RETRIEVAL-AUGMENTED LINGUISTIC CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Linguistic cues such as "I believe" and "probably" offer an intuitive interface for communicating confidence, yet a generalisable, principled calibration framework for linguistic confidence expressions remains underexplored. In particular, co-occurring linguistic cues, contextual variation, and subjective audience interpretation pose unique challenges. We therefore model linguistic confidence as a distribution over plausible perceived probability values that a statement is correct, capturing interpretation variability that scalar representations discard. Within this distributional framework, we introduce faithfulness as a complementary evaluation dimension and present Faithfulness Divergence (FD), an information-theoretic metric quantifying the surprise induced in audience beliefs upon truth revelation. Building on these foundations, we present Retrieval-Augmented Linguistic Calibration (RALC), a lightweight post-hoc pipeline that propagates calibrated confidence signals back into natural language via retrieval-augmented rewriting. Across three QA benchmarks and five LLM families, RALC improves in-domain faithfulness and calibration up to 66% and 58%, respectively, outperforming black-box and grey-box calibration baselines.

1 INTRODUCTION

Reliable confidence estimation is fundamental to the trustworthy deployment of large language models (LLMs) in human decision-making pipelines (Steyvers et al., 2025). Without well-calibrated confidence signals, users risk over-relying on model outputs that hallucinate or fail silently (Kim et al., 2024), underscoring the need for confidence frameworks that are both scientifically rigorous and interpretable by human users.

Existing confidence estimation methods represent confidence as scalar probability values, including token-level probability (Lamb et al., 2025), semantic uncertainty (Farquhar et al., 2024), and verbalised scores (Lin et al., 2022a; Tian et al., 2023). However, humans struggle to reason accurately with numerical probabilities (Zhang & Maloney, 2012), motivating the use of linguistic markers such as "may" or "likely" as more natural confidence interfaces. Prior work demonstrates that such markers retain evaluative signal (Yona et al., 2024; Tao et al., 2025a); however, treating them as scalars discards the inherent subjectivity of linguistic interpretation: different readers map the same expression to different perceived probability values (Tao et al., 2025b).

We address this gap by modelling linguistic confidence as a distribution over plausible perceived probability values that a statement is correct, where perception arises from readers' interpretations of the full linguistic content of the statement rather than from a discrete mapping of individual vocabulary items. Treating linguistic confidence as a surrogate for statement correctness induces a binary classification view, in which confidence corresponds to the predictive probability of the true class. Drawing a parallel to evidential deep learning, which models class probabilities with Dirichlet distributions to capture second-order predictive uncertainty (Sensoy et al., 2018), we formalise linguistic confidence for binary correctness using its natural binary special case: a Beta distribution over perceived confidence scores that a statement is correct, where the mean captures the central tendency of perceived confidence across readers and the concentration encodes the strength of agreement.

The standard measure of confidence quality is calibration, assessed through population-level expected calibration error (ECE) (Guo et al., 2017; Wang et al., 2024), quantifying the alignment between confidence and accuracy in expectation. Instance-level metrics such as the Brier score (Brier, 1950)

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.

and negative log likelihood offer pointwise assessment in the classical scalar setting, yet their distributional generalisations still fail to encode variance as the strength of agreement by readers. We therefore introduce *faithfulness* as a complementary dimension of confidence evaluation and present Faithfulness Divergence (FD), a concentration-weighted Bayesian updating cost that quantifies the information-theoretic surprise to confidence beliefs upon truth revelation.

Calibration in linguistic space remains crucial yet largely unsolved. Classical post-hoc calibration methods, including temperature scaling (Guo et al., 2017), Platt scaling (Platt, 1999), histogram binning (Zadrozny & Elkan, 2001), isotonic regression (Zadrozny & Elkan, 2002), Beta calibration (Kull et al., 2017), and distribution-matching approaches (Song et al., 2019; Marx et al., 2023), operate entirely in numerical space and provide no mechanism for propagating calibrated signals back into language. Prompt-conditioned hedging strategies offer a linguistic alternative, yet function as black-box procedures with no principled control over the output (Yona et al., 2024). The closest related work performs discrete hedging word confidence profiling and remapping at the word level in a specialised domain (Wang et al., 2024), overlooking the co-occurrence of multiple linguistic cues within a statement and their contextual interactions. Consequently, a generalisable, continuous, and lightweight post-hoc framework that provides principled guidance for hedging expressions remains underexplored. To address this gap, we introduce Retrieval-Augmented Linguistic Calibration (RALC), a post-hoc pipeline that operates directly in linguistic space to transform raw LLM responses into calibrated and faithful outputs. The pipeline applies Platt scaling (Platt, 1999) on confidence distribution means whilst preserving distributional concentration, and propagates calibrated distributions to language through retrieval-augmented LLM rewriting, employing the retrieval-augmented generation paradigm (Lewis et al., 2020). Furthermore, RALC is compatible with diverse upstream confidence signals beyond linguistic confidence, including token probability and semantic uncertainty.

Our contributions are as follows:

1. We formalise linguistic confidence as a distribution over plausible perceived probability values that a statement is correct, capturing the interplay of linguistic cues and contexts beyond discrete expression mapping and scalar quantification.
2. We introduce *faithfulness* as a new dimension of confidence evaluation and present Faithfulness Divergence (FD), an instance-level metric that quantifies, as information-theoretic surprise, how faithfully a confidence distribution accounts for the ground-truth correctness outcome.
3. We introduce a generalisable retrieval-augmented linguistic confidence calibration pipeline that effectively improves faithfulness and calibration in linguistic space and is compatible with diverse confidence estimation signals.

We evaluate the framework across multiple LLM families and QA benchmarks, including MMLU (Hendrycks et al., 2020), SQuAD 2.0 (Rajpurkar et al., 2018), and TruthfulQA (Lin et al., 2022b). Results demonstrate near-lossless information transfer through the calibration pipeline and substantial improvements in both calibration and faithfulness across models and benchmarks, outperforming the prompt-based calibration baselines.

2 Related work

LLM confidence estimation Existing confidence estimation methods predominantly represent confidence as scalar probability values, including token-level probability aggregation (Lamb et al., 2025; Duan et al., 2024) and consistency-based approaches that infer confidence from the semantic support landscape across repeated samples (Wang et al., 2022; Farquhar et al., 2024). Whilst recent work demonstrates that linguistic cues in model responses preserve evaluative signal as confidence surrogates (Yona et al., 2024; Tao et al., 2025b), their scalar quantification overlooks the inherently subjective nature of linguistic interpretation. Wang et al. (2024) take a step towards distributional representations by mapping individual discrete hedging words to confidence distributions; however, their approach targets word-level remapping rather than statement-level confidence, where multiple linguistic cues co-occur and interact with context. Huang et al. (2024) jointly model confidence and correctness as distributions over ambiguous long-form generation contexts, which is orthogonal to our objective of binary classification with predictive probability distributions.

Confidence evaluation Expected calibration error (ECE) is the dominant metric for evaluating confidence, measuring alignment between scalar confidence scores and accuracy in both classical (Guo et al., 2017) and language model settings (Zhu et al., 2023). Extensions based on entropy (Sumler et al., 2025), variance (Thompson & Desai, 2026), and distributional generalisation (Wang et al., 2024) remain ECE-based and rely on local aggregation, discarding full distributional information at the instance level. Instance-level scoring such as the Brier score (Brier, 1950) and negative log likelihood similarly do not capture variance-scaled misalignment.

Confidence calibration Classical post-hoc calibration methods, including Platt scaling (Platt, 1999), histogram binning (Zadrozny & Elkan, 2001), isotonic regression (Zadrozny & Elkan, 2002), and Beta calibration (Kull et al., 2017), adjust scalar outputs towards empirical accuracy but are confined to numerical space. Distributional calibration methods frame the problem as distribution matching, aligning predicted confidence distributions with empirical label distributions through various mapping strategies (Song et al., 2019; Marx et al., 2023), though they target global rather than instance-level calibration. In the linguistic space, prompt-based strategies have been explored to steer LLM hedging (Yona et al., 2024), but these lack principled control. Internal model steering offers finer-grained calibration of verbal uncertainty (Ji et al., 2025), yet requires access to model internals, limiting applicability to open-source settings. The most closely related approach remaps discrete hedging words at the vocabulary level, without accounting for contextual interactions or producing calibrated full responses (Wang et al., 2024).

3 Confidence estimation and evaluation

3.1 Linguistic confidence estimation

For each input–response pair (X, R), let $y \in \{0, 1\}$ denote the correctness label of R. We define a distributional confidence estimator

$$g : \mathcal{R} \to \mathcal{P}([0, 1]),$$

where $\mathcal{R}$ denotes the space of model responses and $\mathcal{P}([0, 1])$ denotes the space of probability distributions over $[0, 1]$. The estimator g models the plausible probability values that R is correct as a distribution over confidence scores in $[0, 1]$ as perceived by readers (human or model-based evaluators). We abstract g as a model-based or human-based evaluator and parameterise the estimated distribution S as a Beta distribution, $S = \mathrm{Beta}(\alpha, \beta)$. This choice draws a parallel to evidential deep learning (Sensoy et al., 2018), which places a Dirichlet prior over class probabilities to represent second-order uncertainty. Our setting resembles binary classification: each reader produces an interpreted confidence score viewable as a draw of the true-class probability; the Beta distribution is therefore the principled choice, as the binary special case of the Dirichlet and the natural conjugate prior for the Bernoulli likelihood. The mean $\alpha/(\alpha + \beta)$ captures the central tendency of perceived confidence across readers, whilst the concentration $(\alpha + \beta)$ encodes agreement strength: a high mean with low concentration signals inconsistent reader interpretations, whereas the same mean with higher concentration signals consistent agreement.

3.2 Confidence evaluation

Calibration as a dimension of confidence evaluation At the population level, calibration requires confidence to match empirical accuracy in expectation. Letting $p \sim S$ denote the scalar confidence value drawn from the estimated distribution, the classical measure is the expected calibration error,

$$\mathrm{ECE} = \mathbb{E}[|\mathbb{E}[Y \mid p] - p|]\,,$$

with realisations including scalar bin-based (Guo et al., 2017) and distribution-generalised variants (Wang et al., 2024).

Faithfulness as a dimension of confidence evaluation Calibration is necessary but not sufficient in the distributional setting: two predictors may achieve similar average calibration yet convey markedly different instance-level confidence profiles. We therefore introduce *faithfulness*, a human-aligned, instance-level dimension of confidence evaluation.

A confidence distribution is faithful when observing the ground truth induces little surprise relative to prior beliefs. Such surprise is driven by both central-tendency misalignment and the strength of

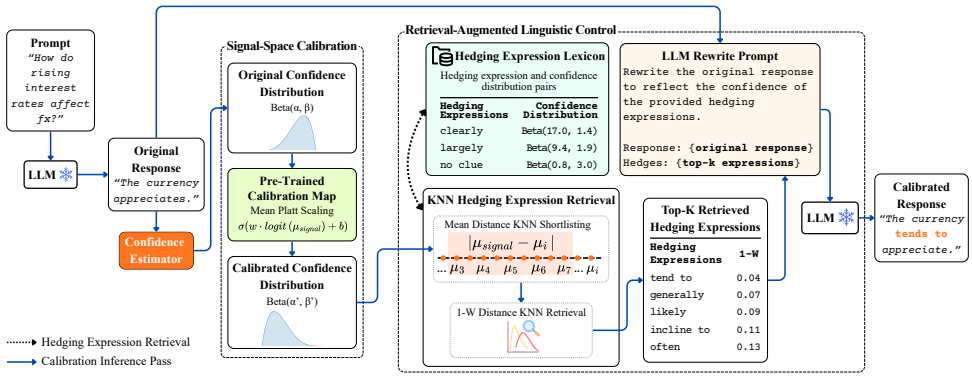

Figure 1: Retrieval-Augmented Linguistic Calibration pipeline overview. In each calibration inference pass (**blue arrow** →), we estimate a confidence distribution for the original response using linguistic confidence, token probability, or semantic uncertainty (Sections 3.1, 4.3). In the signal space, we apply a pre-trained Platt scaling calibration map on the mean to correct miscalibration in the numerical space (Section 4.1). The calibrated distribution is then used as a retrieval signal to find the nearest hedging expressions from a pre-built hedge-confidence-pair lexicon. The KNN retrieval process uses absolute distance in the means for shortlisting and 1-Wasserstein distance for final retrieval to ensure alignment in both central tendency and spread. The retrieved hedging expressions form a rewrite prompt along with the original response, which is passed to an LLM editor to produce a linguistically calibrated response (Section 4.2).

agreement with which that misalignment is held. Grounding this intuition in information theory, we draw on Bayesian surprise (Itti & Baldi, 2009), measured by the KL divergence between posterior and prior distributions, and weight it by the concentration $(\alpha_i + \beta_i)$ as the effective sample size of the prior (Morita et al., 2008) to represent the total surprise of the update.

Formally, for instance i, we update the prior S_i with a single Bernoulli observation of the correctness label y_i to obtain the posterior S_i^*:

$$S_i = \mathrm{Beta}(\alpha_i,\ \beta_i), \qquad S_i^* = \mathrm{Beta}(\alpha_i + y_i,\ \beta_i + 1 - y_i).$$

The *Faithfulness Divergence* (FD) for instance i is then

$$\mathrm{FD}_i := (\alpha_i + \beta_i) \cdot \mathrm{KL}(S_i^* \,\|\, S_i)\,.$$

The KL term quantifies the information-theoretic change of belief required upon observing the outcome; weighting by $(\alpha_i + \beta_i)$ ensures the significance of the change is properly accounted for. FD is non-negative and intended as a *relative* instance-level metric: lower values indicate more faithful confidence communication, whilst higher values signal greater mismatch between expressed confidence and realised correctness. We provide a further discussion of the modelling in Appendix B.

4 Retrieval-Augmented Linguistic Calibration (RALC)

We introduce a novel post-hoc calibration pipeline that operates directly in the linguistic space to transform original LLM responses into calibrated and faithful outputs, Retrieval-Augmented Linguistic Calibration (RALC). RALC transforms a raw response into one whose perceived confidence more faithfully reflects the underlying ground-truth label. Figure 1 illustrates the pipeline mechanics. It estimates a confidence distribution for the original responses, applies post-hoc signal-space calibration to correct miscalibration in the numerical space, and then uses the calibrated distribution as a retrieval signal to search for appropriate hedging expressions that guide the rewriting of the original response into a linguistically calibrated and faithful one.

4.1 Post-hoc signal-space calibration

Definition Under the distributional setting, the confidence estimator g outputs a distribution over confidence scores. Post-hoc calibration learns a mapping t in distribution space so that the calibrated estimator $t \circ g$ reduces expected calibration error and Faithfulness Divergence.

Platt scaling on distribution means We parameterise confidence with a Beta distribution $S = \mathrm{Beta}(\alpha, \beta)$, with mean $\mu = \alpha/(\alpha + \beta)$ and concentration $\kappa = \alpha + \beta$. We apply Platt scaling (Platt, 1999) on the distribution mean, mapping μ to a calibrated mean μ' by fitting a logistic regression of distribution means against binary correctness labels:

$$\mu' = \sigma(w \cdot \mathrm{logit}(\mu) + b)\,,$$

where $\sigma(\cdot)$ denotes the sigmoid function and $(w, b) \in \mathbb{R}^2$ are learned scalar parameters. The concentration $\kappa = \alpha + \beta$ is preserved from the original estimated distribution due to the non-existence of a natural target for concentration calibration.

We reconstruct the calibrated Beta distribution from the calibrated mean μ' and the preserved concentration κ, setting $\alpha' = \mu' \cdot \kappa$ and $\beta' = (1 - \mu') \cdot \kappa$, yielding the calibrated distribution $t(S) = \mathrm{Beta}(\alpha', \beta')$. We also investigate classical alternatives, including temperature scaling (Guo et al., 2017), isotonic regression (Zadrozny & Elkan, 2002), and histogram binning (Zadrozny & Elkan, 2001), and find that Platt scaling is the consistent outperformer in improving calibration and faithfulness. We present an ablation study in Appendix D.7.

Interpretation The signal-space Platt scaling corrects systematic mean misalignment, whilst preserving the concentration that encodes the strength of agreement across readers. The mean is directly supervised by the correctness labels and scaling induces a more faithful and calibrated belief of the correctness outcome. By contrast, the concentration is a linguistic feature that emerges from the interplay of linguistic cues and reader interpretation, and does not have a theoretical target for calibration. As a result, we deliberately confine our calibration design to one mean scaling where a natural target exists.

4.2 Retrieval-augmented linguistic control

Post-hoc signal-space calibration updates the originally estimated confidence distributions but leaves the response language unchanged. To close this gap, we introduce retrieval-augmented linguistic control, which rewrites the original response R into a revised response R' whose perceived confidence aligns with the calibrated signal $S' = t(S)$. Formally, a linguistic calibrator l produces $R' = l(R, S')$ such that re-estimating confidence from R' recovers S'. The full pipeline $l \circ t \circ g$ thus forms a closed loop from raw response to calibrated confidence to linguistically calibrated response.

Linguistic confidence lexicon The retrieval step relies on a lexicon that maps hedging expressions to confidence distributions. Hedging expressions are sourced from state-of-the-art LLMs, including Claude-Sonnet-4.6 (Anthropic, 2026), GPT-5.4 (OpenAI, 2026), and Gemini-3-Flash (Google DeepMind, 2025b). For each hedging expression w_k, GPT-OSS-20B (Agarwal et al., 2025) rewrites a collection of non-verifiable statements to incorporate that expression. The LLM linguistic evaluator ensemble then independently evaluates the perceived confidence of each rewritten statement in 3 model passes per evaluator, producing a set of confidence scores, as outlined in Section 5.1. Fitting a Beta distribution to all confidence scores across passes yields the pair $\big(w_k,\, \mathrm{Beta}(\alpha_k, \beta_k)\big)$. Repeating this procedure across all hedging expressions produces the lexicon $\{(w_k,\, \mathrm{Beta}(\alpha_k, \beta_k))\}_{k=1}^{K}$ used at inference time. Additional details on the lexicon construction are provided in Appendix D.1.

Retrieval-based rewriting Given a calibrated signal S', we retrieve the k nearest hedging expressions from the lexicon via a two-stage process. First, we shortlist candidates by mean distance $|\mu_k - \mu_{S'}|$, retaining expressions whose distributional mean falls within a neighbourhood of $\mu_{S'}$. Second, we rank the shortlisted candidates by the 1-Wasserstein distance via Monte Carlo estimation,

$$d(w_k,\, S') = W_1(\mathrm{Beta}(\alpha_k, \beta_k),\, S')\,,$$

and select the top-k nearest expressions. Mean-distance shortlisting efficiently narrows the candidate set, whilst the subsequent W_1 ranking captures full distributional shape, matching both the central tendency and spread of S', at lower computational cost than applying W_1 over the entire lexicon. The retrieved expressions are then passed alongside R to an LLM editor, which rewrites R into R' to match the target confidence profile, enforcing $g(R') \approx S'$ in practice.

Table 1: Instance-level metric comparison across controlled subsets varying in mean confidence, concentration, and accuracy.

Subset	Acc.	Avg Conf.	Conc.	FD↓	KL	$\mathbb{E}$[Brier]	$\mathbb{E}$[NLL]
(1) high conf., high conc., wrong	0.0	0.812	25.8	2.932	0.168	0.681	2.052
(2) low conf., high conc., right	1.0	0.448	6.0	0.550	0.114	0.348	0.953
(3) high conf., low conc., wrong	0.0	0.656	1.0	0.486	0.590	0.557	3.832
(4) low conf., low conc., right	1.0	0.433	1.0	0.392	0.378	0.442	1.719

4.3 Alternative confidence signals

The RALC pipeline is compatible with confidence signals beyond linguistic confidence, including token probability (Lamb et al., 2025) and semantic uncertainty (Farquhar et al., 2024), serving as the post-hoc signal-space calibration object and retrieval signal to guide the calibration process. We formulate both distributional confidence signals under self-consistency sampling (Wang et al., 2022) and cluster the generated responses semantically.

Length-normalised token probability Let $\mathcal{I}_{\max}$ denote the index set of the largest cluster. For each $R_j \in \mathcal{I}_{\max}$, where r_i is the i-th token and $r_{<i}$ the preceding context, we compute its length-normalised token probability (Lamb et al., 2025):

$$s_j^{\text{tok}} = \exp\left(\frac{1}{|R_j|}\sum_{i=1}^{|R_j|}\log p_\theta(r_i \mid r_{<i}, X)\right) \in [0,1].$$

Fitting a Beta distribution to $\{s_j^{\text{tok}}\}$ via method of moments (Appendix A.1.1) yields S^{tok} that represents confidence as a distribution of token-level probabilities for a particular semantic meaning.

Semantic uncertainty Let $\mathcal{I}_{\max}$ denote the index set of the largest cluster and N be the total number of sampled responses. For each self-consistency sample for a given input, the Beta parameters are set directly from cluster counts,

$$\alpha^{\text{sem}} = |\mathcal{I}_{\max}|, \qquad \beta^{\text{sem}} = N - |\mathcal{I}_{\max}|,$$

Both parameters are clipped to a minimum of 10^{-6} to handle degenerate cases (e.g. all responses falling into a single cluster, which would set $\beta^{\text{sem}} = 0$), yielding $S^{\text{sem}} = \text{Beta}(\max(\alpha^{\text{sem}}, 10^{-6}), \max(\beta^{\text{sem}}, 10^{-6}))$ that represents confidence as a distribution of semantic support for a particular semantic meaning across samples.

5 Experiments

5.1 Setup preliminaries

We evaluate across five open-source language models from different families: GPT-OSS-20B (Agarwal et al., 2025), Llama-3.1-8B-Instruct (Meta AI, 2024), Qwen3-8B (Yang et al., 2025), Mistral-7B-Instruct-v0.3 (Mistral AI Team, 2023), and Gemma-4-31B-IT (Google DeepMind, 2025a), on three benchmarks: MMLU (Hendrycks et al., 2020), SQuAD 2.0 (Rajpurkar et al., 2018), and TruthfulQA (Lin et al., 2022b), covering reasoning-heavy multiple-choice, reading comprehension, and closed-book short-answer question-answering formats. We elicit responses using the Direct QA and Hedged QA prompt templates based on Yona et al. (2024)'s work. Direct QA aims to produce natural, succinct free-form responses, whilst Hedged QA additionally instructs the model to express uncertainty through hedging language and serves as a black-box baseline for RALC. The full templates are in Appendix C.1. We employ GPT-OSS-20B as a model-based grader for correctness and the rewriting model in our RALC pipeline. To estimate linguistic confidence as perceived by readers, we construct an evaluator ensemble of three LLMs (Qwen3-8B, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3) as a human audience surrogate, capturing potential co-occurrences of linguistic cues and their contextual interactions; each model independently evaluates the confidence expressed in a response three times. A Beta distribution is fitted to these scores via method of moments (Appendix A.1.1). We further validate our LLM ensemble against the human-annotated linguistic benchmark of Tao et al. (2025b) (Appendix C.2.2).

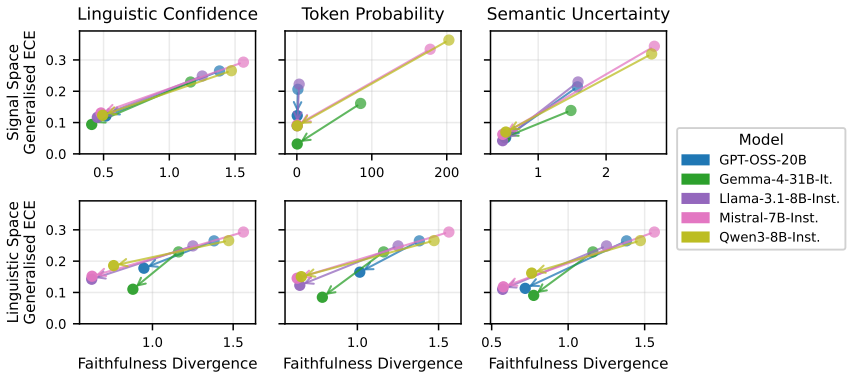

Figure 2: Pre-calibration→post-calibration changes in generalised ECE and Faithfulness Divergence across signal space (**top**) and linguistic space (**bottom**), averaged across MMLU, SQuAD 2.0, and TruthfulQA. Our RALC consistently improves (reduces) both metrics across all confidence signals in both spaces.

5.2 Measuring faithfulness under distributional confidence

To validate the quantification of the *surprise after truth revelation*, we construct four controlled subsets from the linguistic confidence distributions of Llama-3.1-8B-Instruct on SQuAD 2.0, spanning the corners of the distributional confidence space by varying mean confidence, concentration, and accuracy. We compare FD against KL divergence (Kullback & Leibler, 1951), $\mathbb{E}[\text{Brier}]$ (Brier, 1950), and $\mathbb{E}[\text{NLL}]$. Whilst the Brier score and NLL are commonly used in scalar settings, we provide their distributional generalisations in Appendix A.2.

As shown in Table 1, FD is the only metric that recovers the expected surprise ordering: concentrated, misaligned distributions receive the highest penalties (surprise), whilst the same misalignment expressed with low concentration receives lower penalties. KL divergence inverts this ranking by penalising diffuse distributions more heavily regardless of mean misalignment due to a lack of weighting by the effective sample size. $\mathbb{E}[\text{Brier}]$ and $\mathbb{E}[\text{NLL}]$ also fail to recover the expected surprise ordering, as neither encodes variance as an amplifier or mediator of surprise. These results confirm that FD uniquely captures both mean misalignment and dispersion, making it the appropriate instance-level faithfulness metric in the distributional confidence setting under our information-theoretic modelling. In addition to the empirical validation, we provide further theoretical ablation studies on FD in Appendix B.4.

5.3 Retrieval-Augmented Linguistic Calibration (RALC)

5.3.1 In-domain calibration

Signal-space and linguistic-space calibration We evaluate RALC's ability to improve calibration and faithfulness across both signal and linguistic spaces. For each question, we generate 20 responses under self-consistency sampling, cluster them semantically, and identify the largest cluster. The first response in the majority cluster is selected as the original response for linguistic calibration. Linguistic confidence is estimated by our LLM ensemble on this response; token probability and semantic uncertainty distributions are constructed from the cluster according to Section 4.3. We provide the LLM-based clustering prompt in Appendix C.4.

For each confidence signal, we train a Platt scaling calibration map on the first 30% of each dataset, regressing per-response distribution means against binary correctness labels $y \in \{0, 1\}$. We then run RALC inference on the remaining 70% held-out set, applying the pre-trained Platt scaling map to the estimated confidence distributions to obtain calibrated distributions. The calibrated distributions are passed to the retrieval-augmented linguistic control module, which selects the top k=5 from the top 30 mean-based shortlisted hedging expressions to rewrite the original response into a linguistically calibrated one. We provide the LLM rewriting prompt in Appendix D.3 and the choice of k ablation study in Appendix D.5.

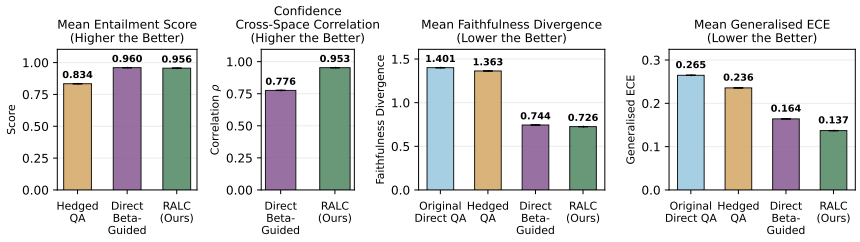

Figure 3: Calibration effectiveness and quality comparison between RALC (averaged across all signals and models), Hedged QA, and Direct Beta-Guided Rewrite across content preservation (entailment), signal-to-language confidence correlation (ρ), linguistic-space Faithfulness Divergence, and linguistic-space generalised ECE. RALC matches Direct Beta-Guided Rewrite on content preservation, achieves a markedly higher signal-to-linguistic-space correlation, and outperforms both baselines on calibration and faithfulness.

We evaluate pre- and post-calibration in the signal space by comparing the Faithfulness Divergence and generalised ECE (Wang et al., 2024) of the confidence distributions before and after the calibration map. In the linguistic space, we estimate the linguistic confidence for both the original and linguistically calibrated responses using the LLM ensemble and evaluate using the same metrics. As shown in Figure 2, RALC consistently improves both metrics across all three confidence signals in both spaces. Additionally, across all signals, semantic uncertainty yields the strongest improvements in calibration and faithfulness as shown in Table 2. Additional results are detailed in Appendix E.

Benchmark against calibration baselines To contextualise RALC's linguistic calibration quality, we benchmark it against two baselines: **Hedged QA**, a black-box baseline that prompts the model to hedge without access to any calibrated signal (Yona et al., 2024), and **Direct Beta-Guided Rewrite**, a grey-box ablation of our pipeline in which the lexicon retrieval step is removed and the calibrated Beta distribution is passed directly to the rewriting model, relying on it to select appropriate hedging language without explicit linguistic grounding. This comparison isolates the contribution of structured lexicon retrieval over uncontrolled and partially controlled generation.

All performance metrics are measured in the linguistic space. We first apply our LLM ensemble to estimate the linguistic confidence expressed in the original uncalibrated Direct QA responses. For each calibration method, we then re-estimate linguistic confidence from the rewritten output and compute all metrics accordingly. Following Farquhar et al. (2024)'s entailment setup, content preservation is scored as $p_{\text{entail}} + 0.5 \cdot p_{\text{neutral}}$ using DeBERTa-v3-Large-MNLI, and evaluated alongside signal-to-language confidence correlation (ρ), Faithfulness Divergence, and generalised ECE. As shown in Figure 3, RALC substantially outperforms Hedged QA across all metrics. Against Direct Beta-Guided Rewrite, RALC matches on content preservation whilst achieving lower Faithfulness Divergence and ECE, and attains a markedly higher Spearman ρ between calibrated signal-space and re-estimated linguistic-space confidence, indicating that structured lexicon retrieval propagates the calibrated signal into language more reliably than unconstrained hedging selection.

5.3.2 Cross-domain confidence calibration

Having established strong in-domain performance, we examine the cross-domain transferability of confidence signals through the lens of RALC. The assessment is independent of the performance of RALC; rather, it uses RALC as a diagnostic framework to evaluate signal stability under domain shift, specifically the reliability of each confidence signal as an input to the pipeline when the Platt scaling map trained on one dataset is applied to a different domain without retraining.

We train the Platt scaling calibration map on each of the three datasets in turn and evaluate the resulting RALC pipeline on all three datasets without retraining, yielding both in-domain (diagonal entries) and cross-domain (non-diagonal entries) conditions. We report pre-to-post-RALC percentage reductions in linguistic-space Faithfulness Divergence and generalised ECE (Wang et al., 2024) in Table 2. Raw metric value changes are provided in Appendix E.1.

Table 2: In-domain and cross-domain linguistic-space calibration metric percentage changes for both Faithfulness Divergence and generalised ECE. We report percentage change relative to the pre-calibration metrics (mean ± std across models). Green text indicates calibration improvement (lower error).

Metric	Signal	Train/Test	MMLU	SQuAD 2.0	TruthfulQA
Faithfulness Divergence Mean Reduction	Linguistic Confidence	MMLU	Δ11.4±19.2%	Δ34.6±11.6%	Δ29.6±12.4%
		SQuAD 2.0	Δ42.7±4.2%	Δ62.0±2.3%	Δ59.8±3.9%
		TruthfulQA	Δ27.0±17.2%	Δ65.8±2.2%	Δ60.5±4.6%
	Token Probability	MMLU	Δ10.2±23.2%	Δ52.6±8.1%	Δ56.4±6.2%
		SQuAD 2.0	Δ32.8±4.9%	Δ64.9±2.4%	Δ70.4±3.3%
		TruthfulQA	Δ48.8±5.1%	Δ64.2±3.1%	Δ64.6±3.0%
	Semantic Uncertainty	MMLU	Δ21.6±13.9%	Δ36.1±12.4%	Δ43.0±9.8%
		SQuAD 2.0	Δ43.9±4.8%	Δ66.0±4.1%	Δ65.5±4.3%
		TruthfulQA	Δ59.9±2.8%	Δ66.4±1.9%	Δ66.1±2.9%
Generalised ECE Mean Reduction	Linguistic Confidence	MMLU	Δ38.1±8.0%	Δ20.7±7.6%	Δ12.2±6.4%
		SQuAD 2.0	Δ21.9±11.4%	Δ43.4±1.8%	Δ35.0±2.3%
		TruthfulQA	Δ21.9±7.1%	Δ47.2±1.9%	Δ39.6±0.9%
	Token Probability	MMLU	Δ46.8±8.7%	Δ38.5±11.1%	Δ37.5±10.1%
		SQuAD 2.0	Δ23.5±2.2%	Δ50.6±1.0%	Δ61.4±1.1%
		TruthfulQA	Δ60.2±2.6%	Δ48.6±2.4%	Δ42.3±2.3%
	Semantic Uncertainty	MMLU	Δ49.8±8.7%	Δ30.8±8.5%	Δ33.7±5.7%
		SQuAD 2.0	Δ30.1±16.9%	Δ58.7±3.8%	Δ49.8±4.1%
		TruthfulQA	Δ62.3±3.1%	Δ56.9±4.0%	Δ45.0±3.1%

All three signals yield improvements across both metrics and all domain pairs. Semantic uncertainty exhibits the strongest cross-domain transferability, producing the highest gains with the lowest variance across models in both in-domain and cross-domain settings. As a result, the empirical evidence supports semantic uncertainty as the most robust confidence signal for RALC.

In contrast, cross-domain calibrators occasionally outperform in-domain ones. This anomaly occurs when the target domain has a weak miscalibration bias, providing insufficient signal for its in-domain calibrator to learn a reliable correction. A cross-domain source with a stronger, more consistent bias learns a more decisive correction that transfers to the target domain, provided both share the same direction of miscalibration. We provide a detailed investigation in Appendix E.2.

6 Discussion and Conclusion

Limitations and future work As a downstream framework, RALC's calibration quality is ultimately bounded by the quality of the upstream confidence signal. Whilst semantic uncertainty is the most robust signal evaluated, its reliance on multi-round self-consistency sampling incurs significant inference cost. Identifying signals that match its performance at a lower computational expense is therefore an important direction for future work. Additionally, the hedging expression lexicon covers common confidence expressions and is not intended to represent the full landscape of linguistic uncertainty cues. Future work could investigate domain-specific hedging vocabularies and audience-adapted confidence scoring, which would enhance the specialisation and expressiveness of RALC in targeted deployment settings.

Conclusion In this work, we introduce a distributional treatment of linguistic confidence as a Beta distribution, define faithfulness as a complementary dimension of confidence evaluation, and present Faithfulness Divergence to quantify it from an information-theoretic perspective. Building on these foundations, we propose Retrieval-Augmented Linguistic Calibration (RALC), a principled post-hoc pipeline that calibrates confidence in the linguistic space, yielding well-calibrated and faithful responses that consistently outperform prompt-based baselines across models and benchmarks.

AI USE STATEMENT

In this work, we used generative AI tools for the following purposes. First, we used AI tools to polish the writing of the manuscript; all content, ideas, and arguments originate from and were thought through by the authors, with AI assistance limited to language editing and phrasing. Second, we used LLMs to assist with debugging and refactoring the experiment pipelines, under extensive manual control and review by the authors at every step. Third, we use LLMs as a judge within the experiment pipeline itself; this is a core, non-standard component of our methodology rather than a writing aid, and is described in full in Section 3, Section 5.1, and Appendix C. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artefacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, the collection of personal data, or crowdsourcing. All datasets used for evaluation (MMLU, SQuAD 2.0, and TruthfulQA) and all language models used (Llama-3.1-8B-Instruct, Qwen3-8B, Mistral-7B-Instruct-v0.3, GPT-OSS-20B, and Gemma-4-31B-IT) are publicly available and are used in accordance with their stated terms. The human-annotated linguistic confidence benchmark used to validate our LLM evaluator ensemble is drawn from an existing published source (Tao et al., 2025b) rather than newly collected data.

RALC is intended to improve the trustworthiness of LLM deployments by producing confidence expressions that more faithfully reflect a model's underlying correctness, reducing rather than amplifying overconfident outputs. We identify no direct path to harmful misuse. The principal limitation with ethical relevance is that RALC's calibration quality is bounded by the quality of its upstream confidence signal, and an insufficiently expressive hedging lexicon could produce miscalibrated hedging language in specialised domains; this is discussed further in Section 6.

REPRODUCIBILITY STATEMENT

We have ensured the reproducibility of our results by providing the hyperparameters and experimental configurations used throughout the paper. All prompts used in the pipeline, including the QA prompts, the grader prompt, the linguistic confidence evaluator prompt, the semantic clustering prompt, and the retrieval-augmented and direct Beta-guided rewriting prompts, are given verbatim in Appendix C and Appendix D. All model names, dataset splits (30% train, 70% test), and hyperparameters (retrieval depth $k = 5$, chosen via the ablation in Appendix D.5; 20 self-consistency samples) are specified in Section 5, with full sampling temperatures and hosting configurations for every model given in Appendix F. The derivations underlying the Beta distribution parameterisation, Faithfulness Divergence, Platt scaling on distribution means, the method-of-moments estimator, and the distributional generalisations of the Brier score and negative log-likelihood are given in full in Appendix A and Appendix B. All three evaluation datasets are publicly available, and the hedging lexicon construction procedure is documented in full in Appendix D.1.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Anthropic. Introducing Claude Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, February 2026.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5050–5063, 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.

Google DeepMind. Gemma 4: Our most intelligent open models. https://deepmind.google/models/gemma/gemma-4/, 2025a.

Google DeepMind. Gemini 3 Flash: Frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/, 2025b.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. Calibrating long-form generations from large language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 13441–13460, 2024.

Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10): 1295–1306, 2009.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3769–3793, 2025.

Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. " i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pp. 822–835, 2024.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pp. 623–631. PMLR, 2017.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Tom A Lamb, Desi R Ivanova, Philip Torr, and Tim GJ Rudner. Semantic-level confidence calibration of language models via temperature scaling. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022a.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022b.

Charlie Marx, Sofian Zalouk, and Stefano Ermon. Calibration by distribution matching: Trainable kernel calibration metrics. *Advances in Neural Information Processing Systems*, 36:25910–25928, 2023.

Meta AI. Llama 3.1 model card and prompt formats. `https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/`, 2024.

Mistral AI Team. Mistral 7B. `https://mistral.ai/news/announcing-mistral-7b`, September 2023.

Satoshi Morita, Peter F Thall, and Peter Müller. Determining the effective sample size of a parametric prior. *Biometrics*, 64(2):595–602, 2008.

OpenAI. Introducing GPT-5.4. `https://openai.com/index/introducing-gpt-5-4/`, March 2026.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *International Conference on Machine Learning*, pp. 5897–5906. PMLR, 2019.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231, 2025.

Daniel James Sumler, Lee Devlin, Simon Maskell, and Richard O Lane. An entropic metric for measuring calibration of machine learning models. *arXiv preprint arXiv:2502.14545*, 2025.

Linwei Tao, Yi-Fan Yeh, Minjing Dong, Tao Huang, Philip Torr, and Chang Xu. Revisiting uncertainty estimation and calibration of large language models. *arXiv preprint arXiv:2505.23854*, 2025a.

Linwei Tao, Yi-Fan Yeh, Bo Kai, Minjing Dong, Tao Huang, Tom A Lamb, Jialin Yu, Philip HS Torr, and Chang Xu. Can large language models express uncertainty like human? *arXiv preprint arXiv:2509.24202*, 2025b.

Andrew Thompson and Vivek Desai. Extending confidence calibration to generalised measures of variation. *arXiv preprint arXiv:2602.12975*, 2026.

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.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, volume 1, pp. 2001, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Hang Zhang and Laurence T Maloney. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in neuroscience*, 6:1, 2012.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. On the calibration of large language models and alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9778–9795, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.654. URL https://aclanthology.org/2023.findings-emnlp.654/.

A THEORY

A.1 BETA DISTRIBUTION ESTIMATION

The Beta distribution is a natural choice for modelling random variables supported on $[0, 1]$, such as confidence scores and empirical accuracies. A Beta distribution with parameters (α, β) has probability density function

$$p(x \mid \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in [0, 1],$$

where $\alpha > 0$, $\beta > 0$, and $\mathrm{B}(\alpha, \beta)$ denotes the Beta function. Its mean, variance, and concentration factor are given by

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \qquad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \qquad \kappa = \alpha + \beta.$$

A.1.1 METHOD OF MOMENTS

Let $\mathcal{S}_Z = \{s_t\}_{t=1}^{T_Z}$ denote the pseudo-observation set for a fixed target Z, with $s_t \in [0, 1]$. Denote the empirical mean and variance by

$$\bar{s}_Z = \frac{1}{T_Z} \sum_{t=1}^{T_Z} s_t, \qquad v_Z = \frac{1}{T_Z - 1} \sum_{t=1}^{T_Z} (s_t - \bar{s}_Z)^2.$$

The method of moments estimates (α, β) by matching these empirical moments to the theoretical mean and variance of the Beta distribution. Solving the resulting system yields

$$\hat{\alpha}_Z = \bar{s}_Z \left(\frac{\bar{s}_Z(1-\bar{s}_Z)}{v_Z} - 1 \right), \qquad \hat{\beta}_Z = (1-\bar{s}_Z) \left(\frac{\bar{s}_Z(1-\bar{s}_Z)}{v_Z} - 1 \right),$$

provided that $v_Z > 0$ and $v_Z < \bar{s}_Z(1-\bar{s}_Z)$. When this condition is not met, two fallback cases are distinguished.

Boundary-degenerate case ($\bar{s}_Z \in \{0, 1\}$, i.e. all observations are 0 or all are 1): the mean-preserving concentration formula cannot be applied since one parameter would be zero. Instead we set

$$\hat{\alpha}_Z = \bar{s}_Z \cdot \kappa, \qquad \hat{\beta}_Z = (1-\bar{s}_Z) \cdot \kappa,$$

with $\kappa = T_Z$, and rely on the clipping step below to lift any zero parameter to 10^{-6}. This yields $\mathrm{Beta}(T_Z, 10^{-6})$ when all observations are 1, placing nearly all mass near 1, and $\mathrm{Beta}(10^{-6}, T_Z)$ when all observations are 0, placing nearly all mass near 0.

Interior-degenerate case ($v_Z = 0$ or $v_Z \geq \bar{s}_Z(1-\bar{s}_Z)$ with $\bar{s}_Z \in (0, 1)$): the observations are constant or insufficiently dispersed at an interior value. We preserve the empirical mean by setting

$$\hat{\alpha}_Z = \bar{s}_Z \cdot \kappa, \qquad \hat{\beta}_Z = (1-\bar{s}_Z) \cdot \kappa,$$

with $\kappa = T_Z$. This produces a Beta distribution with mean $\bar{s}_Z$ and high concentration proportional to the number of observations, reflecting the certainty implied by the consistency of the pseudo-observations.

In all cases, both parameters are clipped to a minimum of 10^{-6} to ensure numerical stability:

$$\hat{\alpha}_Z \leftarrow \max(\hat{\alpha}_Z,\, 10^{-6}), \qquad \hat{\beta}_Z \leftarrow \max(\hat{\beta}_Z,\, 10^{-6}).$$

A.2 CLASSICAL INSTANCE-LEVEL CALIBRATION METRICS WITH DISTRIBUTION GENERALISATION

A.2.1 EXPECTED BRIER SCORE

Let $S = \mathrm{Beta}(\alpha, \beta)$ denote the predictive confidence distribution, let $p \sim S$ denote the scalar confidence value drawn from it, and let $y \in \{0, 1\}$ be the ground-truth label. The Expected Brier Score is defined as

$$\mathbb{E}_{p \sim S}\left[(p - y)^2\right].$$

This admits a closed form:

$$\mathbb{E}[(p-y)^2] = \mathrm{Var}(p) + (\mathbb{E}[p]-y)^2,$$

where

$$\mathbb{E}[p] = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{Var}(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Hence,

$$\mathbb{E}[(p-y)^2] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \left(\frac{\alpha}{\alpha+\beta} - y\right)^2.$$

A.3 Expected Negative Log-Likelihood (NLL)

Given $S = \mathrm{Beta}(\alpha, \beta)$, $p \sim S$, and label $y \in \{0, 1\}$, we define the distributional Negative Log-Likelihood as the expected Bernoulli log-loss under S:

$$\mathcal{L}_{\mathrm{NLL}}(S, y) = \mathbb{E}_{p\sim S}\left[-\log p(y \mid p)\right],$$

where $p(y \mid p) = p^y(1-p)^{1-y}$.

This yields the closed form:

$$\mathcal{L}_{\mathrm{NLL}}(S, y) = \begin{cases} \psi(\alpha+\beta) - \psi(\alpha), & y = 1, \\ \psi(\alpha+\beta) - \psi(\beta), & y = 0, \end{cases}$$

where $\psi(\cdot)$ denotes the digamma function.

Equivalently, this can be written as

$$\mathcal{L}_{\mathrm{NLL}}(S, y) = y\left[\psi(\alpha+\beta) - \psi(\alpha)\right] + (1-y)\left[\psi(\alpha+\beta) - \psi(\beta)\right].$$

B FURTHER DISCUSSION ON FAITHFULNESS

B.1 DISTRIBUTIONAL REPRESENTATION OF LINGUISTIC CONFIDENCE

We represent response-level linguistic confidence as $S = \mathrm{Beta}(\alpha, \beta)$, where the mean $\mu = \alpha/(\alpha+\beta)$ captures the consensus perceived confidence across readers and the concentration $\kappa = \alpha + \beta$ captures the strength of that consensus. Two responses may elicit identical mean confidence yet differ substantially in concentration: one may induce consistent perceptions across readers whilst the other induces highly variable ones. Scalar representations discard this distinction; the distributional representation preserves it, and it is precisely this information that faithfulness evaluation requires.

B.2 FAITHFULNESS DIVERGENCE

Faithfulness measures the degree of surprise induced by truth revelation. A response that elicits high mean confidence with high concentration is highly surprising when incorrect, as it represents a strongly held prior belief that the ground truth contradicts. The same misalignment expressed with low concentration is less surprising, since the prior was weakly held and requires only a modest update. Faithfulness Divergence (FD) operationalises this intuition.

Formally, for instance i with correctness label y_i, the estimated confidence distribution serves as the prior $S_i = \mathrm{Beta}(\alpha_i, \beta_i)$. Upon observing y_i, the prior is updated by a single Bernoulli observation to yield the posterior $S_i^* = \mathrm{Beta}(\alpha_i + y_i,\ \beta_i + 1 - y_i)$. The KL divergence $\mathrm{KL}(S_i^* \| S_i)$ quantifies the normalised magnitude of the required belief revision. However, KL divergence alone does not account for the strength of agreement underlying the prior. Under the Beta–Bernoulli model, the concentration $\alpha_i + \beta_i$ is interpretable as the effective sample size of the prior: a larger concentration encodes a more strongly held belief, and an identical KL divergence therefore represents a larger total epistemic adjustment when the prior is more concentrated (Morita et al., 2008). FD scales the KL divergence by this effective sample size,

$$\mathrm{FD}_i := (\alpha_i + \beta_i) \cdot \mathrm{KL}(S_i^* \,\|\, S_i)\,,$$

yielding a scalar that quantifies the degree of surprise induced by truth revelation, scaled by the strength of agreement encoded in the prior.

B.3 SCOPE AND INTENDED USE

FD is designed as a diagnostic measurement tool, not as a proper scoring rule or a training objective. It quantifies the surprise induced at the instance level when the ground truth is revealed, providing a complementary lens on confidence quality that population-level metrics such as ECE do not capture. The goal of a well-calibrated response is not to minimise FD in isolation, but to express confidence that is consistent with the model's actual uncertainty; FD measures how far a given response falls from that standard. It should therefore be interpreted alongside calibration metrics rather than treated as a sole optimisation target.

B.4 Additional Faithfulness Divergence ablation studies

In addition to the empirical ablation in Table 1, we conduct theoretical ablations to further illustrate the unique properties of Faithfulness Divergence and its alignment with our definition of surprise upon truth revelation, compared against alternative metrics including KL divergence, expected Brier score, and expected NLL. Two controlled settings are examined: varying the confidence mean under a fixed concentration, and varying concentration under a fixed mean, each evaluated against a binary ground-truth label.

Figure 4 shows that all four metrics correctly capture mean deviation from the ground-truth label as increasing surprise, assigning monotonically higher values as the confidence mean moves further from the outcome. However, Figure 5 demonstrates that only FD encodes concentration as an amplifier of misalignment, assigning monotonically higher surprise as concentration increases for a fixed misaligned mean, whilst the alternative metrics behave otherwise. FD therefore uniquely quantifies surprise upon truth revelation in accordance with our definition, which requires that a more strongly held incorrect belief be treated as more surprising.

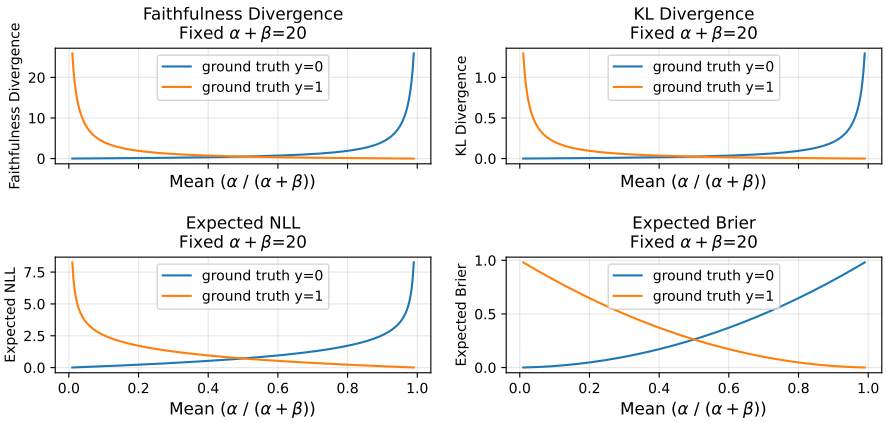

Figure 4: Faithfulness Divergence, KL divergence, expected Brier score, and expected NLL under varying confidence means with fixed concentration ($\alpha + \beta = 20$) against a binary ground-truth label. All metrics increase monotonically as the mean deviates from the ground-truth label, reflecting greater surprise upon truth revelation for more misaligned beliefs.

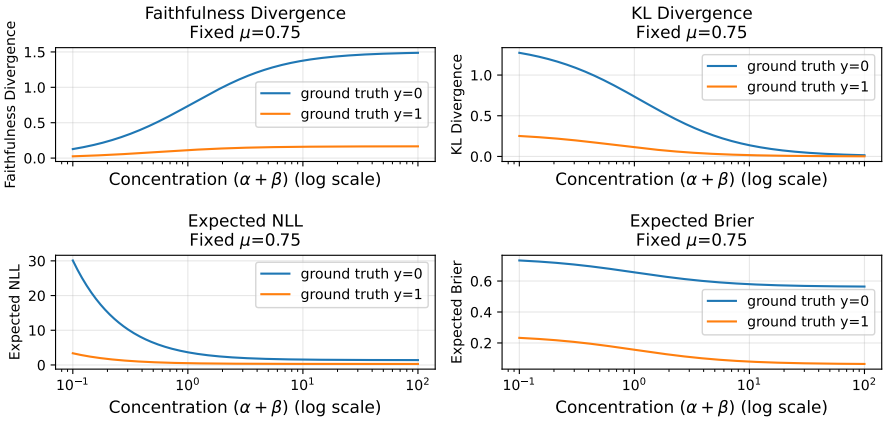

Figure 5: Faithfulness Divergence, KL divergence, expected Brier score, and expected NLL under varying concentration with a fixed misaligned mean ($\mu = 0.75$) against a binary ground-truth label. Only Faithfulness Divergence correctly increases monotonically with concentration, encoding the intuition that a strongly held incorrect belief induces greater surprise upon truth revelation than a weakly held one of equal mean.

C CONFIDENCE EVALUATION IMPLEMENTATION

C.1 QA PROMPTS

We detail the exact prompts used to sample responses across MMLU, SQuAD 2.0, and TruthfulQA with different confidence estimation methods, including linguistic confidence, token probability, and semantic uncertainty.

C.1.1 MMLU

Direct QA prompt

Answer the following question using a succinct (at most one sentence) and full answer.

{ question }

{% for choice in choices %}

{ 'ABCDEFGHIJKLMNOPQRSTUVWXYZ'[loop.index0] }. { choice }

{% endfor %}

Answer: [Return the answer letter only]

Hedged QA prompt

Answer the following question using a succinct (at most one sentence) and full answer. If you are uncertain about your answer to the question, convey this uncertainty linguistically by precisely hedging this answer. Please include the option letter in your answer.

Question: { question }

{% for choice in choices %}

{ 'ABCDEFGHIJKLMNOPQRSTUVWXYZ'[loop.index0] }. { choice }

{% endfor %}

Answer:

C.1.2 SQUAD 2.0

Direct QA prompt

Answer the following question using a succinct (at most one sentence) and full answer.

Title: { title }

Background: { context }

Question: { question }

Answer:

Hedged QA prompt

Answer the following question using a succinct (at most one sentence) and full answer. If you are uncertain about your answer to the question, convey this uncertainty linguistically by precisely hedging this answer.

Title: { title }

Background: { context }

Question: { question }

Answer:

C.1.3 TRUTHFULQA

Direct QA prompt

Answer the following question using a succinct (at most one sentence) and full answer.

Question: {question}

Answer:

Hedged QA prompt

Answer the following question using a succinct (at most one sentence) and full answer. If you are uncertain about your answer to the question, convey this uncertainty linguistically by precisely hedging this answer.

Question: {question}

Answer:

C.2 LLM LINGUISTIC CONFIDENCE EVALUATOR

C.2.1 EVALUATOR PROMPT

An LLM linguistic confidence evaluator ensemble is prompted as follows. We use an ensemble of models rather than a single evaluator to capture the complex linguistic relationships arising from co-occurring cues and their contextual interactions within a statement, whilst averaging out idiosyncratic biases of individual models. To align with human perception, we provide the LLMs with human-annotated linguistic cues and their associated confidence profiles from Tao et al. (2025b) as reference. The LLM is then asked to return a confidence score between 0 and 100 based solely on the linguistic cues present in the sentence, without using any external or prior knowledge to assess the knowledge conveyed by the sentence. The extracted output score is then normalised to $[0, 1]$.

LLM linguistic confidence evaluator prompt

Please provide only a confidence score between 0 and 100, based solely on the degree of confidence expressed in the tone and linguistic cues of the following sentence, without using any external or prior knowledge to assess the knowledge conveyed by the sentence. Please pay attention to the hedging language used and the overall assertiveness of the statement. If the sentence abstains from answering by pointing out the insufficiency of information with a firm tone, the score should be towards 100. If the sentence contains random guesses or abstention, the score should be towards 0. If the sentence is stated with strong certainty or no hedging, the score should be towards 100. If the sentence does not contain any hedging language or is a succinct, decisive short answer, the score should be towards 100, too.

To align with human perception, here are some human-annotated linguistic cues with perceived confidence profiles (mean and standard deviation) for your reference: {human_annotated_cues}

Here is the sentence: {sentence}

Confidence Score: [Return only a number between 0 and 100 without any additional text or explanation]

C.2.2 LLM ENSEMBLE VS. HUMAN BENCHMARK

We compare the LLM-ensemble confidence scores against human annotations on the benchmark of Tao et al. (2025b). The benchmark consists of human-annotated confidence scores across various statements. We employ our LLM ensemble to generate confidence scores for each statement in a similar manner to human annotators. The results in Figure 6 and Table 3 confirm that the ensemble largely matches human confidence judgements for common hedging expressions.

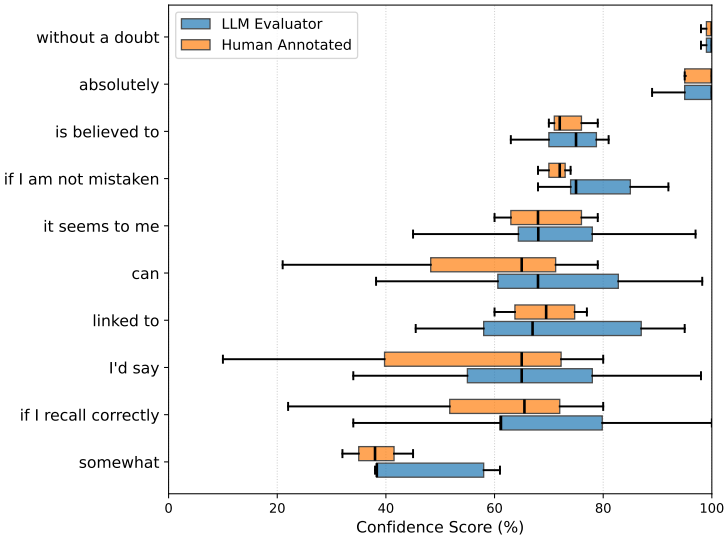

Figure 6: LLM vs. human perceived linguistic confidence on the human-annotated benchmark of Tao et al. (2025b). The LLM ensemble largely follows human confidence annotations across confidence levels.

Table 3: Rank and linear correlations between LLM-ensemble and human-annotated confidence scores on the benchmark of Tao et al. (2025b). All p-values are $< 10^{-10}$, rejecting the null hypothesis of zero correlation ($H_0 : \rho = 0$) and confirming that the ensemble reliably reproduces human perception of linguistic confidence cues.

Metric	**Coefficient**	**p-value**
Spearman ρ	0.8535	$< 10^{-10}$
Pearson r	0.8450	$< 10^{-10}$
Kendall τ	0.6909	$< 10^{-10}$

C.3 GRADER PROMPT

LLM grader prompt

If the predicted answer matches, implies or covers the correct answer, the grade is CORRECT.

If the predicted answer does not match, imply or cover the correct answer, the grade is INCORRECT. Do NOT grade it as INCORRECT if the predicted answer abstains from answering (e.g. "I don't know the answer..." or "I have no idea...").

If the predicted answer is empty, none or abstention (e.g. "I don't know the answer..." or "I have no idea..."), grade the predicted answer as NOT_ATTEMPTED instead of CORRECT or INCORRECT. If the predicted answer makes an attempt (even random guesses), do not grade it as NOT_ATTEMPTED.

Ignore any explanation or linguistic cues present in the predicted answer. Don't apologise or correct yourself if there was a mistake; we are just trying to grade the answer.

Question: {question}

Correct answer: {correct_answer}

Predicted answer: {"" if predicted_answer is None else predicted_answer}

Grade the predicted answer of this new question as one of:

A: CORRECT

B: INCORRECT

C: NOT_ATTEMPTED

Just return one of the letters "A", "B", or "C", with no text around it.

C.4 LLM-BASED SEMANTIC CLUSTERING PROMPT

LLM-based semantic clustering prompt

You are a strict JSON generator. Group semantically equivalent candidate responses to the same question. Ignore any linguistic markers of uncertainty or hedging and focus solely on the core meaning of the responses.

Return a JSON object with a single key `"semantic_ids"`, a list of integers aligned with the response order. Responses that are semantically equivalent (bidirectional entailment) must share the same integer id. Use 0-based ids. Semantic ids represent the semantic cluster assignment for each response. Return ONLY the JSON object, no extra text.

For instance, given the question and candidate responses:

Question: What is the capital of France?

Candidate responses:
0: 'I guess Paris is the capital of France.'
1: 'Paris is the capital city of France.'
2: 'The capital of France is Berlin.'

The correct JSON output would be:

```
{"semantic_ids": [0, 0, 1]}
```

Now, please group the following candidate responses to the given question and return the JSON object:

Question: {question}

Candidate responses: {responses}

```
{"semantic_ids": [...]}
```

D POST-HOC LINGUISTIC CALIBRATION IMPLEMENTATION

D.1 LEXICON CONSTRUCTION

The lexicon is constructed in three stages: hedging expression sourcing, confidence score collection, and Beta distribution fitting.

Hedging expression sourcing A curated set of K hedging expressions spanning the full confidence spectrum is generated by prompting Claude-Sonnet-4.6 to produce words and phrases that humans use to convey varying degrees of certainty, from expressions of complete ignorance (e.g. "I have no idea", "my random guess is") to expressions of near-certain belief (e.g. "without a doubt", "I can confirm").

Hedging expression sourcing prompt

Generate a Python list of words or expressions that humans use to convey the level of confidence, certainty, or hedging in their statements (without a subject, only the linguistic cues). These words should include common hedging phrases, adverbs, and qualifiers that indicate varying degrees of certainty or uncertainty, from extremely low confidence (like I do not know, my random guess is, etc) to high confidence (certain, sure, definitely).

The list should be comprehensive and cover a wide range of expressions used in everyday language as well as in academic and professional contexts.

Confidence score collection For each hedging expression w_k, GPT-OSS-20B (Agarwal et al., 2025) generates 20 candidate sentences by rewriting a randomly selected non-verifiable statement drawn from a fixed pool of 12 non-verifiable template sentences, with instructions to incorporate w_k naturally and to avoid introducing additional hedging cues. Each generated sentence is then independently evaluated by three LLM evaluators (Llama-3.1-8B-Instruct (Meta AI, 2024), Qwen3-8B (Yang et al., 2025), Mistral-7B-Instruct-v0.3 (Mistral AI Team, 2023)), each prompted to assign a perceived-confidence score on a 0–100 scale based solely on the linguistic cues present, ignoring factual content. Human-annotated reference profiles from Tao et al. (2025b) are provided in-context to anchor model ratings to human perception. Each evaluator scores every sentence 3 times (temperature $= 1$), yielding up to $20 \times 3 \times 3 = 180$ raw scores per hedging expression.

Non-verifiable template sentences

"There is a correlation between X and Y."
"It rains tomorrow."
"The experiment shows a significant effect."
"The new policy improves the economy."
"The medication is effective in treating the disease."
"The new product is successful in the market."
"The neighbour is home."
"The movie is good."
"The restaurant serves delicious food."
"The city is the oldest in the country."
"The book is informative."
"The report is not accurate."

Non-verifiable sentence rewriting prompt

Given a linguistic cue: "{word}", rewrite one of the following non-verifiable statements to naturally include this cue to convey the intended level of confidence, certainty, or hedging. Please do not use other hedging words, hedging phrases or linguistic cues in the sentence other than the specified linguistic cue.

Example sentences to rewrite: {selected_sentence}

Do not use other hedging words or linguistic cues in the sentence. Do not combine linguistic cues. Do not include labels like "Example:" or "Sentence:". Just provide the statement.

Beta distribution fitting All raw scores for a given expression are aggregated across sentences, evaluators, and repeated passes, then normalised to $[0, 1]$ and clipped to $(10^{-6},\, 1 - 10^{-6})$ to avoid boundary degeneracy. A Beta distribution is fitted to the pooled scores by maximum likelihood estimation (with fixed support $[0, 1]$), yielding the lexicon entry $\big(w_k,\, \mathrm{Beta}(\alpha_k, \beta_k)\big)$. The resulting lexicon $\{(w_k,\, \mathrm{Beta}(\alpha_k, \beta_k))\}_{k=1}^{K}$ is used at inference time for Wasserstein-distance-based retrieval.

D.2 Sample hedging expressions from the lexicon

Figure 7 shows sample hedging expressions from the lexicon, with their corresponding Beta distributions over perceived confidence by our LLM ensemble.

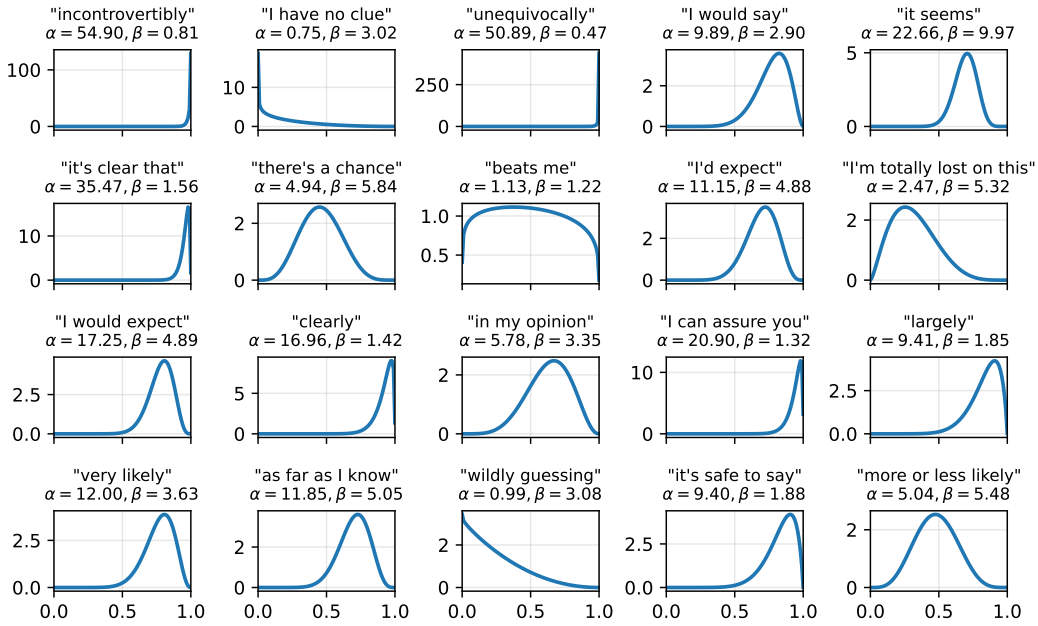

Figure 7: Sample hedging expressions from the lexicon, with their corresponding Beta distributions over perceived confidence by our LLM ensemble.

D.3 Retrieval-augmented linguistic calibration rewriting prompt

Retrieval-augmented linguistic calibration prompt

Given an original response and a list of target hedging words with their confidence profiles (Beta Distributions), rewrite the response to appropriately reflect the confidence level indicated by the set of target hedging words. You must preserve the original meaning of the response, as we are only adjusting the tone to match the confidence level suggested by the hedging words. Ensure the new response sounds natural and fluent.

Original response: My answer to the question is: "{response}"

Target hedging words with confidence profiles: {hedges}

Please return only the rewritten sentence without any explanation.

New response:

D.4 Direct Beta-Guided Rewrite calibration prompt

Direct Beta-Guided Rewrite prompt

Given an original response and a Beta distribution, rewrite the response to appropriately reflect the confidence level indicated by the given Beta distribution by using hedging language. You must preserve the original meaning of the response, as we are only adjusting the tone to match the confidence level suggested by the hedging words. Ensure the new response sounds natural and fluent.

Original response: My answer to the question is: "{response}"

Target Beta distribution: Beta(alpha={alpha:.2f}, beta={beta:.2f})

Please return only the rewritten sentence without any explanation.

New response:

D.5 Choice of k for hedging expression retrieval

Given the pre-constructed lexicon of hedging expressions, we perform an ablation study on the choice of k for the KNN retrieval of hedging expressions in the RALC pipeline. The following figure shows the impact on Faithfulness Divergence and generalised ECE for different choices of k across linguistic confidence (LC), token probability (TP), and semantic uncertainty (SU) as retrieval signals for Llama-3.1-8B-Instruct on the TruthfulQA dataset. Figure 8 shows that both metrics are not highly sensitive to the choice of k, with $k = 5$ showing consistently better marginal performance in the exploration landscape with the lowest Faithfulness Divergence and generalised ECE. Therefore, we choose $k = 5$ for both the in-domain and cross-domain calibration experiments in this work.

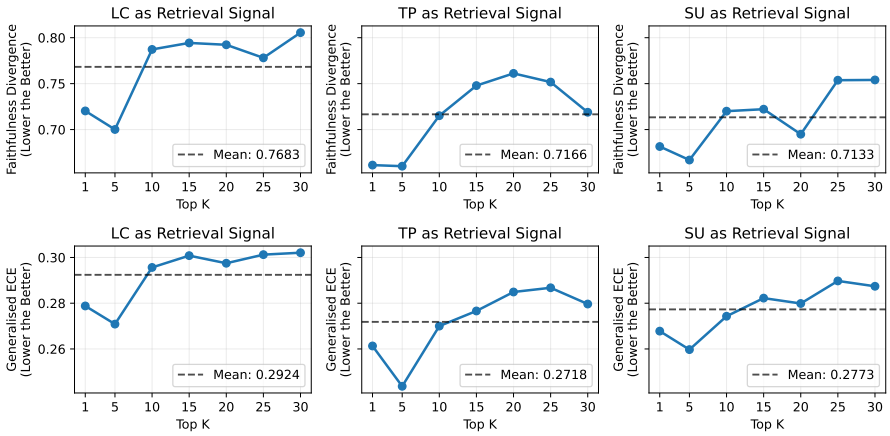

Figure 8: Impact of the choice of k for the KNN retrieval of hedging expressions in RALC pipeline on Faithfulness Divergence and generalised ECE across different confidence signals for Llama-3.1-8B-Instruct on the TruthfulQA dataset. The results show that both metrics are not highly sensitive to the choice of k within a reasonable range, with $k = 5$ showing consistently better marginal performance in the exploration range.

D.6 Confidence distribution profiles across different estimators

Figure 9 shows the distribution of confidence standard deviations across responses for each signal. Linguistic confidence exhibits the highest variability, whilst token probability and semantic uncertainty each produce a substantial proportion of zero-variance distributions, arising when responses share identical token probability profiles or collapse into a single semantic cluster. These degenerate cases are handled by clipping the (α, β) parameters as specified in Appendix A.1.1, ensuring compatibility with both the calibration map and the Wasserstein-based retrieval step.

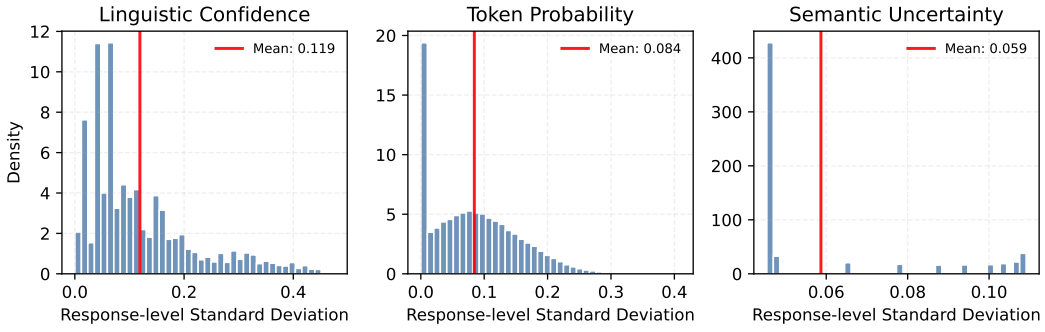

Figure 9: For each confidence signal for Direct QA responses, we plot the distribution of the standard deviation of the confidence distribution across responses.

D.7 CALIBRATION MAP ABLATION STUDY

Table 4: Ablation over signal-space calibration maps, averaged across all datasets and models (± 1 standard deviation). Bold green denotes the best (lowest) value per column. Platt scaling achieves the lowest error in five of six columns and is adopted in the RALC pipeline.

	Linguistic confidence		Token probability		Semantic uncertainty	
Method	Gen. ECE	FD	Gen. ECE	FD	Gen. ECE	FD
Uncalibrated	$0.280_{\pm 0.107}$	$1.487_{\pm 0.628}$	$0.262_{\pm 0.111}$	$68.879_{\pm 96.520}$	$0.267_{\pm 0.114}$	$2.112_{\pm 0.762}$
Platt scaling (Platt, 1999)	$\mathbf{0.109}_{\pm 0.018}$	$\mathbf{0.486}_{\pm 0.042}$	$\mathbf{0.085}_{\pm 0.032}$	$\mathbf{0.500}_{\pm 0.060}$	$\mathbf{0.060}_{\pm 0.014}$	$0.506_{\pm 0.051}$
Isotonic regression (Zadrozny & Elkan, 2002)	$0.114_{\pm 0.019}$	$0.685_{\pm 0.530}$	$0.090_{\pm 0.034}$	$1.646_{\pm 1.495}$	$0.065_{\pm 0.017}$	$0.718_{\pm 0.439}$
Histogram binning (Zadrozny & Elkan, 2001)	$0.110_{\pm 0.019}$	$0.488_{\pm 0.042}$	$0.090_{\pm 0.039}$	$0.514_{\pm 0.104}$	$0.060_{\pm 0.017}$	$\mathbf{0.502}_{\pm 0.047}$
Temperature scaling (Guo et al., 2017)	$0.127_{\pm 0.022}$	$0.547_{\pm 0.148}$	$0.102_{\pm 0.032}$	$0.646_{\pm 0.380}$	$0.080_{\pm 0.031}$	$0.520_{\pm 0.056}$

We ablate the signal-space calibration map across Platt scaling (Platt, 1999), isotonic regression (Zadrozny & Elkan, 2002), histogram binning (Zadrozny & Elkan, 2001), and temperature scaling (Guo et al., 2017), applied to the distribution means of all three confidence signals and evaluated on generalised ECE and Faithfulness Divergence in the signal space. The results are averaged across all datasets and models and reported in Table 4. Platt scaling is the best performer in the signal space across both metrics and all confidence signals, whilst isotonic regression exhibits instability on small calibration sets, histogram binning trails on FD despite being competitive on ECE, and temperature scaling performs worst overall. We therefore adopt Platt scaling as the signal-space calibration map in the RALC pipeline.

D.8 CONFIDENCE SIGNAL PROPAGATION QUALITY

We evaluate the quality of RALC by measuring the correlation between the calibrated confidence signal and the linguistic confidence in the rewritten responses perceived by our LLM evaluator ensemble for each confidence signal. Our pipeline accurately propagates the calibrated confidence signal into language, as evidenced by a positive Spearman's correlation ρ consistently above 0.9 across all confidence signals.

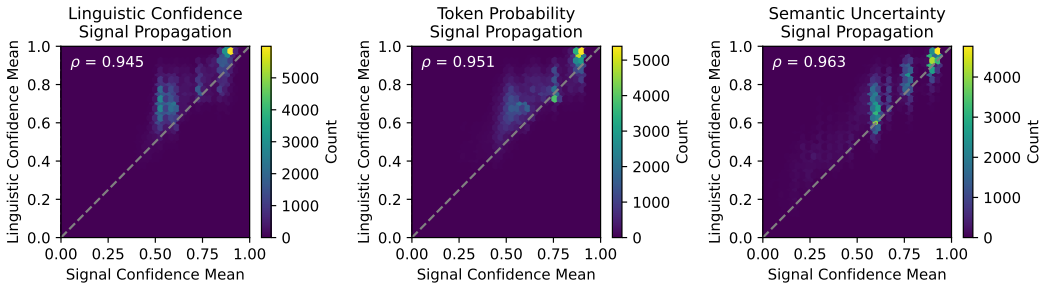

Figure 10: We evaluate the quality of RALC by measuring the correlation between the calibrated confidence signal and the linguistic confidence in the rewritten responses perceived by our LLM evaluator ensemble. Across all confidence signals, our pipeline effectively propagates the calibrated confidence signal into language, as evidenced by a positive correlation consistently well above 0.9.

E ADDITIONAL RESULTS

E.1 IN-DOMAIN AND CROSS-DOMAIN CALIBRATION RESULTS

Table 5 provides the absolute metric value changes corresponding to the percentage changes in Table 2, broken down by confidence estimator, training dataset, and evaluation dataset. Each entry reports the mean reduction in Faithfulness Divergence or generalised ECE across models (mean $\pm$ std), with green indicating improvement and red indicating deterioration. In-domain results appear on the diagonal; off-diagonal entries reflect cross-domain transfer.

Table 5: In-domain and cross-domain linguistic-space calibration metric value changes for both Faithfulness Divergence and generalised ECE. We report the value change relative to the pre-calibration metrics (mean $\pm$ std across models). Green text indicates calibration improvement (lower error), red indicates deterioration.

Metric	**Signal**	**Train/Test**	MMLU	SQuAD 2.0	TruthfulQA
Faithfulness Divergence Mean Reduction	Linguistic Confidence	MMLU	Δ0.1057$\pm$0.1793	Δ0.6384$\pm$0.2359	Δ0.5436$\pm$0.2420
		SQuAD 2.0	Δ0.3598$\pm$0.0779	Δ1.1440$\pm$0.1398	Δ1.0996$\pm$0.1566
		TruthfulQA	Δ0.2904$\pm$0.1946	Δ1.2138$\pm$0.1430	Δ1.1127$\pm$0.1695
	Token Probability	MMLU	Δ0.0952$\pm$0.2094	Δ0.9714$\pm$0.1862	Δ1.0355$\pm$0.1556
		SQuAD 2.0	Δ0.2768$\pm$0.0674	Δ1.1987$\pm$0.1472	Δ1.2931$\pm$0.1621
		TruthfulQA	Δ0.5249$\pm$0.0658	Δ1.1848$\pm$0.1555	Δ1.1890$\pm$0.1479
	Semantic Uncertainty	MMLU	Δ0.2011$\pm$0.1424	Δ0.6669$\pm$0.2484	Δ0.7894$\pm$0.2090
		SQuAD 2.0	Δ0.3700$\pm$0.0836	Δ1.2180$\pm$0.1775	Δ1.2042$\pm$0.1722
		TruthfulQA	Δ0.6437$\pm$0.0766	Δ1.2246$\pm$0.1344	Δ1.2160$\pm$0.1473
Generalised ECE Mean Reduction	Linguistic Confidence	MMLU	Δ0.0786$\pm$0.0242	Δ0.0655$\pm$0.0258	Δ0.0449$\pm$0.0244
		SQuAD 2.0	Δ0.0394$\pm$0.0224	Δ0.1372$\pm$0.0147	Δ0.1286$\pm$0.0145
		TruthfulQA	Δ0.0725$\pm$0.0252	Δ0.1492$\pm$0.0196	Δ0.1462$\pm$0.0124
	Token Probability	MMLU	Δ0.0966$\pm$0.0272	Δ0.1217$\pm$0.0368	Δ0.1381$\pm$0.0382
		SQuAD 2.0	Δ0.0421$\pm$0.0063	Δ0.1600$\pm$0.0158	Δ0.2257$\pm$0.0245
		TruthfulQA	Δ0.1994$\pm$0.0167	Δ0.1538$\pm$0.0194	Δ0.1561$\pm$0.0147
	Semantic Uncertainty	MMLU	Δ0.1028$\pm$0.0273	Δ0.0974$\pm$0.0309	Δ0.1240$\pm$0.0265
		SQuAD 2.0	Δ0.0540$\pm$0.0323	Δ0.1855$\pm$0.0295	Δ0.1832$\pm$0.0258
		TruthfulQA	Δ0.2065$\pm$0.0184	Δ0.1798$\pm$0.0251	Δ0.1662$\pm$0.0165

E.2 FURTHER INVESTIGATION ON CROSS-DOMAIN CALIBRATION ANOMALIES

Table 2 reveals that cross-domain calibrators occasionally outperform in-domain ones. We investigate this anomaly by examining the miscalibration bias of each dataset, defined as the gap between mean expressed confidence and mean accuracy, which determines how much signal is available for the calibration map to learn from.

We measure the per-dataset miscalibration bias across confidence signals and models (Table 6; Figure 11), and correlate the bias difference between each source–target pair with the observed cross-domain advantage (Figure 12).

Table 6: Mean miscalibration bias (mean confidence $-$ mean accuracy) per dataset and confidence signal, averaged over models.

Dataset	Ling. Conf.	Token Prob.	Sem. Unc.
MMLU	0.049	0.225	0.177
SQuAD 2.0	0.285	0.278	0.326
TruthfulQA	0.360	0.229	0.326

All three datasets are systematically miscalibrated, but the magnitude differs considerably. Figure 12 shows a strong negative relationship between the source–target bias difference and the cross-domain advantage. Transfer pairs whose source and target share a similar miscalibration bias show little

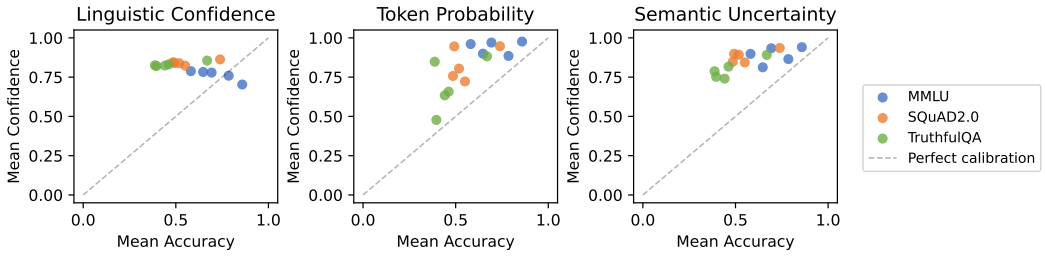

Figure 11: Mean confidence vs. mean accuracy per (dataset, model) pair. All datasets are systematically miscalibrated (above the diagonal), but the magnitude of bias varies considerably across domains.

performance gap relative to in-domain calibration, whilst larger differences tend to favour in-domain calibration.

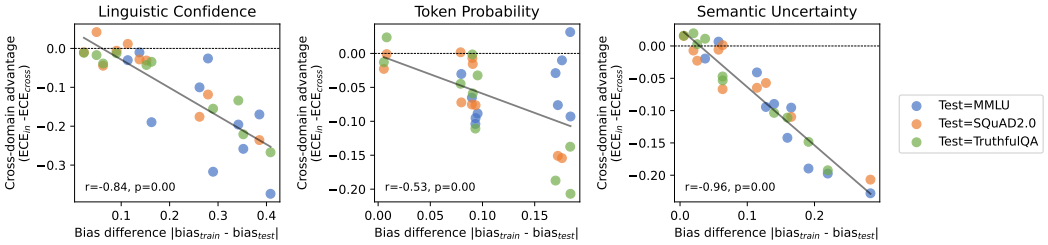

Figure 12: Miscalibration bias difference $|\bar{b}_{\text{train}} - \bar{b}_{\text{test}}|$ vs. cross-domain advantage ($\text{ECE}_{\text{in}} - \text{ECE}_{\text{cross}}$). Colour indicates the test dataset. Transfer pairs with similar miscalibration biases achieve performance closer to in-domain calibration.

This pattern follows from the learning dynamics of the calibration map. When a target domain has a weak bias, the in-domain calibrator has little signal to learn from and fits an unreliable correction. A cross-domain source with a stronger, more consistent bias learns a more decisive correction; provided the two domains share the same direction of miscalibration, this correction transfers effectively even if its magnitude differs. Cross-domain transfer therefore outperforms in-domain calibration precisely when in-domain data is least informative.

E.3 Detailed in-domain calibration vs. Hedged QA comparison

Table 7 reports pre-to-post changes in linguistic-space Faithfulness Divergence and generalised ECE for RALC and the Hedged QA baseline, broken down by dataset, model, and confidence signal. Each cell shows the original Direct QA value alongside the post-intervention value; linguistic confidence is estimated by the LLM ensemble for all three response types. RALC outperforms Hedged QA in expectation across both metrics, with few exceptions due to model-specific signal characteristics, whilst Hedged QA shows limited and inconsistent improvements.

Table 7: Detailed in-domain calibration vs. Hedged QA comparison.

Dataset	Model	Signal	**In-Domain Calibration**		**Hedged QA**	
			Faithfulness Divergence (Orig. → Calib.)	Generalised ECE (Orig. → Calib.)	Faithfulness Divergence (Orig. → Hedged)	Generalised ECE (Orig. → Hedged)
MMLU	Mistral-7B-Inst.	Linguistic Conf.	1.198 → 0.557	0.239 → 0.120	1.198 → 1.227	0.239 → 0.251
		Token Prob.	1.198 → 0.578	0.239 → 0.109		
		Semantic Unc.	1.198 → 0.471	0.239 → 0.113		
	Gemma-4-31B-IT	Linguistic Conf.	1.047 → 1.077	0.284 → 0.124	1.047 → 0.577	0.284 → 0.146
		Token Prob.	1.047 → 0.956	0.284 → 0.093		
		Semantic Unc.	1.047 → 0.530	0.284 → 0.080		
	Llama-3.1-8B-Inst.	Linguistic Conf.	0.925 → 0.602	0.190 → 0.120	0.925 → 0.961	0.190 → 0.199
		Token Prob.	0.925 → 0.665	0.190 → 0.096		
		Semantic Unc.	0.925 → 0.461	0.190 → 0.092		
	GPT-OSS-20B	Linguistic Conf.	0.579 → 1.118	0.152 → 0.128	0.579 → 0.530	0.152 → 0.146
		Token Prob.	0.579 → 1.333	0.152 → 0.138		
		Semantic Unc.	0.579 → 0.563	0.152 → 0.089		
	Qwen3-8B	Linguistic Conf.	0.902 → 0.770	0.167 → 0.148	0.902 → 0.851	0.167 → 0.160
		Token Prob.	0.902 → 0.643	0.167 → 0.113		
		Semantic Unc.	0.902 → 0.564	0.167 → 0.145		
TruthfulQA	Mistral-7B-Inst.	Linguistic Conf.	1.782 → 0.710	0.374 → 0.227	1.782 → 1.654	0.374 → 0.350
		Token Prob.	1.782 → 0.727	0.374 → 0.248		
		Semantic Unc.	1.782 → 0.477	0.374 → 0.238		
	Gemma-4-31B-IT	Linguistic Conf.	1.272 → 0.831	0.209 → 0.118	1.272 → 1.271	0.209 → 0.204
		Token Prob.	1.272 → 0.625	0.209 → 0.100		
		Semantic Unc.	1.272 → 0.469	0.209 → 0.096		
	Llama-3.1-8B-Inst.	Linguistic Conf.	1.828 → 0.712	0.427 → 0.269	1.828 → 1.756	0.427 → 0.394
		Token Prob.	1.828 → 0.636	0.427 → 0.247		
		Semantic Unc.	1.828 → 0.467	0.427 → 0.254		
	GPT-OSS-20B	Linguistic Conf.	2.265 → 0.678	0.392 → 0.230	2.265 → 2.143	0.392 → 0.375
		Token Prob.	2.265 → 0.645	0.392 → 0.227		
		Semantic Unc.	2.265 → 0.475	0.392 → 0.179		
	Qwen3-8B	Linguistic Conf.	2.054 → 0.707	0.442 → 0.270	2.054 → 1.999	0.442 → 0.413
		Token Prob.	2.054 → 0.623	0.442 → 0.242		
		Semantic Unc.	2.054 → 0.464	0.442 → 0.247		
SQuAD 2.0	Mistral-7B-Inst.	Linguistic Conf.	1.978 → 0.701	0.350 → 0.185	1.978 → 1.943	0.350 → 0.343
		Token Prob.	1.978 → 0.682	0.350 → 0.181		
		Semantic Unc.	1.978 → 0.487	0.350 → 0.116		
	Gemma-4-31B-IT	Linguistic Conf.	1.288 → 0.649	0.166 → 0.094	1.288 → 1.300	0.166 → 0.167
		Token Prob.	1.288 → 0.593	0.166 → 0.074		
		Semantic Unc.	1.288 → 0.480	0.166 → 0.103		
	Llama-3.1-8B-Inst.	Linguistic Conf.	1.594 → 0.643	0.305 → 0.160	1.594 → 1.584	0.305 → 0.302
		Token Prob.	1.594 → 0.623	0.305 → 0.145		
		Semantic Unc.	1.594 → 0.495	0.305 → 0.122		
	GPT-OSS-20B	Linguistic Conf.	2.265 → 0.764	0.389 → 0.232	2.265 → 2.214	0.389 → 0.381
		Token Prob.	2.265 → 0.665	0.389 → 0.192		
		Semantic Unc.	2.265 → 0.474	0.389 → 0.137		
	Qwen3-8B	Linguistic Conf.	2.103 → 0.751	0.370 → 0.225	2.103 → 2.033	0.370 → 0.360
		Token Prob.	2.103 → 0.670	0.370 → 0.189		
		Semantic Unc.	2.103 → 0.480	0.370 → 0.176		

E.4 IN-DOMAIN CALIBRATION RELIABILITY DIAGRAMS

Figures 13 and 14 present the in-domain calibration reliability diagrams for MMLU and TruthfulQA, respectively, across confidence signals and models. The left column shows the original linguistic confidence of the Direct QA responses. The rest of the columns show the linguistic confidence of the rewritten responses through RALC guided by different confidence signals, including linguistic confidence (LC), token probability (TP), and semantic uncertainty (SU). Additionally, we report the generalised ECE (Wang et al., 2024) and Faithfulness Divergence (FD) along with the reliability diagrams. The results show that RALC effectively reduces miscalibration across all confidence signals and models.

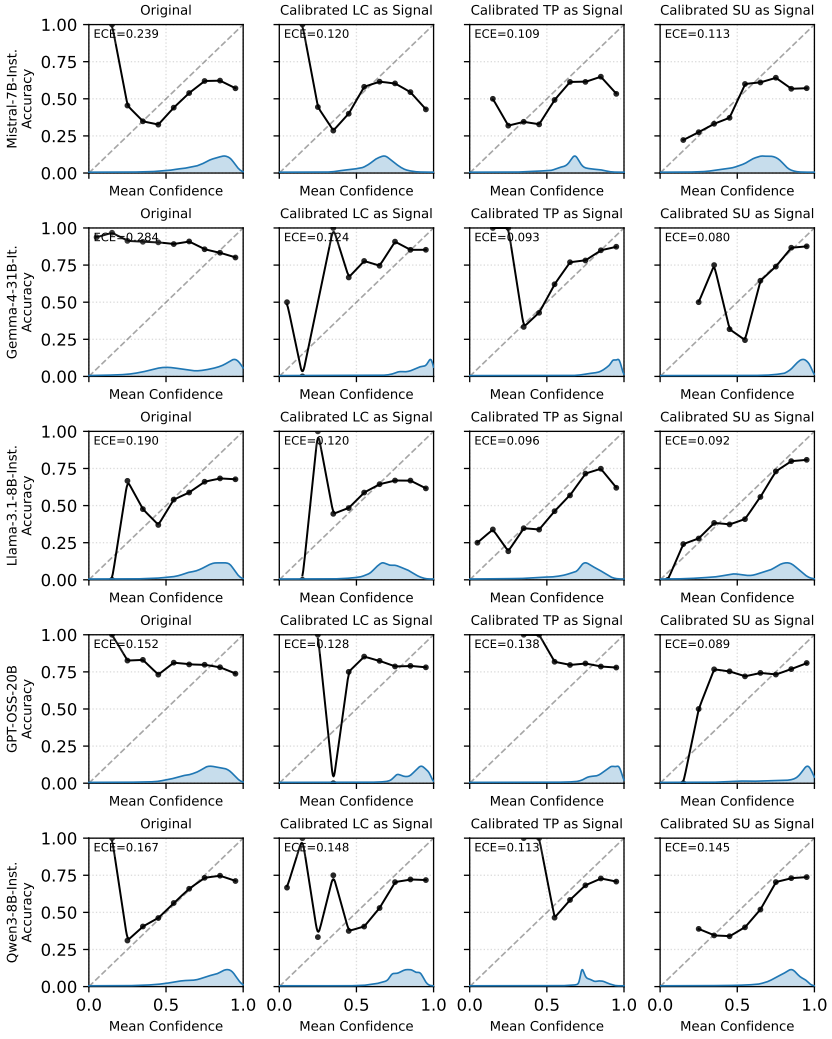

Figure 13: In-domain calibration reliability diagrams for MMLU across confidence signals and models. The left column shows the original linguistic confidence of the Direct QA responses. The rest of the columns show the linguistic confidence of the rewritten responses through RALC guided by different confidence signals, including linguistic confidence (LC), token probability (TP), and semantic uncertainty (SU). We report the generalised ECE (Wang et al., 2024) along with the reliability diagrams. The results show that RALC effectively reduces miscalibration across all confidence signals and models.

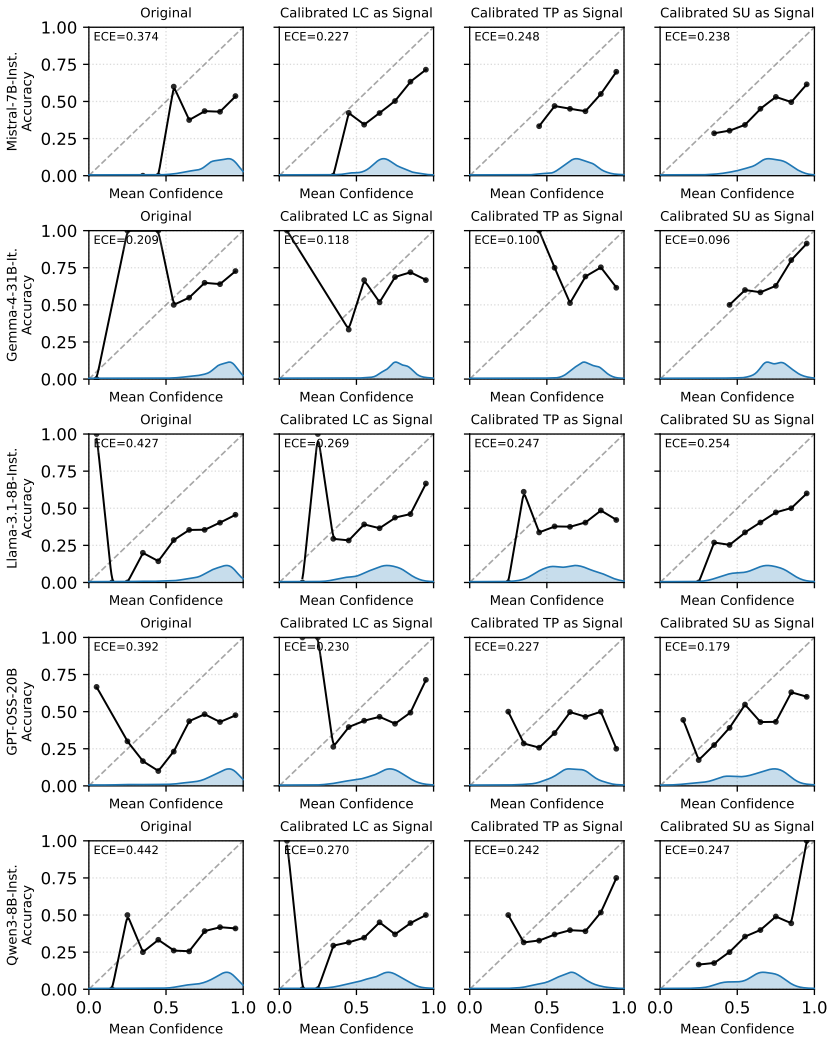

Figure 14: In-domain calibration reliability diagrams for TruthfulQA across confidence signals and models. The left column shows the original linguistic confidence of the Direct QA responses. The rest of the columns show the linguistic confidence of the rewritten responses through RALC guided by different confidence signals, including linguistic confidence (LC), token probability (TP), and semantic uncertainty (SU). We report the generalised ECE (Wang et al., 2024) along with the reliability diagrams. The results show that RALC effectively reduces miscalibration across all confidence signals and models.

E.5 HEAD-TO-HEAD COMPARISON WITH WANG ET AL.

The closest related method, the hedging-word confidence remapping of Wang et al. (2024), has a major coverage limitation: its discrete, word-level remapping process means that the absence of matching hedging language in a response blocks any assessment of calibration or faithfulness improvement for that response. We compare RALC directly against this method on TruthfulQA across four models under matched conditions (30%/70% train/test split; the baseline implemented via regex matching against its published lexicon), spanning our three confidence signals (linguistic confidence, LC; token probability, TP; semantic uncertainty, SU).

Table 8 reports each method's coverage alongside the percentage change in generalised ECE and Faithfulness Divergence. Δ values for the baseline are coverage-weighted, since it only estimates confidence for the under-8% of responses matching its lexicon, so its raw subset-level change is not directly comparable to RALC's full-coverage change. RALC's reductions translate directly into dataset-wide gains across all four models, whereas the baseline's collapse to marginal, or occasionally adverse, effects once its sub-8% coverage is applied, confirming limited coverage as the core limitation of word-level remapping relative to RALC's retrieval-augmented rewriting.

Table 8: Head-to-head comparison between RALC and the hedging-word remapping method of Wang et al. (2024) on TruthfulQA. Coverage denotes the proportion of responses for which the method produces a confidence estimate. Δ ECE and Δ FD report the percentage change relative to the pre-calibration metric; green indicates improvement, red indicates deterioration. Δ values for the baseline are coverage-weighted.

Model	Method	Coverage	Δ ECE	Δ FD
Llama-3.1-8B	LC	100%	✓ 37.0%	✓ 61.1%
	SU	100%	✓ 40.4%	✓ 74.5%
	TP	100%	✓ 42.1%	✓ 65.2%
	Wang et al.	5.4%	✓ 4.29%	✓ 4.65%
Mistral-7B	LC	100%	✓ 39.2%	✓ 60.1%
	SU	100%	✓ 36.5%	✓ 73.2%
	TP	100%	✓ 33.7%	✓ 59.2%
	Wang et al.	4.2%	× 1.65%	✓ 2.04%
GPT-OSS-20B	LC	100%	✓ 41.4%	✓ 70.1%
	SU	100%	✓ 54.5%	✓ 79.0%
	TP	100%	✓ 42.2%	✓ 71.5%
	Wang et al.	7.5%	✓ 0.66%	× 83.79%
Qwen3-8B	LC	100%	✓ 38.9%	✓ 65.6%
	SU	100%	✓ 44.2%	✓ 77.4%
	TP	100%	✓ 45.2%	✓ 69.7%
	Wang et al.	1.6%	✓ 0.79%	✓ 1.38%

F LLM CONFIGURATIONS

All five evaluation targets, the LLM ensemble, and the LLM rewriter in the RALC pipeline are configured with a temperature of 1 to encourage diverse outputs. The LLM cluster selector and grader are configured with a temperature of 0 to encourage deterministic outputs. All models are hosted locally (single RTX 4090 GPU) or through cloud APIs.