty penalties.

To reconcile these competing objectives, we propose **Hierarchical Gated RL** (**HG-RL**). HG-RL establishes a hierarchical dependency between global structure and local feedback: granular claim-level rewards are unlocked only after the response satisfies a holistic quality gate.

Formally, consider a response $\mathbf{y} = (y_1, \ldots, y_T)$ partitioned into K disjoint token spans $\mathcal{S}_1, \ldots, \mathcal{S}_K \subset \{1, \ldots, T\}$, each corresponding to an atomic statement. We evaluate generation quality via a global score R_{glo} alongside statement-level rewards $\{R_i\}_{i=1}^K$. Regulated by a compliance threshold δ, the gating indicator $g = \mathbb{I}(R_{\text{glo}} \geq \delta)$ determines the token-level reward r_t:

$$r_t = (1 - g)\frac{R_{\text{glo}}}{T} + g \sum_{i=1}^{K} \frac{R_i}{|\mathcal{S}_i|} \mathbb{I}(t \in \mathcal{S}_i). \tag{5}$$

Under this formulation, non-compliant outputs ($g = 0$) suppress granular feedback, compelling the policy to optimize for global validity. Once the response clears the gate ($g = 1$), calibration rewards are routed strictly to their constituent spans, with the length normalizer $|\mathcal{S}_i|$ eliminating verbosity bias. Finally, to stabilize training, token advantages are normalized across multiple rollouts for the same underlying statement.

Table 2: Long-form evaluation results on WIKIEU and UNCLE.

Method	WIKIEU				UNCLE			
	Acc ↑	AUROC ↑	ECE ↓	Brier ↓	Acc ↑	AUROC ↑	ECE ↓	Brier ↓
Qwen2.5-7B-Instruct								
BASELINE								
⌞ CB-SFT	.559	.598	.219	.295	.383	.577	.296	.337
⌞ HG-RL	**.614**	.725	**.066**	.198	.412	.737	.240	.260
STRATIFIED								
⌞ CB-SFT	.545	.709	.123	.230	.393	.686	.176	.254
⌞ HG-RL	.601	**.827**	.078	**.164**	**.427**	**.804**	**.163**	**.204**
Llama-3.1-8B-Instruct								
BASELINE								
⌞ CB-SFT	.622	.579	.238	.307	.470	.593	.248	.307
⌞ HG-RL	**.687**	.672	**.071**	.195	**.531**	.730	.189	.242
STRATIFIED								
⌞ CB-SFT	.605	.704	.125	.229	.490	.683	**.117**	.241
⌞ HG-RL	.666	**.788**	.093	**.171**	.519	**.762**	.169	**.225**

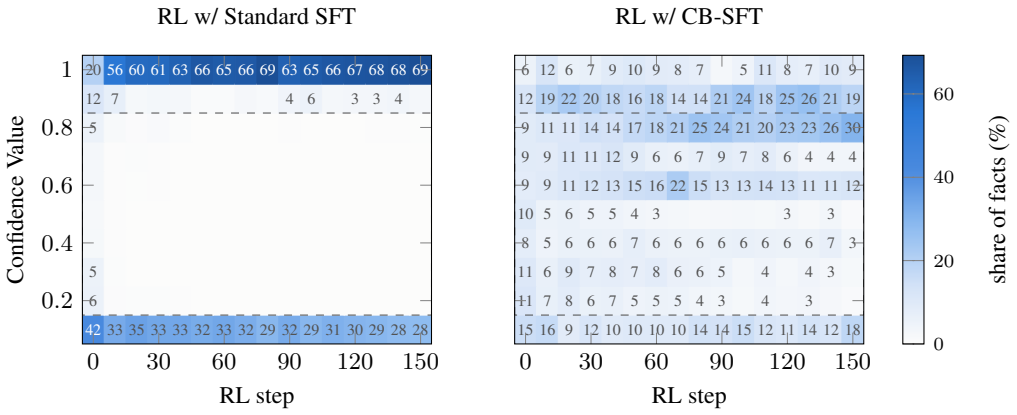


Figure 5: Evolution of statement-level confidence distributions during HG-RL training on the WIKIEU test set. Heatmaps display statement density across confidence tiers over RL steps for policies initialized from standard SFT (left) versus CB-SFT (right).

4.4 RESULTS

Beyond WIKIEU, we evaluate out-of-distribution generalization on the long-form benchmark UNCLE (Yang et al., 2025). As summarized in Table 2, STRATIFIED consistently outperforms the baseline format across both benchmarks and backbone architectures on AUROC, confirming that decoupling epistemic estimation from surface verbalization is essential for long-form reliability.

CB-SFT successfully preserves expressive diversity across confidence levels: under vanilla SFT, mid-band statements (0.2–0.8) virtually vanish during subsequent RL (26.3% → 0.3%), whereas CB-SFT maintains balanced generative support throughout training (67.0% → 54.7%; Figure 5). Building on these warm-up priors, RL optimization proves indispensable for discriminative calibration, driving marked AUROC gains over SFT checkpoints. Furthermore, **HG-RL** substantially outperforms the **Global-RL** baseline, reaching an AUROC of 0.848 versus 0.805, while reducing both Brier score (0.145 vs. 0.168) and ECE (0.031 vs. 0.048; Figure 6). Throughout training, the validation pass rate of the quality gate remains consistently high across both backbones ($\geq$ 0.89 for Qwen2.5-7B from step 0; Figure 9), indicating that the gate selectively filters early degenerate rollouts without starving the policy of granular span-level learning signals.

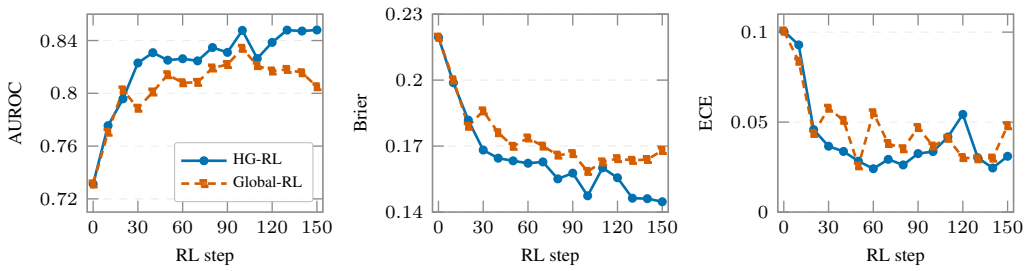


Figure 6: Confidence estimation performance (AUROC, Brier score, and ECE) across RL training steps on the WIKIEU test set, evaluated on numeric confidence in epistemic states for direct comparison. We compare HG-RL against the Global-RL baseline, where claim rewards are pooled into a single sequence-level scalar.

Table 3: Out-of-distribution transfer results on HotpotQA stratified by native question difficulty.

Method	Difficulty	Acc ↑	AUROC ↑	ECE ↓	Brier ↓
BASELINE	Easy	0.370	0.711	0.283	0.293
	Medium	0.200	0.677	0.290	0.280
	Hard	0.236	0.723	0.294	0.275
STRATIFIED	Easy	0.354	0.715	0.095	0.195
	Medium	0.198	0.691	0.088	0.143
	Hard	0.246	0.730	0.051	0.156

To probe complex reasoning under distribution shift, we further evaluate our framework on HotpotQA (Yang et al., 2018). Because the provided evidence sentences cannot reliably support the factual validation of isolated intermediate claims, we average statement confidence scores to evaluate against final answer correctness. As shown in Table 3, although AUROC improvements are modest, the calibration metrics improve substantially. These results confirm that our framework generalizes robustly to multi-step reasoning queries that differ substantially from the training distribution.

We provide comprehensive analyses in the Appendix to further substantiate our findings. Specifically, CB-SFT substantially enriches hedging diversity and epistemic coverage (§D). Evaluator-swap experiments with GPT-5.6-Luna confirm that our conclusions remain robust across different confidence judges (§G). Additionally, though STRATIFIED enables more precise linguistic expressions (§F), it incurs a decoding overhead from structured tags, and paired bootstrap tests establish the statistical significance of all key improvements (§H).

5 RELATED WORK

Uncertainty quantification in LLMs encompasses predictive probabilities (Kadavath et al., 2022; Farquhar et al., 2024), sampling consistency across stochastic decodings (Wang et al., 2023b; Manakul et al., 2023; Kuhn et al., 2023), and direct numerical verbalization (Tian et al., 2023; Xiong et al., 2024). Beyond raw numbers, communicating uncertainty through natural language hedging provides an intuitive, human-aligned interface (Mielke et al., 2022; Lin et al., 2022). In this direction, recent research has actively probed the alignment between internal belief states and surface expressions, introducing benchmarks and projection protocols to analyze how linguistic markers behave across varied tasks (Yona et al., 2024; Liu et al., 2025; Liu & Cohan, 2026; Ackerman, 2026). Although these studies also examine linguistic expressions, they differ in their alignment objective: they focus on aligning verbalizations with internal model beliefs (*faithful confidence*) rather than empirical correctness (*factual confidence*). In contrast, we focus on conventional factual confidence, as it directly helps users determine whether or not to trust model outputs in downstream decision-making. Beyond this conceptual distinction, while prior work predominantly documents empirical verbalization behaviors under standard autoregression, we conduct an in-depth mechanistic inves-

tigation into the fundamental divergence between canonical uncertainty quantification and surface linguistic hedging.

For model training, in the absence of ground-truth confidence, prior studies typically adopt empirical correctness frequencies across sample rollouts as confidence proxies for SFT (Chaudhry et al., 2025; Eikema et al., 2025). Beyond imitation, RL has emerged as a promising paradigm to optimize non-differentiable uncertainty metrics (Damani et al., 2026; Bani-Harouni et al., 2026). However, existing methods mainly target short responses with limited claims and adopt a monolithic sequence-level scalar reward (Band et al., 2024; Zhang et al., 2025; 2026). Liu et al. (2026) decoupled numeric confidence estimation from verbal hedging via post-hoc rewriting, bypassing end-to-end policy optimization. In this work, we improve these methods by introducing **CB-SFT** as warm-up and **HG-RL** to balance granular, statement-level credit with global discourse validity.

6 Discussion and Limitations

While this work demonstrates empirical improvements in eliciting human-like verbalized uncertainty, several limitations remain. We discuss these aspects below to contextualize our results and guide future work.

Extension to Larger Model Scales. Our empirical investigations focus on open-source models at the 7B and 8B parameter scale (Qwen2.5-7B and Llama-3.1-8B). Because untrained base models exhibit severe miscalibration and unreliable verbalized uncertainty, which obscures meaningful comparisons between the theoretical efficacy of different methods, we explicitly adopt a post-training regime. We prioritize these models because they serve as representative and widely adopted open-source backbones in recent calibration studies (Damani et al., 2026; Zhang et al., 2026), frequently outperforming much larger models (Wang et al., 2025) after optimization. Due to computational constraints, we were unable to train larger models under this RL regime. Nevertheless, because larger models share the identical underlying paradigm, we expect the pathologies of verbal uncertainty, including ungrounded priors and prompt brittleness, to persist across model scales.

Generalization across Diverse Tasks. Our evaluation centers on knowledge-intensive question answering, where atomic propositions can be validated against reference knowledge from knowledge bases such as Wikipedia and Wikidata. Extending this paradigm to unconstrained, open-ended generation presents notable practical challenges: it requires robust claim-boundary segmentation in continuous prose, and verification often necessitates web retrieval, which can introduce external noise into both training and evaluation.

Relation to Test-Time Computation. Stratified incurs additional test-time computation due to the explicit serialization of epistemic states prior to emitting hedged text. This design shares a conceptual motivation with thinking mechanisms (Guo et al., 2025), where the model plans before generating the final response. However, unconstrained thinking processes complicate fine-grained credit assignment in RL because both free-form reasoning traces and response sentences must be parsed for localized rewards. In addition, Stratified offers distinct interface advantages, such as allowing client applications to visually highlight low-confidence spans to alert users.

7 Conclusion

In this work, we systematically investigate how generation order and modality govern verbal uncertainty in large language models. Our findings demonstrate that pre-assertion verbal confidence fails to condition downstream outputs, acting as an ungrounded prior, while surface verbal hedging remains brittle under prompt perturbations. To resolve this generative tension, we introduce the Stratified framework, which grounds factual claims and numeric confidence in structured epistemic states prior to emitting hedged prose. Extending this design to long-form discourse, we pair confidence-balanced SFT (**CB-SFT**) to prevent mid-band representation collapse with hierarchical gated RL (**HG-RL**) to reconcile granular span-level credit assignment with global response quality. Empirical evaluations across three benchmarks and two backbone models demonstrate that Stratified consistently outperforms direct surface hedging, preserving fact accuracy while substantially enhancing statement-level confidence estimation.

AI USE STATEMENT

In this work, we used generative AI tools solely for code assistance, text polishing, and proofreading. All foundational research contributions, including scientific ideation, methodological formulation, experimental design, and paper organization, were conceived and conducted entirely by the authors without AI assistance. For all AI-assisted tasks, we manually reviewed and tested any generated code to ensure its correctness and alignment with our proposed methodology. Similarly, all AI-polished text was carefully examined to maintain complete fidelity to our implementation and empirical findings, ensuring the absence of hallucinations. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

Implementation details, including data preprocessing and training hyperparameters, are elaborated in §A and §B. Statistical significance tests supporting our empirical findings are reported in §H. Furthermore, the source code to reproduce all experimental results is provided in the supplementary material.

REFERENCES

Christopher Ackerman. Evidence for limited metacognition in llms. In *International Conference on Learning Representations*, volume 2026, pp. 155228–155253, 2026.

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. Linguistic calibration of long-form generations. In *Forty-first International Conference on Machine Learning*, 2024.

David Bani-Harouni, Chantal Pellegrini, Paul Stangel, Ege Özsoy, Kamilia Zaripova, Nassir Navab, and Matthias Keicher. Rewarding doubt: A reinforcement learning approach to calibrated confidence expression of large language models. In *International Conference on Learning Representations*, volume 2026, pp. 117927–117943, 2026.

Arslan Chaudhry, Sridhar Thiagarajan, and Dilan Gorur. Finetuning language models to emit linguistic expressions of uncertainty. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*, 2025.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training lms to reason about their uncertainty. In *International conference on learning representations*, volume 2026, pp. 59456–59493, 2026.

Bryan Eikema, Evgenia Ilia, José GC de Souza, Chrysoula Zerva, and Wilker Aziz. Teaching language models to faithfully express their uncertainty. *arXiv preprint arXiv:2510.12587*, 2025.

Wade Fagen-Ulmschneider. Perception of probability words. `https://waf.cs.illinois.edu/visualizations/Perception-of-Probability-Words/`, 2023.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Ken Hyland. Hedging in scientific research articles. 1998.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

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.

Elise Kärkkäinen. *Epistemic stance in English conversation*, volume 115. John Benjamins Publishing Company Amsterdam & Philadelphia, PA, 2003.

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

George Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508, 1973.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Gabrielle Kaili-May Liu and Arman Cohan. Can llms use linguistic uncertainty markers to reliably reflect intrinsic confidence? *arXiv preprint arXiv:2605.28778*, 2026.

Gabrielle Kaili-May Liu, Gal Yona, Avi Caciularu, Idan Szpektor, Tim GJ Rudner, and Arman Cohan. Metafaith: Faithful natural language uncertainty expression in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29600–29644, 2025.

Gabrielle Kaili-May Liu, Avi Caciularu, Gal Yona, Idan Szpektor, and Arman Cohan. Reinforcement learning with metacognitive feedback elicits faithful uncertainty expression in llms. *arXiv preprint arXiv:2606.32032*, 2026.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 12076–12100, 2023.

Qwen Team. Qwen2.5 technical report, 2025.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

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

Ante Wang, Weizhi Ma, and Yang Liu. Let the model distribute its doubt: Confidence estimation through verbalized probability distribution. *arXiv preprint arXiv:2511.14275*, 2025.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. 2023a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023b.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827, 2024.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Dong Yu, Nigel Collier, and Deqing Yang. Uncle: Benchmarking uncertainty expressions in long-form generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30328–30344, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7752–7764, 2024.

Caiqi Zhang, Ruihan Yang, Zhisong Zhang, Xinting Huang, Sen Yang, Dong Yu, and Nigel Collier. Atomic calibration of llms in long-form generations. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 148–169, 2025.

Caiqi Zhang, Xiaochen Zhu, Chengzu Li, Nigel Collier, and Andreas Vlachos. Lovec: Reinforcement learning for better verbalized confidence in long-form generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33336–33363, 2026.

A Dataset Details

TriviaQA uses $50{,}000$ training examples and the full $7{,}785$-question validation split for evaluation.

WikiEU is curated from Wikidata via a sitelink-ordered SPARQL query covering $10{,}111$ entities ($246{,}744$ relational triplets (s, p, o)). Triplets sharing the same subject and predicate are aggregated into predicate groups $(s, p, \mathcal{O})$. To account for multi-valued relations, we exhaustively retrieve all valid Wikidata objects associated with each (s, p) pair. After filtering ambiguous predicates and ill-formed entities, $9{,}480$ entities retain at least one valid predicate group. Each group is empirically probed using 16 rollouts from Qwen2.5-7B-Instruct at temperature $T{=}0.8$; the empirical success frequency serves as an intrinsic difficulty proxy. Trivial groups (0% or 100% accuracy) are down-sampled rather than discarded to maintain broad coverage across intermediate uncertainty tiers. Data splits are strictly entity-disjoint, allocating 500 entities for evaluation, $1{,}000$ for SFT, and $7{,}980$ for RL. To compose diverse query instances for both SFT and RL, we randomly subsample triplet subsets for each entity. This process yields $6{,}300$ augmented queries for SFT (with responses synthesized into fluent multi-claim discourse via DeepSeek-V4-Flash) and $25{,}405$ augmented queries for RL training. Table 4 summarizes the complete curation funnel and an example is shown in Figure 7.

Table 4: WIKIEU construction funnel and partition statistics. Triplets sharing identical subject-predicate pairs are aggregated into predicate groups $(s, p, \mathcal{O})$. In Phase 3, zero-accuracy groups are capped at one per entity (6,000 globally), and 60% of perfect-accuracy groups are retained.

Pipeline Stage	Asset / Metric	Count
Phase 1: Knowledge Extraction & Semantic Filtering		
Wikidata SPARQL retrieval	Unique candidate entities	10,111
	Relational triplets (s, p, o)	246,744
Constraint & schema filtering	Valid predicate groups $(s, p, \mathcal{O})$	90,269
Phase 2: Empirical Difficulty Probing		
Behavioral rollout evaluation	Evaluated predicate groups	90,269
	Unfamiliar / Intractable (Acc $= 0$)	61,690
	Graded uncertainty ($0 <$ Acc < 1)	14,757
	Highly familiar (Acc $= 1$)	13,822
Phase 3: Difficulty Rebalancing		
Bimodal down-sampling	Retained predicate groups $(s, p, \mathcal{O})$	27,898
	Ground-truth assertions (s, p, o)	32,007
	Retained entities (≥ 1 valid group)	9,480
Phase 4: Entity-Disjoint Partitions		
Test split	Entities / Evaluation queries	500 / 500
SFT split	Entities / Augmented queries	1,000 / 6,300
RL split	Entities / Augmented queries	7,980 / 25,405

Example

Question
In a paragraph, introduce the album *DAMN.*, including publication date, language of work or name, genre, number of parts of this work.

PreVerb
DAMN. is an intriguing work to introduce; *it may be that* it was released on April 14, 2017, and *I'm quite certain that* the language of the work is English. *It could be that* its genre is hip-hop, and *there's a slim chance that* it consists of 14 parts.

Stratified
<fact statement="DAMN. was released on April 14, 2017" confidence="0.4"> *Perhaps* DAMN. was released on April 14, 2017. </fact>
<fact statement="DAMN. is in English" confidence="0.9"> As for its language, *I'm quite certain that* it is in English. </fact>
<fact statement="The album's genre is hip-hop" confidence="0.3"> In terms of genre, *there's a chance that* it's hip-hop. </fact>
<fact statement="The album contains 14 tracks" confidence="0.2"> And finally, *it's a wild guess, but* it contains 14 tracks. </fact>

Figure 7: An illustrative example from WIKIEU, featuring a constructed query and responses synthesized across different formats via DeepSeek-V4-Flash.

UNCLE provides 1,066 open-domain long-form questions, evaluated under its native atomic fact verification protocol.

HotpotQA transfer (Table 3) evaluates out-of-distribution multi-hop reasoning on 1,500 bridge-type questions sampled from the distractor training split (500 questions per difficulty tier: easy, medium, and hard), excluding examples with answer leakage in the question stem. Models answer closed-book under greedy decoding without gold context; final correctness is determined by GPT-OSS-120B, and calibration is evaluated by averaging statement confidences against final-answer correctness.

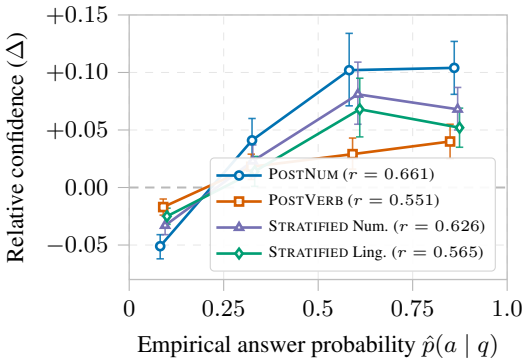


Figure 8: Relative confidence shift (Δ) as a function of empirical answer probability $\hat{p}(a \mid q)$ under answer prefilling. Markers plot binned means; dashed line indicates $\Delta = 0$; error bars denote bootstrap 95% CIs.

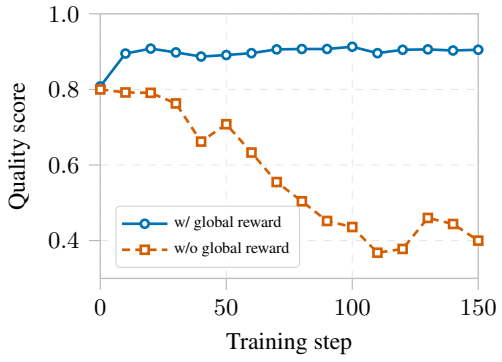


Figure 9: Averaged quality score across training steps on WIKIEU with and without the HG-RL global reward gate.

B TRAINING DETAILS

Supervised Fine-Tuning. Both backbones are fine-tuned for one epoch (98 steps) on 6,300 instances in either STRATIFIED or standard unconstrained formats using Fully Sharded Data Parallel (FSDP). Training uses a batch size of 64, a maximum sequence length of 1,024 tokens, and a learning rate of 5×10^{-6} with cosine decay following a 10% linear warmup. Under the CB-SFT objective (§4.2), atomic statement spans are identified using GPT-OSS-120B for the standard format, while for STRATIFIED they are directly delineated by structured XML markup tags. Training is conducted on 4 NVIDIA A100 (80GB) GPUs.

Reinforcement Learning. All Part I checkpoints are initialized from the base instruct model, whereas Part II checkpoints initialize directly from their corresponding SFT warm-up models. Policies are trained for 200 steps on TriviaQA and 150 steps on WIKIEU. Training employs $n = 8$ rollouts per prompt, a batch size of 256 (64 per mini-batch), a learning rate of 10^{-6}, low-variance KL regularization with coefficient 10^{-4}, and an entropy bonus of 10^{-3}. Maximum prompt and response lengths are both set to 1,024 tokens. For the HG-RL quality gate, the compliance threshold is set to $\delta = 0.7$. For STRATIFIED policy optimization, the verbalization-alignment penalty weight is fixed to $\lambda = 0.2$. Training is conducted on 4 NVIDIA A100 (80GB) GPUs, with 2 additional GPUs dedicated to deploying GPT-OSS-120B for reward computation.

C ANSWER-PREFILL INTERVENTION ANALYSIS

While §3 shows that fronted confidence exerts no causal effect on downstream sequence selection ($P(y \mid do(c), x) \approx P(y \mid x)$), this section tests the converse: does conditioning directly on a realized proposition causally drive the emitted confidence, $P(c \mid do(y{=}a), x)$?

For each question q, we prefill the context with candidate answer a sampled across empirical generation probabilities $\hat{p}(a \mid q) \in [0, 1]$ (16 rollouts at $T{=}1.0$), greedily decode the subsequent confidence statement, and compute the confidence shift Δ relative to each question's mean confidence, thereby controlling for question difficulty as a covariate.

As shown in Figure 8, prefilled answer prefixes causally drive posterior confidence updates across all formats ($r \geq 0.551$). This confirms that post-assertion confidence functions as a true posterior estimate $P(c \mid x, y)$ conditioned on realized answer tokens, directly explaining why post-assertion elicitation consistently outperforms pre-assertion designs in Table 1.

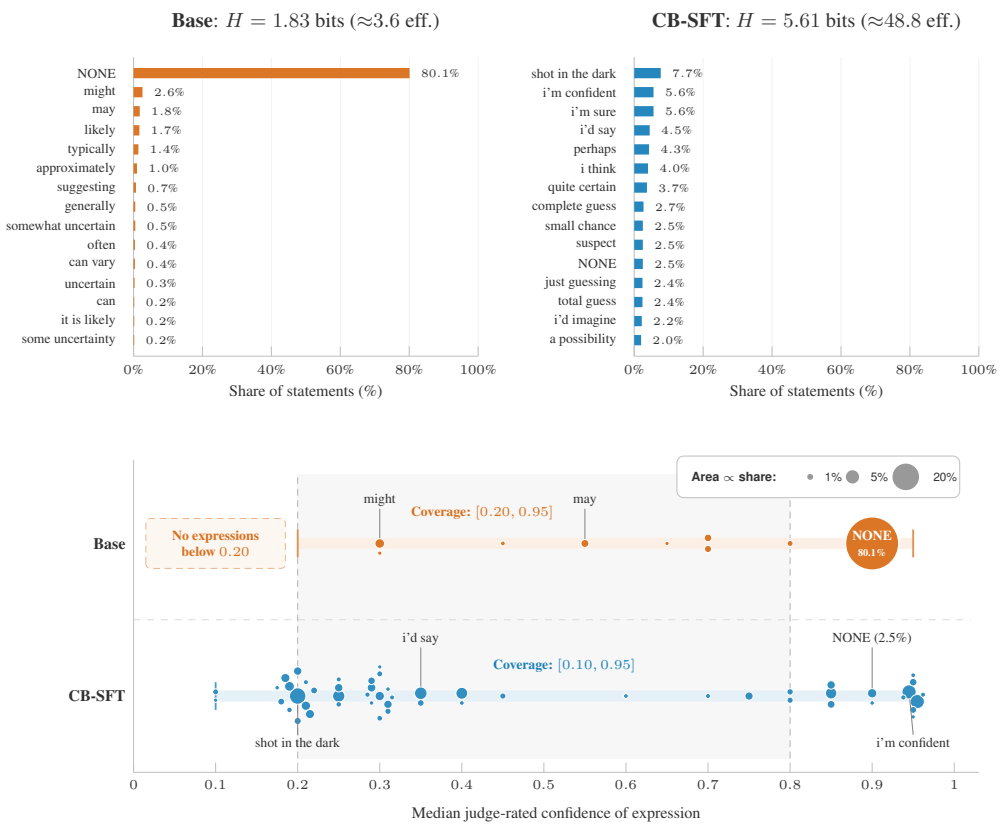


Figure 10: Hedging-expression diversity on the WIKIEU QA test set using Qwen2.5-7B-Instruct. **Top:** Share of the top-15 hedging expressions on a shared 0–100% axis for the zero-shot base model (left) versus CB-SFT (right). Unhedged statements (NONE) are counted as a distinct category; H denotes the Shannon entropy of the full expression distribution. **Bottom:** Implied median judge-rated confidence (`gpt-oss-120b`) per expression, where marker area is proportional to empirical share and horizontal rules indicate total coverage ranges.

D EXPRESSION DIVERSITY

Figures 10 and 11 present the diversity of hedging expressions across both model backbones on WIKIEU (§4). The QWEN base model rarely hedges, leaving the vast majority of statements without an epistemic marker (NONE: 80.1%, $H = 1.83$ bits); CB-SFT substantially broadens this vocabulary, yielding higher entropy ($H = 5.61$ bits) and balanced coverage across confidence bins. While the LLAMA base model adheres much better to the hedging instruction (NONE: 1.3%, $H = 4.52$ bits), it concentrates 98.3% of its hedged statements within the mid-confidence band (0.20–0.80) and lacks expressions below a median judged confidence of 0.20. In contrast, CB-SFT extends low-end and high-end coverage to $[0.12, 0.97]$, enabling the model to express a richer and more calibrated continuum of epistemic states.

E QUALITY GATE ANALYSIS

Figure 9 traces the averaged quality score during training. Without global reward, the response quality drops quickly as LLMs learn to hack the statement-level reward through selectively generating facts, resulting in incomplete responses to the instruction, validating the motivation of our HG-RL method.

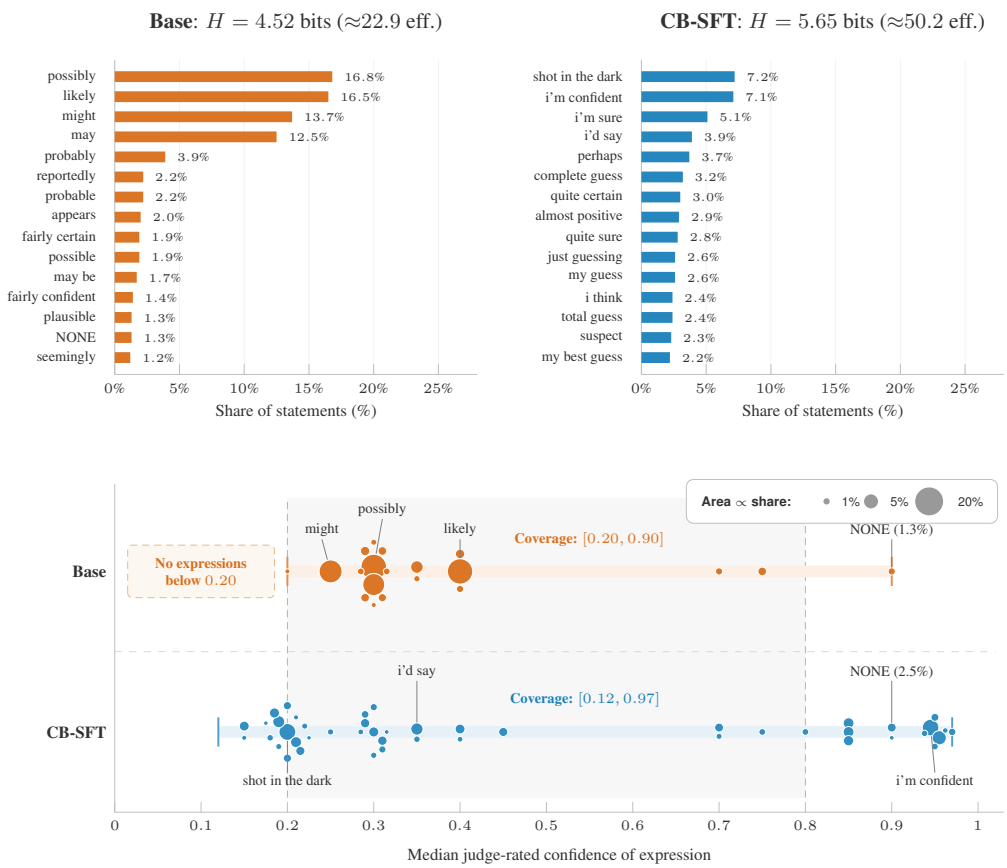


Figure 11: Counterpart of Figure 10 evaluated on Llama-3.1-8B-Instruct on the WIKIEU QA test set, showing Top-15 expression shares on a shared 0–28% axis (top) and confidence coverage across $[0.12, 0.97]$ (bottom).

F TOKEN OVERHEAD ANALYSIS

Table 5 evaluates generation lengths across both model backbones and benchmarks to quantify the computational footprint of structured elicitation.

In terms of substantive content, STRATIFIED (Stripped) yields slightly more concise assertions than BASELINE, although the difference remains modest (e.g., 49 vs. 54–66 tokens on WIKIEU; 118–126 vs. 120–136 tokens on UNCLE).

Because the markup tags are stripped prior to presentation, user-facing readability and cognitive load remain unaffected. Meanwhile, the raw tag decoding footprint ($+79\%$ to $+133\%$ tokens relative to BASELINE) does incur a non-trivial inference cost. However, this decoding overhead directly enables more precise epistemic expression, making it a worthwhile trade-off for high-stakes deployment.

G JUDGE ROBUSTNESS

Table 6 and Table 7 repeat the long-form evaluations with the confidence judge swapped from GPT-OSS-120B to GPT-5.6-Luna under the same protocol. For each result, we randomly sample 500 instances for evaluation. Both models provide very close results, again validating our conclusions.

Table 5: Mean response token counts corresponding to Table 2, measured with native tokenizers across response sets. Stripped denotes STRATIFIED outputs with markup tags stripped; Raw denotes raw tagged generations. Tag overhead measures raw outputs relative to stripped text, while Raw/BASE compares raw outputs against BASELINE.

Backbone	Dataset	Stage	Mean Response Tokens			Overhead (%)	
			BASE	STRAT (stripped)	STRAT (tagged)	Tags	Raw/BASE
Qwen2.5-7B	WIKIEU	SFT	54	49	113	+130	+108
		RL	56	49	115	+132	+104
	UNCLE	SFT	120	123	280	+128	+133
		RL	126	122	277	+126	+119
Llama-3.1-8B	WIKIEU	SFT	61	49	112	+128	+84
		RL	66	51	118	+130	+79
	UNCLE	SFT	124	118	267	+125	+114
		RL	136	126	286	+127	+111

Table 6: Judge-swap validation with GPT-5.6-Luna as the confidence judge, mirroring the layout of Table 2. Evaluations are computed over an identical 500-statement subsample per cell: *oss* = GPT-OSS-120B, *luna* = GPT-5.6-Luna.

Method	WIKIEU						UNCLE					
	AUROC ↑		ECE ↓		Brier ↓		AUROC ↑		ECE ↓		Brier ↓	
	oss	*luna*	*oss*	*luna*	*oss*	*luna*	*oss*	*luna*	*oss*	*luna*	*oss*	*luna*
Qwen2.5-7B-Instruct												
BASELINE												
⌞ SFT	.621	.608	.206	.224	.283	.292	.593	.614	.280	.257	.320	.311
⌞ RL	.709	.705	.091	.100	.213	.217	.802	.798	.235	.224	.231	.228
STRATIFIED												
⌞ SFT	.710	.715	.128	.138	.228	.232	.705	.706	.151	.142	.243	.242
⌞ RL	.818	.796	.067	.071	.167	.174	.791	.794	.160	.162	.210	.210
Llama-3.1-8B-Instruct												
BASELINE												
⌞ SFT	.599	.586	.216	.266	.292	.318	.577	.593	.279	.261	.316	.314
⌞ RL	.699	.714	.086	.085	.189	.190	.727	.734	.190	.190	.243	.243
STRATIFIED												
⌞ SFT	.728	.736	.131	.129	.222	.222	.725	.729	.130	.100	.221	.218
⌞ RL	.749	.736	.124	.110	.189	.186	.765	.755	.195	.193	.227	.228

H STATISTICAL SIGNIFICANCE ANALYSIS

We assess the statistical significance of our findings via paired bootstrap resampling ($B = 2{,}000$) across question IDs. Results for Part I and Part II are shown in Table 8 and Table 9, respectively.

Table 7: Judge-swap validation on HotpotQA with GPT-5.6-Luna as the confidence judge, mirroring the layout of Table 3. Evaluations are computed over an identical 500-statement subsample per cell: *oss* = GPT-OSS-120B, *luna* = GPT-5.6-Luna.

Method	**Difficulty**	**AUROC ↑**		**ECE ↓**		**Brier ↓**	
		oss	*luna*	*oss*	*luna*	*oss*	*luna*
BASELINE	Easy	.711	.711	.283	.282	.292	.289
	Medium	.677	.679	.290	.295	.280	.284
	Hard	.723	.726	.294	.298	.275	.277
STRATIFIED	Easy	.715	.740	.095	.103	.195	.195
	Medium	.691	.727	.088	.079	.143	.140
	Hard	.730	.746	.051	.056	.156	.156

Table 8: Paired-bootstrap significance analysis of Part I findings. Δ represents the AUROC difference between evaluation conditions; 95% percentile CIs and two-sided p-values are computed from B=2,000 resamples paired by question ID.

Contrast	ΔAUROC	95% CI	p
*Position effect (*POST *vs.* PRE*)*			
POSTNUM vs. PRENUM	$+0.079$	$[+0.070, +0.090]$	0.001
POSTVERB vs. PREVERB	$+0.060$	$[+0.051, +0.069]$	0.001
Prompt perturbation (Perturbed vs. Original)			
POSTVERB			
Para-A	-0.046	$[-0.054, -0.037]$	0.001
Para-B	-0.202	$[-0.215, -0.188]$	0.001
Para-C	-0.224	$[-0.237, -0.211]$	0.001
Para-D	-0.100	$[-0.110, -0.090]$	0.001
POSTNUM			
Para-A	-0.006	$[-0.011, -0.002]$	0.009
Para-B	-0.009	$[-0.014, -0.004]$	0.001
Para-C	-0.013	$[-0.018, -0.007]$	0.001
Para-D	-0.021	$[-0.027, -0.016]$	0.001
STRATIFIED NUM.			
Para-A	-0.002	$[-0.008, +0.003]$	0.490
Para-B	-0.012	$[-0.017, -0.007]$	0.001
Para-C	-0.011	$[-0.016, -0.006]$	0.001
Para-D	-0.017	$[-0.022, -0.012]$	0.001
STRATIFIED LING.			
Para-A	-0.016	$[-0.023, -0.009]$	0.001
Para-B	-0.006	$[-0.011, \ 0.000]$	0.044
Para-C	-0.008	$[-0.013, -0.002]$	0.004
Para-D	-0.011	$[-0.017, -0.005]$	0.001

Table 9: Paired-bootstrap significance analysis of Part II findings. Δ denotes metric differences between conditions (negative values indicate improvement for Brier score and ECE); 95% percentile CIs and two-sided p-values are derived from B=2,000 resamples. Contrasts for Table 2 evaluate tag-stripped STRATIFIED outputs against BASELINE across matching prompts.

Contrast	Δ	95% CI	p
*Main comparison: AUROC (*STRATIFIED *vs.* BASELINE*)*			
Qwen			
WIKIEU, SFT	$+0.112$	$[+0.074, +0.149]$	0.001
WIKIEU, RL	$+0.102$	$[+0.070, +0.134]$	0.001
UNCLE, SFT	$+0.109$	$[+0.090, +0.128]$	0.001
UNCLE, RL	$+0.067$	$[+0.052, +0.082]$	0.001
Llama			
WIKIEU, SFT	$+0.125$	$[+0.085, +0.164]$	0.001
WIKIEU, RL	$+0.117$	$[+0.082, +0.149]$	0.001
UNCLE, SFT	$+0.090$	$[+0.070, +0.110]$	0.001
UNCLE, RL	$+0.031$	$[+0.017, +0.045]$	0.001
*Main comparison: Accuracy (*STRATIFIED *vs.* BASELINE*)*			
Qwen			
WIKIEU, SFT	-0.013	$[-0.035, +0.010]$	0.275
WIKIEU, RL	-0.012	$[-0.030, +0.006]$	0.180
UNCLE, SFT	$+0.010$	$[+0.000, +0.020]$	0.040
UNCLE, RL	$+0.015$	$[+0.006, +0.024]$	0.001
Llama			
WIKIEU, SFT	-0.018	$[-0.040, +0.005]$	0.130
WIKIEU, RL	-0.021	$[-0.041, +0.001]$	0.059
UNCLE, SFT	$+0.020$	$[+0.008, +0.032]$	0.001
UNCLE, RL	-0.012	$[-0.022, -0.003]$	0.015
Credit assignment (HG-RL vs. Global-RL, WIKIEU*)*			
AUROC	$+0.043$	$[+0.024, +0.064]$	0.001
Brier	-0.023	$[-0.034, -0.013]$	0.001
ECE	-0.017	$[-0.040, +0.004]$	0.118
Optimization gain (HG-RL vs. SFT init, AUROC)			
Overall	$+0.139$	$[+0.111, +0.170]$	0.001
Balancing objective (CB-SFT vs. Standard SFT)			
Mid-band statement share after RL	$+0.545$	$[+0.515, +0.574]$	0.001
ECE (SFT checkpoint)	-0.072	$[-0.100, -0.041]$	0.001
Accuracy (SFT checkpoint)	-0.004	$[-0.024, +0.016]$	0.728
*HotpotQA transfer (*STRATIFIED *vs.* BASELINE*, Brier)*			
Easy	-0.098	$[-0.133, -0.065]$	0.001
Medium	-0.137	$[-0.169, -0.101]$	0.001
Hard	-0.118	$[-0.149, -0.085]$	0.001

I REINFORCEMENT LEARNING REWARD JUDGES

This section details the judge prompt templates used to supply reward signals during reinforcement learning. Because of the structured markup in STRATIFIED, individual factual statements can be extracted via direct tag parsing rather than through an auxiliary judge model. However, an alignment reward is required to ensure that verbalized hedges faithfully reflect the underlying epistemic state. At evaluation time, all markup tags are stripped, and generations are assessed through the exact same evaluation pipeline as BASELINE to guarantee a strictly fair comparison.

BASELINE Statement Correctness Judge

You are an objective grading assistant. Your task is to evaluate whether a model-generated paragraph correctly answers a specific question.

Evaluation Protocol:

1. **Extract the Relevant Statement:**
   - Read the paragraph and locate the sentence or clause that addresses the question.
   - Copy the extracted statement EXACTLY as it appears in the paragraph.
   - If the paragraph does not address the question at all, set "answer" to "missing" and "is_correct" to "missing".
2. **Compare with Ground Truth:**
   - The extracted statement is CORRECT ("yes") if it clearly states the ground truth answer (semantic equivalence).
   - The extracted statement is INCORRECT ("no") if it states a contradictory or incorrect answer.

Data to Evaluate:

- **Question:** "`{question}`"
- **Ground Truth:** `{ground_truth}`
- **Paragraph:** "`{paragraph}`"

Output format (JSON ONLY, no other text):

```
{"answer": "extracted statement here", "is_correct": "yes"}
```

STRATIFIED Statement Correctness Judge

You are given one ground-truth fact (subject, predicate, object) and candidate fact statements from a model's structured response. Each candidate is one line of a structured tag list.

1. Find the candidate whose statement DIRECTLY asserts the ground-truth SUBJECT and PREDICATE: it states what the subject's value for that specific property is. Do NOT pick a candidate that only provides a broader category or a different property. If several candidates discuss it, pick the most specific one. If NO candidate discusses the subject+predicate, set fact_index to null.
2. Judge the chosen candidate's statement:
   - "yes" if it states the correct object.
   - "no" if it states a different object.

Hedging expressions do NOT affect factual correctness.

Data to Evaluate:

- **Triplet:** subject="`{subject}`", predicate="`{predicate}`", object="`{object}`"
- **Candidate facts:**
  `{facts_list}`

Output Format (JSON ONLY, no other text):

```
{"fact_index": 0, "is_correct": "yes"}
```

Hedge-to-Confidence Judge

You are an expert evaluator of natural language uncertainty expressions (hedge words).
Analyze the provided text and infer what confidence level the hedge words imply.

Evaluation Guidelines:

- Focus ONLY on the uncertainty/hedging language: NOT on whether the statement is factually correct.
- If the input contains both a factual claim and an explicit certainty statement, focus on the sentence or clause expressing confidence.
- A direct factual statement without any hedge markers implies confidence ≈ 0.95.
- If the text includes explicit percentage numbers (e.g., "I am 80% sure"), implied confidence is ≈ 0.80.
- If the model states it is guessing randomly among N options, implied confidence is $\approx 1/N$.
- If the statement is missing, empty, or unreadable, implied confidence is 0.0.

Confidence Inference Mapping:

Hedge Level	Example Phrases	Confidence
No hedging (certain)	"The answer is X", "I am certain", "There is no doubt"	0.90–1.00
Light hedging	"I'm fairly confident", "almost certain", "should be X"	0.75–0.90
Moderate hedging	"I think", "probably", "likely", "I'm not entirely sure"	0.50–0.75
Strong hedging	"might be", "could be", "not very confident", "unsure"	0.25–0.50
Very strong / guess	"no idea", "wild guess", "just guessing", "shot in the dark"	0.05–0.25

Data to Evaluate:

- **Text:** "{text}"

Output Format (JSON ONLY, no other text):

{"implied_confidence": 0.85}

STRATIFIED Verbalization Alignment Judge

You are an expert evaluator of natural language responses. Your task is to evaluate how well a model's free-text response aligns with its stated answer and confidence level.

Evaluation Criteria:

1. **Answer Presence:** Verify whether the response conveys the core factual claim of the stated answer. Score 0 if omitted or contradicted.
2. **Hedge-Confidence Alignment:** Verify whether verbal hedges accurately reflect the stated numeric confidence of {confidence_pct}% according to standard mapping rubrics.
3. **Overall Alignment Score (0.0–1.0):**
   - 1.0: Answer is clearly present AND hedge words match stated confidence.
   - 0.7–0.9: Answer present; hedge words slightly deviate from stated confidence.
   - 0.4–0.6: Answer present but hedging significantly mismatches stated confidence.
   - 0.1–0.3: Answer missing or wrong, or hedge words severely contradict stated confidence.
   - 0.0: Answer missing AND hedge words completely contradictory.

Data to Evaluate:

- **Stated Answer:** "{extracted_answer}"
- **Stated Confidence:** {confidence_pct}%
- **Model Response:** "{response_text}"

Output Format (JSON ONLY, no other text):

```
{"answer_in_response": "yes", "alignment_score": 0.85}
```

Response Quality Judge

Score the writing quality of this response from 0.0 to 1.0.

Lower the score when:

- The same claim is stated more than twice.
- Sentences do not form coherent, connected prose.
- The response contradicts itself.

Hedging expressions never lower the quality score. Residual tag fragments are formatting artifacts and do not affect the score.

Examples:
Good (1.0): "I think Bergisel is a mountain in Austria. It might be best known for ski jumping."
Bad (0.0): "Bergisel is a mountain in Austria. I'm certain Bergisel is a mountain. Bergisel is definitely in Austria."

Data to Evaluate:

- **Response:** "{response}"

Output Format (JSON ONLY, no other text):

```
{"quality_score": 0.75, "reason": "brief"}
```

J CONFIDENCE MAPPING

To construct SFT annotations for WIKIEU, knowledge triplets are paired with target confidence tiers and mapped to diverse linguistic hedges. Table 10 presents the complete 10-tier Confidence Expression Mapping. To maximize lexical diversity and prevent position or frequency bias during generation, candidate phrases within each tier are shuffled randomly before prompt formatting.

The prompt template below injects this mapping during teacher-model generation to synthesize fluent narrative responses from discrete fact-confidence pairs.

WIKIEU Response Synthesis Prompt

You are an expert data generation assistant. Write the assistant's answer to the user's question below.

The answer as a whole must be a natural, fluent conversational response: not a list, not a stiff concatenation of facts. Weave ALL provided facts into the narrative:

- Each fact has an assigned confidence level (1.0 = certain, 0.1 = total guess). Express each fact using a phrase from the Confidence Expression Mapping that matches its level; never leave a fact out and never invent hedges of your own.
- Phrase facts naturally instead of echoing raw attribute labels.
- Do not use the level number itself or the word "confidence".

Data to Ingest:

- **User Question:** {question}
- **Facts with Assigned Confidence:** {facts}
- **Expression Mapping Table:** {mapping_tab}

Output Constraint:

Output ONLY the answer text itself: no JSON, no code fences, no commentary.

Confidence	Approved Linguistic Hedging Expressions
1.0	State the fact directly as a plain statement with no hedging phrase; "I am completely certain that…"; "I know for a fact that…"; "Without a doubt, …"; "I'm certain that…"; "I have no doubt that…"; "I'm positive that…"; "I'm virtually certain that…"
0.9	"I'm all but sure that…"; "I'm sure that…"; "I'm almost positive that…"; "I'm quite certain that…"; "There's little doubt that…"; "I'm convinced that…"
0.8	"I'm quite sure that…"; "I'm confident that…"; "I'm pretty sure that…"
0.7	"I'm fairly sure that…"; "I'd bet that…"; "I have a strong feeling that…"; "Most likely, …"; "It's likely that…"; "It's probable that…"; "I'm fairly confident that…"; "Probably …"; "There's a good chance that…"; "Chances are that…"; "I believe that…"; "I expect that…"; "As far as I know, …"
0.6	"I'd say that…"; "I think that…"; "It seems like…"; "My gut tells me that…"; "The way I see it, …"
0.5	"I suspect that…"; "I wouldn't be surprised if…"; "It can be …"; "I don't have much to go on, but…"; "I'd imagine …"; "Something tells me that…"
0.4	"It's entirely possible that…"; "My best guess is that…"; "It may be that…"; "Maybe …"; "Perhaps …"
0.3	"It might be that…"; "It could be that…"; "There's a chance that…"; "I can see a scenario where…"; "There's a possibility that…"; "If I had to guess, …"; "My guess is that…"
0.2	"There's a small chance that…"; "There's a slim chance that…"; "There's an outside chance that…"; "It's a long shot, but…"; "I'm just guessing here, but…"; "It's a wild guess, but…"
0.1	"This is a total shot in the dark, but…"; "This is a complete guess, but…"; "I'm completely guessing here, but…"; "This is a total guess, but…"

Table 10: Full confidence expression mapping projecting confidence tiers (0.1–1.0) to linguistic hedging phrases for WIKIEU synthesis.