

Self-Aware Knowledge Probing: Evaluating Language Models’ Relational Knowledge through Confidence Calibration

Christopher Kissling¹, Elena Merdjanovska^{1,2} and Alan Akbik^{1,2}

¹Humboldt-Universität zu Berlin

²Science of Intelligence

c_kissling@icloud.com, {elena.merdjanovska, alan.akbik}@hu-berlin.de

Abstract

Knowledge probing quantifies how much relational knowledge a language model (LM) has acquired during pre-training. Existing knowledge probes evaluate model capabilities through metrics like prediction accuracy and precision. Such evaluations fail to account for the model’s reliability, reflected in the calibration of its confidence scores. In this paper, we establish an empirical framework to evaluate the calibration of relational knowledge in a closed answer-set setting. We cover three modalities of model confidence: (1) intrinsic confidence, (2) paraphrase consistency and (3) linguistic confidence. Our extensive analysis of 21 causal and 8 masked language models reveals that most models, but especially those pre-trained with the masking objective, are overconfident. The best-calibrated scores come from confidence estimates that account for inconsistencies due to statement rephrasing. Moreover, calibration improves when increasing the number of answer options.

1 Introduction

Knowledge probes serve as the primary diagnostic tool for quantifying the relational knowledge internalized by language models during pre-training. By querying models with factual triplets—typically formatted as cloze-style statements such as “Balach Marri died in [MASK]”—researchers can assess a model’s grasp of diverse semantic relations (Petroni et al., 2019; Youssef et al., 2023). Such evaluations are critical for understanding how internalized knowledge correlates with performance on complex downstream reasoning tasks.

However, existing knowledge probing frameworks (Youssef et al., 2023; Kalo and Fichtel, 2022; Wiland et al., 2024) focus almost exclusively on attainment: the accuracy and precision with which

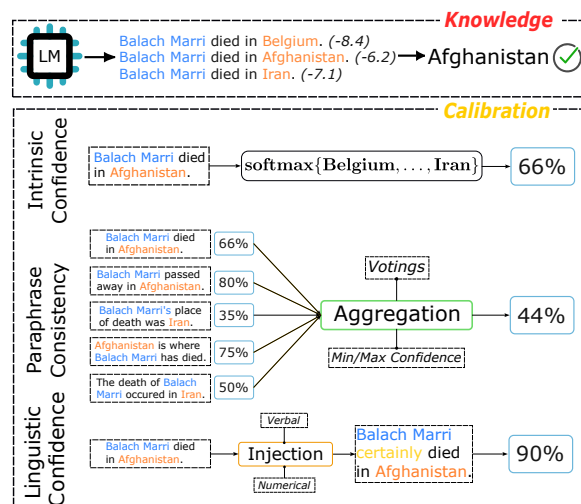


Figure 1: Using knowledge facts (like the place of death of a particular person), we measure an LM’s awareness of its relational knowledge in three modalities: (1) intrinsic confidence, (2) paraphrase consistency of predictions across multiple rephrasings of the same statement, and (3) linguistic confidence using epistemic markers.

a model retrieves a fact. While accurate recall is imperative, it provides an incomplete picture of model reliability. A truly reliable model must be “self-aware”, exhibiting confidence scores that are well-calibrated with its probability of being correct (Guo et al., 2017). In a perfectly calibrated system, a confidence score of 0.8 implies an 80% likelihood of factual correctness. Without such alignment, even high-accuracy models are prone to silent failures and overconfident hallucinations, as they lack the mechanisms to signal the limits of their own knowledge.

Knowledge calibration probing. In this paper, we empirically evaluate calibration techniques for relational knowledge. Unlike prior work that treats knowledge as a binary state of presence or absence, we evaluate the epistemic self-awareness of language models through calibration. We examine three distinct modalities of model confi-

<https://github.com/diamir707/knowledge-calibration>

dence: (1) *intrinsic confidence*, derived from raw log-probabilities; (2) *paraphrase consistency*, measured through agreement across semantically equivalent statements; and (3) *linguistic confidence*, evaluated through the model’s response to explicit linguistic markers of uncertainty (see Figure 1).

This multi-faceted approach allows for a granular analysis of how model architecture and scale influence the reliability of factual retrieval. Specifically, we leverage a closed-set probing setup (Wiland et al., 2024) to facilitate a direct, head-to-head comparison between causal language models (CLMs) and masked language models (MLMs). While these architectures differ fundamentally in their training objectives, our framework provides a unified environment to evaluate their respective epistemic self-awareness. To ensure our findings are robust across diverse knowledge domains and model families, we evaluate 23 distinct models ranging from 109M to 12B parameters. Among CLMs, we include base-variants and its corresponding instruction-tuned versions.

Contributions. Our primary contributions include:

- We provide the first large-scale comparison of calibration between CLMs and MLMs, analyzing 23 models to isolate the impact of pre-training objectives on epistemic awareness.
- We evaluate the efficacy of six confidence estimation strategies—ranging from raw log-probabilities to multi-prompt aggregations.
- We investigate the impact of explicit linguistic confidence expressions on model reliability.

Our analysis reveals a critical "reliability gap" in current architectures; specifically, that MLMs—despite competitive factual accuracy—remain fundamentally less aware of their own knowledge than CLMs. Furthermore, we found a clear dependency between the number of distractors and calibration, showing that LMs need to be exposed to a sufficiently large number of answer options to achieve good calibration. Finally, our results suggest a clear scale-dependency for linguistic confidence: while smaller models are "semantically blind" to expressions of doubt, larger models begin to align their internal confidence with linguistic cues. To facilitate further research into model reliability, we provide our proposed framework as an extensible benchmark for evaluating the epistemic awareness of future language models.

2 Related Work

The evaluation of model reliability in factual retrieval lies at the intersection of two research areas: uncertainty quantification for language models and relational knowledge probing.

2.1 Confidence Estimation for LMs

Confidence estimation methods can be broadly categorized into three modalities, mirroring the evaluation framework we propose in this paper.

Intrinsic confidence. The most direct approach utilizes the model’s internal probability distribution. Early work on calibration in neural networks (Guo et al., 2017) focused on softmax-normalized scores, a technique recently extended to LMs for tasks like hallucination detection (Huang et al., 2025) and selective prediction (Duan et al., 2024). Our work specifically examines how these intrinsic signals behave when the answer space is constrained to a fixed set of factual candidates.

Paraphrase consistency. This modality relies on the intuition that a model “knows” a fact if it provides the same answer across various semantically equivalent formulations. Techniques such as *Self-Consistency* (Wang et al., 2023) and *Semantic Entropy* (Kuhn et al., 2023) leverage multiple model outputs to derive uncertainty. While consistency is often used as a proxy for reliability in knowledge bases (Zheng et al., 2024), our framework evaluates whether these consistency signals are well-calibrated with factual truth.

Linguistic confidence. A growing body of research explores “self-reported” or “verbalized” confidences, where the model expresses its certainty through natural language (Tian et al., 2023). Previous studies have found that LMs are often reluctant to express uncertainty (Zhou et al., 2024) or fail to align their verbalized scores with their internal log-probabilities (Xiong et al., 2024). We extend this line of research to relational knowledge statements by injecting explicit epistemic markers (e.g., *certainly*, *possibly*) into the probing templates.

2.2 Relational Knowledge Probing

The field of knowledge probing was pioneered by the LAMA benchmark (Petroni et al., 2019), which used cloze-style prompts to quantify the factual knowledge of MLMs. Subsequent benchmarks like KAMEL (Kalo and Fichtel, 2022) and Myriad-LAMA (Zhao et al., 2024) expanded the diversity of relations and entities.

Benchmarks with cloze-style prompts provide a highly controlled environment to probe for relational knowledge. Unlike knowledge benchmarks in a question-answering format (Lin et al., 2022b; Hendrycks et al., 2021), which introduce confounding variables like fluency or style (Rogers et al., 2023), cloze-style probes minimize linguistic uncertainties. This allows us to specifically isolate an LM’s core factual capabilities.

2.3 The BEAR Framework

The BEAR probe (Wiland et al., 2024) establishes a unified knowledge probing mechanism for both CLMs and MLMs. Given a relation instance—such as (BALACH MARRI, died-in, AFGHANISTAN)—BEAR provides a set of K candidate objects, comprising the true answer and plausible distractors (e.g. IRAN). Using a relation-specific template, a statement is generated for each candidate—for instance, *Balach Marri died in Afghanistan* and *Balach Marri died in Iran*. The probe ultimately tests whether an LM can distinguish the true fact from the distractors by calculating the full log-likelihood of each statement.

Scorings. For CLMs, a sentence-level log-likelihood is simply the sum of the token log-likelihoods in the sequence:

$$\log P(x) = \sum_{t=1}^{|x|} \log P(x_t \mid x_{<t})$$

For MLMs, BEAR utilizes pseudo log-likelihoods (PLLs) (Salazar et al., 2020). We adopt the refinement by Kauf and Ivanova (2023), which masks tokens one-by-one and accounts for multi-token words by masking all future within-word tokens:

$$\text{PLL}(x) = \sum_{t=1}^{|x|} \log P(x_t \mid x_{\setminus t})$$

Answer selection. Once a log-likelihood score ℓ_i is assigned to every candidate statement o_i in the answer set, the model’s prediction $\hat{y}$ is the object that maximizes this score: $\hat{y} := \arg \max_{i \in \{1, \dots, K\}} \ell_i$.

Dataset Characteristics. The dataset contains facts found in Wikidata, covering the following domains: *Biographical*, *Arts*, *Geographic*, *Movies*, *Economic*, *Historical*, *Sports*, *Scientific* and *Political*. It should be noted that BEAR focuses on information prevalent enough to likely appear in pre-training corpora beyond Wikipedia. The resulting potential for test-set contamination is not

necessarily a limitation; rather, for the purposes of knowledge probing, such exposure is expected. Finally, BEAR avoids key limitations of earlier probes, such as restrictions to single-subtoken answers, imbalanced answer distributions, and reliance on rare Wikidata entries.

3 Confidence Estimation

In this section, we present the confidence estimates we calculated in order to assess calibration. For this, we first normalize the answer log-likelihoods using the softmax function (σ), yielding probabilities which can be interpreted as the model’s confidence in the correctness of a given answer (Figure 1).

3.1 Intrinsic Confidence

Estimates in this category use a single template to calculate each instance’s log-likelihood.

Base-Confidence. We define a baseline estimate as the maximum value after softmax normalization:

$$C_{\text{Base}}(\hat{y}) := \max_{i \in \{1, \dots, K\}} \sigma(\ell)_i \quad (1)$$

where ℓ represents the vector of log-likelihoods of the answer options. We also used this baseline estimate to find the token log-likelihood reduction strategy which maximizes accuracy and minimizes calibration error (Appendix C).

Margin-Confidence. Another single-template estimate is the confidence margin (Lin et al., 2022c), also commonly used in the context of image classification (Pleiss et al., 2020; Joshi et al., 2009). With *Margin-Confidence*, we compute the difference between the largest and second-largest value returned by the softmax, reflecting whether the model’s answer is clearly preferred over the second-best answer:

$$C_{\text{Margin}}(\hat{y}) := \sigma(\ell)_{[1]} - \sigma(\ell)_{[2]} \quad (2)$$

Here, the subscripts [1] and [2] refer to the largest and second-largest values returned by the softmax.

3.2 Paraphrase Consistency

Paraphrase consistency estimates rely on multiple rephrasings of each relation-specific template.¹ Repeating the probe across the templates yields multiple answers $\tilde{y}_1, \dots, \tilde{y}_5$, requiring an aggregation step to obtain a single prediction $\hat{y}$ (Figure 1).

¹Different from consistency often mentioned in knowledge editing literature (Dong et al., 2025), where it refers to logical consistency.

Answer Aggregation. The answers can be aggregated through voting schemes or confidence-based aggregation. Voting schemes require a minimum level of agreement (for example, at least 2 answers need to be the same). If agreement is insufficient, the prediction is rejected, yielding no answer and zero calibration error by design. Confidence-based aggregation always produces a final answer, which in some cases can rely only on a single candidate. In our experiments, we use confidence-based aggregation, selecting the predicted answer with minimum C_{Base} .

Average-Confidence. We define *Average-Confidence* as the per-template average of the selected answer:

$$C_{\text{Average}}(\hat{y}) := \frac{\sum_{i=1}^5 \mathbb{1}(\tilde{y}_i = \hat{y}) \cdot C_{\text{Base}}(\tilde{y}_i)}{5} \quad (3)$$

Consistency-Confidence. Instead of averaging the softmax scores, confidence can be reflected through the relative frequency of the answers:

$$C_{\text{Consistency}}(\hat{y}) := \frac{\sum_{i=1}^5 \mathbb{1}(\tilde{y}_i = \hat{y})}{5} \quad (4)$$

Hence, $C_{\text{Consistency}}$ bypasses the step of softmax normalization and is discretely distributed. It therefore more closely aligns with how humans naturally express confidence, typically in multiples of five (Xiong et al., 2024).

Entropy-Confidence. The certainty of the predicted answer can also be measured through the entropy over the predicted answer distribution. For each unique predicted answer $\tilde{y}_j \in \tilde{Y}$, we compute the unnormalized average C_{Base} :

$$s(\tilde{y}_j) := \frac{\sum_{i=1}^5 \mathbb{1}(\tilde{y}_i = \tilde{y}_j) \cdot C_{\text{Base}}(\tilde{y}_i)}{\sum_{i=1}^5 \mathbb{1}(\tilde{y}_i = \tilde{y}_j)} \quad (5)$$

Then, we normalize it to obtain a confidence distribution:

$$q(\tilde{y}_j) := \frac{s(\tilde{y}_j)}{\sum_{\tilde{y}_k \in \tilde{Y}} s(\tilde{y}_k)} \quad (6)$$

Since the Shannon entropy is a measure of certainty, we subtract it from 1 to obtain our entropy confidence estimate:

$$C_{\text{Entropy}}(\hat{y}) := 1 - \left(- \sum_{\tilde{y}_j \in \tilde{Y}} q(\tilde{y}_j) \log q(\tilde{y}_j) \right) \quad (7)$$

We base this approach on the widely-used semantic entropy (Kuhn et al., 2023). In our approach,

the semantic clusters correspond to the predicted unique answers. Hence, with five templates, there is a maximum of five clusters per instance.

3.3 Linguistic Confidence

To assess linguistic confidence calibration, we evaluate the impact of *verbalized* and *numerical* confidence expressions on both accuracy and calibration. To achieve this, we modify the templates by injecting verbalized and numerical confidence expressions (Zhou et al., 2023). As a confidence score, we use the single-template C_{Base} (Equation 1).

Verbalized. For the verbalized setup, we inject the epistemic marker *possibly* to lower the confidence of the statement and the marker *certainly* to strengthen it. For instance, *Balach Marri certainly died in Afghanistan*.

Numerical. To evaluate the impact of numerical confidence expressions in relational knowledge statements, we modified the relation templates to include formulations such as “*I’m $x\%$ confident that Balach Marri died in Afghanistan*.”, with $x \in \{0, 25, \dots, 100\}$.

4 Experimental Setup

In this section, we give an overview of the datasets, models and evaluation metrics we used.

4.1 Datasets

We base our calibration probe around the BEAR framework (Wiland et al., 2024). The dataset contains 7,731 instances across 60 relations. Each relation has three semantically equivalent template statements. As part of this work, we extended the number of templates to five per relation using ChatGPT (see Appendix B). Out of all relations, 46 have cardinality N:1 (different subjects can share the same answer) and 14 have cardinality 1:1 (each object is the answer to exactly one subject). For the N:1 relations, the average number of answer options per instance is 6.5. Each 1:1 relation consists of 60 instances, with 60 answer options per instance.

Additionally, we run experiments with the Wiki subset of the FACTOR (Muhlgay et al., 2024) dataset, which has one correct and three incorrect completions for each statement.

4.2 Language Models

We evaluated 16 different models, 10 CLMs and 6 MLMs (listed in Appendix A). The MLMs range

Model	ACE ($\downarrow$)				Brier Score ($\downarrow$)				
	C_{Base}	C_{Margin}	C_{Average}	C_{Entropy}	C_{Base}	C_{Margin}	C_{Average}	$C_{\text{Consistency}}$	C_{Entropy}
gemma-2b	0.113	0.053	0.032	0.150	0.162	0.155	0.098	0.157	0.155
gemma-7b	0.077	0.046	0.042	0.172	0.133	0.134	0.083	0.124	0.144
opt-125m	0.299	0.153	0.116	0.106	0.241	0.186	0.130	0.273	0.159
opt-350m	0.265	0.139	0.098	0.109	0.224	0.177	0.127	0.263	0.166
opt-1.3b	0.207	0.100	0.056	0.118	0.214	0.187	0.126	0.232	0.177
opt-6.7b	0.154	0.069	0.024	0.138	0.187	0.168	0.110	0.191	0.163
gemma-3-270m (IT)	0.395	0.235	0.125	0.128	0.303	0.218	0.116	0.220	0.165
gemma-3-1b (IT)	0.350	0.210	0.092	0.198	0.284	0.217	0.108	0.187	0.226
gemma-3-4b (IT)	0.267	0.162	0.053	0.338	0.240	0.193	0.099	0.147	0.329
gemma-3-12b (IT)	0.193	0.117	0.053	0.436	0.185	0.154	0.084	0.117	0.370
Llama-3.2-1B (IT)	0.170	0.072	0.036	0.288	0.190	0.167	0.110	0.189	0.290
Llama-3.2-3B (IT)	0.129	0.057	0.039	0.367	0.167	0.153	0.098	0.150	0.329
Llama-3.1-8B (IT)	0.086	0.040	0.048	0.433	0.130	0.126	0.076	0.110	0.316
Qwen3.5-0.8B (IT)	0.247	0.118	0.083	0.162	0.218	0.176	0.113	0.247	0.204
Qwen3.5-2B (IT)	0.203	0.087	0.049	0.218	0.201	0.172	0.112	0.222	0.250
Qwen3.5-4B (IT)	0.152	0.058	0.040	0.297	0.172	0.153	0.098	0.165	0.283
Qwen3.5-9B (IT)	0.114	0.047	0.034	0.371	0.155	0.146	0.095	0.153	0.320
bert-base-cased	0.454	0.295	0.218	0.129	0.367	0.276	0.205	0.338	0.247
bert-large-cased	0.452	0.302	0.198	0.125	0.373	0.285	0.196	0.310	0.223
roberta-base	0.482	0.321	0.165	0.115	0.387	0.291	0.162	0.265	0.174
roberta-large	0.441	0.291	0.144	0.124	0.367	0.278	0.158	0.254	0.183
xlm-roberta-base	0.585	0.427	0.274	0.104	0.481	0.368	0.225	0.362	0.246
xlm-roberta-large	0.559	0.412	0.249	0.119	0.461	0.359	0.224	0.354	0.259
ModernBERT-base	0.462	0.303	0.143	0.128	0.372	0.279	0.147	0.248	0.194
ModernBERT-large	0.400	0.254	0.116	0.153	0.332	0.255	0.151	0.242	0.220

Table 1: *ACE* and *Brier Score* of our estimates. **Bold values** indicate the best-performing estimate per model. **ACE:** • [0,0.05), • [0.05,0.10), • [0.10,0.20), • ≥ 0.20 ; **Brier Score:** • [0–0.10), • [0.10–0.20), • [0.20–0.25), • ≥ 0.25 . A *Brier Score* of 0.25 corresponds to a *random baseline*, always predicting a confidence of 0.5. For *ACE*, we consider values < 0.1 *well-calibrated*, where the confidence deviates from accuracy by 10 percentage points at most. We do not report *ACE* for $C_{\text{Consistency}}$ due to its discrete distribution.

from a size of 109 million parameters to 561 million parameters. For the CLMs, the smallest model is the opt-125m with 125 million parameters, and the largest is gemma-12b(IT) with 12 billion parameters. We include examples of both base and instruction-tuned CLMs.

4.3 Evaluation Metrics

Adaptive Calibration Error. Due to the limitations of the commonly-used Expected Calibration Error (ECE) (Naeini et al., 2015) with fixed bin widths, we use the *Adaptive Calibration Error* (ACE) (Nixon et al., 2020). ACE accounts for the underlying confidence distributions by defining adaptive bin intervals so that each contains an equal number of predictions. Low ACEs indicate high knowledge awareness.

Calibration Curves. To assess over- and under-confidence, we rely on calibration curves for which

we plot the confidence scores against accuracy, i.e. the fractions of correct predictions in each confidence bin. For both calculating ACE and plotting the calibration curves, we use quantile binning with 20 bins.

Brier Score. We include the *Brier Score* (Brier, 1950) to reflect both calibration and discriminative power. Unlike ACE, which would be zero for a random guessing binary classifier that always predicts 50% confidence, the Brier Score penalizes such uninformative estimates (Tian et al., 2023). We use correctness labels $1(y_i = \hat{y}_i)$ as a target (Tian et al., 2023; Lin et al., 2022a).

5 Experiment 1: Calibration Evaluation

This experiment evaluates the calibration of the confidence estimates for each model.

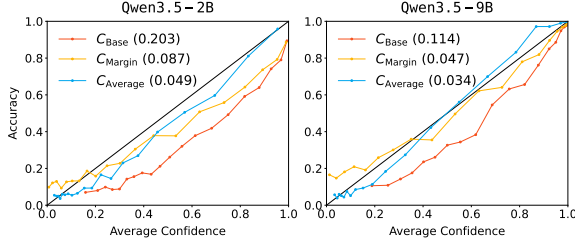


Figure 2: Calibration curves of our continuous confidence estimates. *ACE* in parentheses.

5.1 Comparison of Confidence Estimates

Table 1 presents the *ACEs* and the *Brier Scores* for all models and each confidence estimate. Figure 2 shows the calibration curves for selected models.

C_{Margin} scores are more reliable than C_{Base} . From the intrinsic confidence estimates, derived using only a single template, C_{Margin} emerged as a consistently well-calibrated method across all LMs. With C_{Margin} , gemma-7b, Llama-3.1-8B(IT) and Qwen3.5-9B(IT) achieve an *ACE* of under 0.05, indicating good calibration and reliability. This efficacy is further illustrated by the calibration curves in Figure 2, where we see that C_{Base} is markedly overconfident. On the other hand, the C_{Margin} curves align more closely with the identity line, which signifies perfect calibration.

For all models included in Figure 2, C_{Base} does not produce enough scores in the lower confidence ranges to derive a calibration curve that covers the full $[0, 1]$ interval. C_{Margin} , however, does extend the curve into these lower confidence regions without limiting the model’s ability to output enough high-confidences, allowing for evaluation across the entire range. We also observe that C_{Margin} can introduce underconfidence in the lower range of the confidence scores, pushing the calibration curve above the identity line. From a practical perspective, an underconfident model is generally preferable than an overconfident one.

We confirm the benefits of C_{Margin} through the Wiki-FACTOR dataset. In Table 4 in Appendix F, we see that C_{Margin} consistently outperforms C_{Base} across all models and both evaluation metrics, highlighting it as a more reliable confidence estimate.

C_{Average} is the best-calibrated estimate. The paraphrase consistency estimate C_{Average} clearly outperforms the other approaches across all models in terms of both *Brier Score* and *ACE*. For example, for opt-6.7b, C_{Average} achieves an *ACE* of 0.024, and for Qwen3.5-9B(IT), an *ACE* of 0.034,

which is close to perfect calibration. This is reflected in the calibration curve shown in Figure 2, which deviates only slightly from the identity line. We further note that for larger models (for example, gemma-7b and Qwen3.5-9B(IT)), the difference between C_{Average} and C_{Margin} is often small, with C_{Margin} providing better-aligned scores than C_{Average} for Llama-3.1-8B(IT).

The other paraphrase consistency estimates: $C_{\text{Consistency}}$ and C_{Entropy} provide very miscalibrated scores with low discriminative power, showing that simple answer agreement is not enough to capture model confidence.

The largest models are also most calibrated.

For most model families, we see that by increasing model size, *ACEs* and *Brier Scores* decrease. This applies to all confidence estimation methods. The model families with the overall lowest *ACE* and *Brier Scores* are LLaMA and gemma, which indicates that these models’ confidences best reflect their factual knowledge. Newer gemma-3 models exhibit lower accuracy (see Appendix A) and worse calibration than the original gemma family. In general, all models larger than 1B parameters, achieve an $ACE \leq 0.1$ with their best estimate, which we consider well-calibrated. Among the MLMs, we found that the ModernBERT’s C_{Base} confidences are comparable to BERT and RoBERTa, however with C_{Average} ModernBERT is better-calibrated. Overall, xlm-roberta-large deviates from the other MLMs with worse accuracy and calibration, demonstrating the negative impact of multi-lingual training data.

5.2 Impact of Answer Set Size

This section investigates the influence of distractor count on intrinsic confidence. We analyze the average C_{Base} confidence, accuracy and *ACE* as a function of the number of answer options, sampled as subsets of the 60 options in the 1:1 relations in BEAR. Figure 3 shows these relationships for selected models, and we observed similar patterns for the other models of each type.

More answer options lead to better calibration. We find that increasing the number of answer options (i.e. more incorrect answers as distractors) significantly reduces calibration error in CLMs smaller than 1B (see Qwen3.5-0.8B and gpt2-medium in Figure 3). While both average confidence and accuracy decline as options increase, they do so at different rates: confidence drops faster than accuracy, reducing *ACE*. In

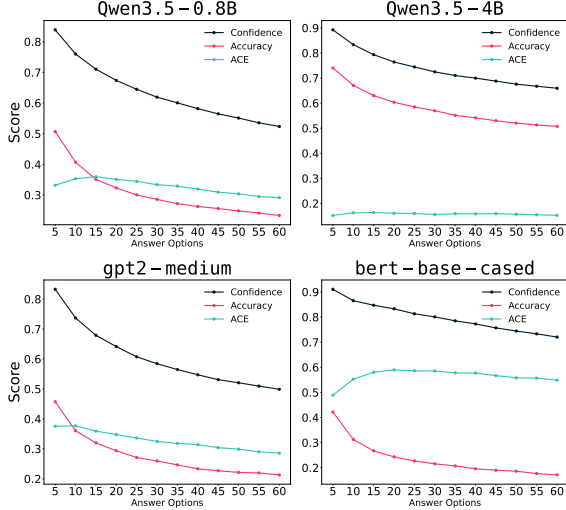


Figure 3: Average C_{Base} confidence, accuracy and ACE as a function of the number of answer options (including the correct answer). Each point represents the mean over three repeated samplings.

larger CLMs and all MLMs this effect is less pronounced, but still present. For Qwen3.5-4B and bert-base-cased, accuracy drops faster than confidence at low answer counts, thereby increasing ACE . After 15–20 answer options, ACE stabilizes, with no further calibration gains for larger CLMs and MLMs. Therefore, robust confidence estimation for both model types requires a fairly high number of distractors, albeit at a higher computational cost.

We confirm these findings through the Wiki-FACTOR benchmark. While FACTOR initially exhibits significantly higher calibration errors than BEAR (see Table 4 in Appendix F), we primarily attribute this discrepancy to FACTOR’s limited four-option answer format. To demonstrate this, in Figure 4 we plot ACE against $Accuracy$ for FACTOR and for the 1:1 relations of BEAR with five sampled answer options, i.e. one more than on FACTOR. We find that matching the number of answer options across datasets recovers a nearly identical relationship between accuracy and ACE . Hence, for identical accuracies, we find similar ACE values on BEAR and FACTOR. For example, gpt2-large achieves an accuracy of 48.97% on BEAR with an ACE of 0.33, while opt-6.7b has a comparable accuracy on FACTOR of 49.33% and a nearly identical ACE of 0.323. Similar parity is observed for gpt2-x1 and gpt2, confirming that calibration is impacted by the distractor count.

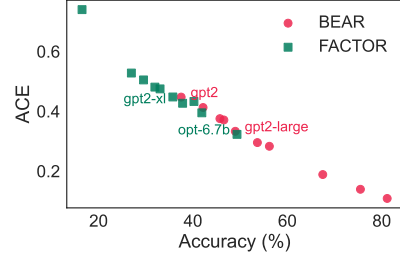


Figure 4: $Accuracy$ against ACE of C_{Base} for the 1:1 relations of BEAR and the Wiki-FACTOR dataset. For BEAR, we sampled the number of answer options to five per instance. On FACTOR, each question consists of four answer options.

Model	Acc.	C_{Margin}	C_{Average}
opt-125m (125M)	15.6	0.186	0.130
roberta-base (125M)	16.0	0.291	0.162
gpt2-medium (355M)	18.4	0.179	0.110
bert-base-cased (109M)	18.1	0.276	0.205
gpt2-large (812M)	21.5	0.163	0.116
roberta-large (355M)	21.1	0.278	0.158
opt-350m (350M)	19.0	0.177	0.127
bert-large-cased (335M)	19.3	0.285	0.196
gemma-3-270m-it (270M)	15.5	0.224	0.116
ModernBERT-base (149M)	16.8	0.289	0.152
Qwen3.5-0.8B (800M)	21.6	0.182	0.115
ModernBERT-large (335M)	22.0	0.262	0.148

Table 2: *Brier Scores* ($\downarrow$) for matched CLM/MLM pairs. Accuracy (in %) is the average over all templates.

5.3 Comparison of CLMs and MLMs

To examine the differences in calibration between the two pre-training objectives, we matched CLMs and MLMs with similar accuracies. The goal was to minimize the influence of differing accuracies on calibration. The matched model pairs are listed in Table 2, together with the *Brier Scores* of the best-performing estimates C_{Margin} and C_{Average} .

Better Calibration of CLMs. Although MLMs with the same or even lower number of parameters can achieve a similar, and in some cases even higher accuracy, we find that CLMs are much better calibrated using most of our estimates (see Table 2). Across all five model pairs, we observe substantial calibration differences, often reaching differences of approximately 0.1 to 0.2 for C_{Margin} and C_{Average} . For C_{Margin} and C_{Average} , all CLMs achieve a *Brier Score* of 0.2 or lower, while none of the MLMs do. **Softmax MLM scores do not reflect model confidence.** MLMs do not explicitly model a sentence-level log-likelihood during their pre-training phase

gemma-3-1b-pt	0.006	0.038	0.022	0.015	0.012	0.011	0.021
gemma-3-1b-it	-0.014	0.003	0.033	0.019	0.024	0.016	0.021
gemma-3-4b-pt	-0.005	0.022	0.019	0.011	0.011	0.004	0.026
gemma-3-4b-it	-0.018	-0.008	0.014	-0.003	-0.005	-0.007	0.003
gemma-3-12b-pt	-0.008	0.018	0.008	-0.010	-0.003	-0.003	0.023
gemma-3-12b-it	-0.007	-0.005	0.031	0.006	0.009	-0.001	0.005
Qwen3.5-0.8B	-0.000	0.018	0.027	0.027	0.022	0.027	0.019
Qwen3.5-2B	-0.006	0.043	0.041	0.037	0.026	0.031	0.032
Qwen3.5-4B	-0.023	0.012	0.004	0.001	0.002	0.006	0.006
Qwen3.5-9B	-0.020	0.020	0.013	-0.000	0.000	0.006	0.018
ModernBERT-base	-0.013	0.004	0.005	0.021	-0.009	-0.004	-0.020
ModernBERT-large	-0.016	-0.012	0.009	0.004	-0.015	-0.017	-0.013
Average	-0.010	0.013	0.019	0.011	0.006	0.006	0.012
	possibly	certainly	0%	25%	50%	75%	100%

Figure 5: Difference in ACE between the non-injected and injected templates. The x-axis shows the applied injection. Negative values (green) show an improvement.

and we rely on the pseudo log-likelihoods, as explained in subsection 2.3. While this scoring is sufficient to distinguish between correct and incorrect answers for the BEAR probe, we believe that the absence of a well-defined sentence-level log-likelihood is the main reason behind the large calibration discrepancies between CLMs and MLMs.

6 Experiment 2: Impact of Linguistic Confidence Expressions

This experiment examines the impact of adding epistemic markers to the statements, by injecting both verbal and numerical confidence expressions. **Injections show inconsistent calibration effects.** We find that verbal weakeners such as *possibly* improve calibration more consistently than numerical weakeners such as “*I’m 0% confident...*”, which exhibit lower and more variable effects across models. One possible explanation is that such expressions are less common in the training data and therefore less reliably interpreted as signals of uncertainty. We also note that instruction-tuning does not systematically shift log-likelihoods in line with the intended meaning of the injections: for gemma-3-4b-it and gemma-3-12b-it, *certainly* reduces overconfidence and thereby lowers ACE .

Verbal markers shift calibration curves. We confirm the expected positive impact of *possibly* on ACE by plotting calibration curves for the base and instruction-tuned variants of gemma-3-4b and

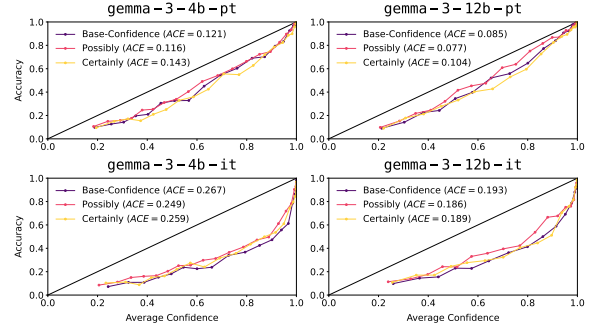


Figure 6: Calibration curves of C_{Base} derived from the original and verbal injected templates.

gemma-3-12b (Figure 6). Compared to the unaltered Base-Confidence, the injection of *possibly* shifts the calibration curve closer to the identity line, thereby reducing ACE . We also observe the somewhat unexpected finding for the instruction-tuned models, where the calibration curve under *certainly* can lie above the curve of the non-injected baseline.

7 Conclusion

This study provides a comprehensive empirical analysis of the reliability of LMs’ relational knowledge through the lens of calibration. Our evaluation of three distinct confidence estimation modalities—intrinsic confidence, paraphrase consistency and linguistic confidence—reveals that obtaining well-calibrated confidences for relational knowledge is challenging. Our findings strongly indicate that LMs lack a deeper understanding of their own knowledge boundaries; regardless of the pre-training objective and model, logit-based intrinsic confidences are consistently overconfident, a tendency more pronounced in MLMs. We demonstrate that the prediction margin can serve as a confidence signal superior to standard top-label confidence, and that calibration error significantly diminishes as the number of answer options increases.

Confidence estimates based on paraphrase consistency yield the best calibration, which reveals the LMs’ high sensitivity to formulation changes. If this sensitivity is not incorporated in the confidence score, the LMs remain largely unreliable from a calibration perspective, especially for low-accuracy models. Since accounting for this sensitivity in practice requires costly prompt perturbations, the self-awareness of LMs in their factual knowledge remains limited.

Limitations

Following are some limitations in our results that we want to address.

For our purposes, quantifying calibration with respect to the top answer is sufficient, since BEAR contains no N:M relations. However, a complete assessment of reliability would also account for such relations. Calibration must then be evaluated over all answer options, for example using the *Brier Score* with one-hot-encoded ground-truth labels y_i or the Static Calibration Error (Nixon et al., 2020).

Our calibration probing serves mainly as a diagnostic tool and is quite specific to closed-answer set scenarios. The calibration error values are therefore unlikely to transfer directly to more practical settings. Such scenarios typically lack predefined answer options, particularly across different formulations, necessary to derive paraphrase consistencies. Moreover, the predefined answer options limit our evaluation to output uncertainty, but unreliable responses can also stem from input uncertainty, such as vaguely formulated prompts.

Acknowledgments

Elena Merdjanovska and Alan Akbik are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC 2002/2 “Science of Intelligence” – project number 390523135. Alan Akbik is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Emmy Noether grant “Eidetic Representations of Natural Language” (project number 448414230).

References

- Glenn W. Brier. 1950. [Verification of forecasts expressed in terms of probability](#). *Monthly Weather Review*, 78(1):1–3.
- Zilu Dong, Xiangqing Shen, Zinong Yang, and Rui Xia. 2025. [ChainEdit: Propagating ripple effects in LLM knowledge editing through logical rule-guided chains](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13558–13571, Vienna, Austria. Association for Computational Linguistics.
- Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. [Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models](#). *Preprint*, arXiv:2307.01379.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). *Preprint*, arXiv:1706.04599.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). *Preprint*, arXiv:2009.03300.
- Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025. [Look before you leap: An exploratory study of uncertainty analysis for large language models](#). *IEEE Transactions on Software Engineering*, 51(2):413–429.
- Ajay J. Joshi, Fatih Porikli, and Nikolaos Panikolopoulos. 2009. [Multi-class active learning for image classification](#). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379.
- Jan-Christoph Kalo and Leandra Fichtel. 2022. [Kamel: Knowledge analysis with multitoken entities in language models](#). In *Automated Knowledge Base Construction*.
- Carina Kauf and Anna Ivanova. 2023. [A better way to do masked language model scoring](#). *Preprint*, arXiv:2305.10588.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *The Eleventh International Conference on Learning Representations*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. [Teaching models to express their uncertainty in words](#). *Preprint*, arXiv:2205.14334.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. [Truthfulqa: Measuring how models mimic human falsehoods](#). *Preprint*, arXiv:2109.07958.
- Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. 2022c. [Towards collaborative neural-symbolic graph semantic parsing via uncertainty](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4160–4173, Dublin, Ireland. Association for Computational Linguistics.
- Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. [Generating benchmarks for factuality evaluation of language models](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66, St. Julian’s, Malta. Association for Computational Linguistics.
- Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. 2008. Accuracy-rejection curves (arcs) for comparing classification methods with a

- reject option. In *Proceedings of the JMLR: Workshop and Conference Proceedings, Machine Learning in Systems Biology*, volume 8, pages 65–81, Paris, France. JMLR.org.
- Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. [Obtaining well calibrated probabilities using bayesian binning](#).
- Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. 2020. [Measuring calibration in deep learning](#). *Preprint*, arXiv:1904.01685.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. [Language models as knowledge bases?](#) *Preprint*, arXiv:1909.01066.
- Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. 2020. [Identifying mislabeled data using the area under the margin ranking](#). *Preprint*, arXiv:2001.10528.
- Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. [Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension](#). *ACM Computing Surveys*, 55(10):1–45.
- Julian Salazar, Davis Liang, Toan Q. Nguyen, and Kartrin Kirchhoff. 2020. [Masked language model scoring](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. [Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback](#). *Preprint*, arXiv:2305.14975.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Jacek Wiland, Max Ploner, and Alan Akbik. 2024. [Bear: A unified framework for evaluating relational knowledge in causal and masked language models](#). *Preprint*, arXiv:2404.04113.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. [Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms](#). *Preprint*, arXiv:2306.13063.
- Paul Youssef, Osman Koraş, Meijie Li, Jörg Schlöterer, and Christin Seifert. 2023. [Give me the facts! a survey on factual knowledge probing in pre-trained language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15588–15605, Singapore. Association for Computational Linguistics.
- Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024. [What matters in memorizing and recalling facts? multifaceted benchmarks for knowledge probing in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13186–13214, Miami, Florida, USA. Association for Computational Linguistics.
- Danna Zheng, Mirella Lapata, and Jeff Z. Pan. 2024. [How reliable are llms as knowledge bases? rethinking factuality and consistency](#). *Preprint*, arXiv:2407.13578.
- Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. [Relying on the unreliable: The impact of language models’ reluctance to express uncertainty](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.
- Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. [Navigating the grey area: How expressions of uncertainty and overconfidence affect language models](#). *Preprint*, arXiv:2302.13439.

A Overview of Models

Table 3 lists all language models included in our experiments.

B Extending the BEAR Templates

To better justify the paraphrase consistency estimates, we extended the number of templates from three to five per relation. To generate the additional templates, we utilized the free version of ChatGPT. We used the following prompt:

As a research assistant, your task is to create an evaluation dataset to assess the relational knowledge of language models. You are provided with a specific relation label, and for each relation label, three semantically equivalent templates are already given, where '[X]' is a placeholder for the subject and '[Y]' for the object (answer). Your objective is to craft two additional semantically similar cloze sentence templates that capture the semantic meaning of the existing templates. Ensure that these new sentence templates are straightforward and devoid of superfluous elements.

For instance, given the following existing templates: "[X] was authored by [Y].", "[X] is a written work by [Y].", "The author of [X] is [Y].", an additional template might be: "[Y] wrote [X]."

Present your response in the JSON file format as you received it. Do not provide only the added templates, but include all five templates for each relation. You will be provided with the relation labels and the corresponding templates one by one in the following prompts.

Model	# Param.	Objective	Accuracy
gemma-2b	2.0b	CLM	50.0
gemma-7b	7.0b	CLM	62.4
gpt2	137M	CLM	12.8
gpt2-medium	355M	CLM	18.4
gpt2-large	812M	CLM	21.5
gpt2-xl	1.6b	CLM	25.3
opt-125m	125M	CLM	15.6
opt-350m	350M	CLM	19.0
opt-1.3b	1.3b	CLM	31.0
opt-6.7b	6.7b	CLM	42.8
gemma-3-270m (IT)	270M	CLM	15.5
gemma-3-1b (IT)	1.0B	CLM	23.6
gemma-3-4b (IT)	4.0B	CLM	40.5
gemma-3-12b (IT)	12.0B	CLM	53.9
Llama-3.2-1B (IT)	1.0B	CLM	37.6
Llama-3.2-3B (IT)	3.0B	CLM	48.7
Llama-3.1-8B (IT)	8.0B	CLM	62.2
Qwen3.5-0.8B (IT)	800B	CLM	21.6
Qwen3.5-2B (IT)	2.0B	CLM	30.7
Qwen3.5-4B (IT)	4.0B	CLM	42.6
Qwen3.5-9B (IT)	9.0B	CLM	51.0
bert-base-cased	109M	MLM	18.1
bert-large-cased	335M	MLM	19.3
roberta-base	125M	MLM	16.0
roberta-large	355M	MLM	21.1
xlm-roberta-base	279M	MLM	11.2
xlm-roberta-large	561M	MLM	13.5
ModernBERT-base	149M	MLM	16.8
ModernBERT-large	335M	MLM	22.0

Table 3: Models considered in this work. (IT) marks an instruction-tuned variant.

C Reduction Strategy Evaluation

We compare different strategies to reduce a set of token log-likelihoods to a single sentence-level log-likelihood. Following are the four different strategies, i.e. variants of C_{Base} (Equation 1):

1. *Sum*: summation of all tokens in the sequence: tokens corresponding to the object, subject, and template.
2. *Mean*: averaging over all tokens in the sequence.
3. *Sum (A)*: summation over (answer) object tokens only.
4. *Mean (A)*: averaging over answer (object) tokens only.

We compare these strategies in terms of accuracy and *ACE*. We additionally calculate their harmonic

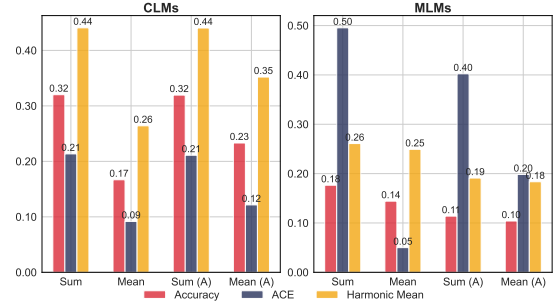


Figure 7: Performance of different token log-likelihood reduction strategies, measured by Accuracy, *ACE* and harmonic mean. Higher harmonic mean is better. Scores are averaged across CLMs and MLMs.

mean weighting accuracy and *ACE* equally:

$$H := \frac{2 \cdot \text{ACC} \cdot (1 - \text{ACE})}{\text{ACC} + (1 - \text{ACE})} \in [0, 1]$$

where we subtracted the *ACE* from 1 so that higher values reflect better overall performance.

In Figure 7, we show the three metrics for each reduction strategy, averaged separately across CLMs and MLMs. We selected the full-sequence summation — *Sum* — as our preferred approach, as it resulted in highest *H* scores for both CLMs and MLMs. *Mean* and *Mean (A)* do result in better calibration scores than the summation variants, however this comes with a significant decrease in accuracy, making it overall less beneficial.

Furthermore, as shown in Figure 8, mean reduction yields strongly right-skewed confidence distributions, which can result in artificially low calibration errors, especially for low-accuracy models. For high-accuracy models, however, the effect is reversed: calibration error increases because confidences remain too low despite correct predictions. Moreover, the choice of reduction strategy itself limits the model’s ability to express high confidences. Under full-sequence mean reduction, the confidences of opt-1.3b do not exceed 0.5.

D Correlation between ACE and Accuracy

We find an overall negative correlation between accuracy and *ACE*, exemplary shown in Figure 9 for C_{Margin} and $C_{\text{Average}}^{\text{Min}}$. Since larger models generally achieve higher accuracy, this result also implies a negative correlation between calibration error and model size. With some exceptions, this trend holds across pre-training objectives and model families. From the perspective of overall trustworthiness,

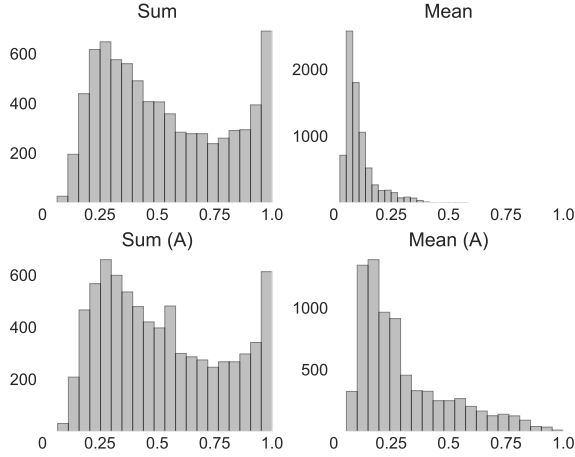


Figure 8: Distributions of C_{Base} for opt-1.3b and the different reduction strategies.

this is concerning: it suggests that gains in calibration primarily arise from higher accuracy rather than from well-aligned confidences.

E Selective Prediction Evaluation

We further evaluate the practical usability of the confidence estimates for selective prediction — where we discard low-confidence data points in order to increase the accuracy on the remaining ones. Figure 10 shows the accuracy-rejection curves—plotting the share of discarded samples against the accuracy of the remaining ones — of C_{Margin} and C_{Average} for selected models. These curves show that the confidence estimates are effective for selective prediction. For example, for gemma-7b, applying a confidence threshold of 0.1 to C_{Average} increases accuracy by about 15 percentage points while rejecting less than 30% of the dataset.

F Wiki-FACTOR Benchmark

While paraphrase consistency and linguistic confidence are specific to template-based probes such as BEAR, intrinsic confidence estimates apply to any multiple-choice setting. For a further benchmark, we implemented the estimates C_{Base} and C_{Margin} on the Wiki-FACTOR dataset (Muhlgay et al., 2024).

FACTOR defines a multiple-choice task in which each question consists of a multi-sentence context, a factual continuation and three non-factual distractors. In the original setup, the model’s answer is chosen as the completion with the highest mean log-probability. As shown on BEAR, mean reduction often results in very low

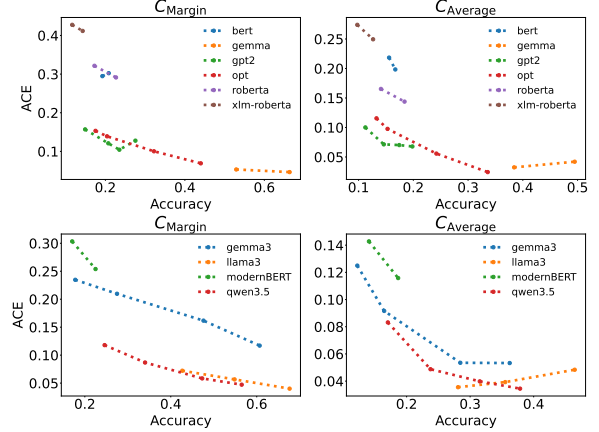


Figure 9: ACE -accuracy scatter plot, demonstrating negative correlation between the two metrics. The first row shows the bert, gemma, gpt2, opt, roberta and xlm-roberta model families. The second row shows the gemma-3, llama3, modern-bert and qwen3.5 model families.

confidence scores, which are not indicative of a well-calibrated model. We observe the same behavior on FACTOR: log-likelihoods do not span the full confidence interval and often differ only marginally, resulting in random guessing confidence splits. Combined with C_{Margin} , the vast majority of confidences would be close to zero. We therefore aggregate token log-likelihoods by summation as we did it for BEAR.

Higher calibration errors on FACTOR. In Table 4 we present the ACE s on FACTOR for C_{Base} and C_{Margin} for the CLMs. The first thing that stands out is that the calibration errors are significantly higher than the ones we obtained on BEAR. On BEAR we found an average ACE of 0.213 for C_{Base} and 0.107 for C_{Margin} for the CLMs. In contrast, on FACTOR the ACE s are 0.475 and 0.446, respectively.

C_{Margin} also improves on FACTOR. We further note that C_{Margin} also consistently improves compared to C_{Base} on FACTOR, but with much smaller magnitude. On BEAR we found an average improvement in ACE among the CLMs of 0.106, while on FACTOR, the average improvement of C_{Margin} is only 0.03. However, we find again a more similar performance if we restrict BEAR to only five answer options per instance, as we did it before. Then average improvement of C_{Margin} is 0.07.

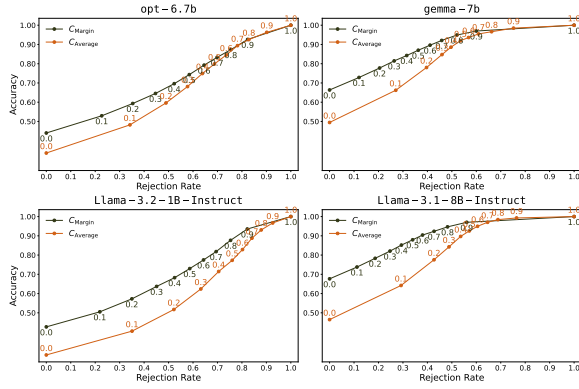


Figure 10: Accuracy Rejection Curves (Nadeem et al., 2008) (fraction of rejected answers versus accuracy among the non-rejected answers) for confidence thresholds of 0.1, . . . , 0.9. For a given threshold, curves closer to the upper left indicate better performance.

Model	Accuracy	ACE ($\downarrow$)		Brier Score ($\downarrow$)	
		C_{Base}	C_{Margin}	C_{Base}	C_{Margin}
gemma-2b	40.21	0.433	0.387	0.457	0.432
gemma-7b	16.50	0.740	0.690	0.753	0.720
gpt2	26.95	0.528	0.482	0.545	0.513
gpt2-medium	31.93	0.481	0.438	0.504	0.477
gpt2-large	35.74	0.448	0.433	0.484	0.468
gpt2-xl	37.81	0.427	0.388	0.459	0.439
opt-125m	29.53	0.505	0.481	0.535	0.515
opt-350m	33.03	0.475	0.452	0.506	0.485
opt-1.3b	41.88	0.395	0.385	0.436	0.427
opt-6.7b	49.33	0.323	0.321	0.377	0.376
gemma-3-270m	26.9	0.139	0.247	0.217	0.264
gemma-3-4b	44.9	0.164	0.402	0.275	0.416
gemma-3-12b	43.0	0.175	0.388	0.277	0.405
Qwen3.5-2B	45.9	0.189	0.407	0.282	0.436
Qwen3.5-4B	53.9	0.243	0.490	0.309	0.504
Qwen3.5-9B	59.0	0.284	0.544	0.325	0.547

Table 4: ACE and Brier Score for C_{Base} and C_{Margin} , as well as accuracy on the Wiki-FACTOR (Muhlgay et al., 2024) dataset. **Bold values** indicate the best-performing estimate per model. **ACE:** ● [0,0.05), ● [0.05,0.10), ● [0.10,0.20), ● ≥ 0.20 ; **Brier Score:** ● [0–0.10), ● [0.10–0.20), ● [0.20–0.25), ● ≥ 0.25 . All gemma-3 and Qwen3.5 models are the instruction-tuned variants.

G Impact of Forced Epistemicity on Accuracy

While forced epistemicity can improve reliability for some models, we find that the injections significantly reduce accuracy, as shown in Figure 11. For instance, gemma-7b shows a drop of 15 percentage points in accuracy for the injection of "I'm 25% confident....", the largest decrease observed overall. The only LM where we observe a slight improvement compared to the unaltered template

bert-base-cased	-0.011	-0.009	-0.024	-0.021	-0.017	-0.016	-0.011
bert-large-cased	-0.030	-0.005	-0.039	-0.040	-0.034	-0.035	-0.027
gemma-2b	-0.030	-0.024	-0.107	-0.099	-0.098	-0.079	-0.092
gemma-7b	-0.054	-0.032	-0.130	-0.150	-0.142	-0.110	-0.129
gpt2	-0.013	-0.012	-0.015	-0.012	-0.010	-0.011	-0.008
gpt2-medium	-0.020	-0.020	-0.027	-0.024	-0.023	-0.021	-0.021
gpt2-large	-0.017	-0.020	-0.031	-0.031	-0.030	-0.031	-0.032
gpt2-xl	-0.025	-0.025	-0.042	-0.043	-0.040	-0.037	-0.034
opt-125m	-0.013	-0.013	-0.016	-0.013	-0.013	-0.011	-0.013
opt-350m	-0.019	-0.016	-0.037	-0.032	-0.032	-0.031	-0.030
opt-1.3b	-0.029	-0.028	-0.058	-0.063	-0.059	-0.054	-0.054
opt-6.7b	-0.039	-0.032	-0.098	-0.109	-0.095	-0.085	-0.087
roberta-base	-0.014	0.002	-0.018	-0.023	-0.015	-0.011	-0.008
roberta-large	-0.031	-0.001	-0.019	-0.042	-0.027	-0.024	-0.025
xlm-roberta-base	-0.007	-0.014	-0.024	-0.024	-0.024	-0.024	-0.019
xlm-roberta-large	-0.023	-0.009	-0.006	-0.015	-0.008	-0.011	-0.004
Qwen3.5-0.8B	-0.020	-0.013	-0.038	-0.033	-0.027	-0.030	-0.017
Qwen3.5-2B	-0.034	-0.024	-0.048	-0.048	-0.035	-0.032	-0.021
Qwen3.5-4B	-0.033	-0.001	-0.064	-0.052	-0.036	-0.031	-0.015
Qwen3.5-9B	-0.038	-0.004	-0.085	-0.094	-0.071	-0.066	-0.041
Llama-3.1-8B-Instruct	-0.043	-0.012	-0.141	-0.132	-0.090	-0.075	-0.079
Llama-3.2-1B-Instruct	-0.027	-0.025	-0.062	-0.044	-0.038	-0.036	-0.034
Llama-3.2-3B-Instruct	-0.040	-0.027	-0.087	-0.072	-0.068	-0.057	-0.075
ModernBERT-base	-0.008	-0.004	-0.006	-0.015	-0.006	-0.006	0.001
ModernBERT-large	-0.011	0.004	-0.016	-0.007	-0.002	-0.001	0.003
gemma-2b-it	-0.036	-0.017	-0.054	-0.050	-0.048	-0.040	-0.046
gemma-7b-it	-0.053	-0.009	-0.064	-0.084	-0.055	-0.040	-0.034
gemma-3-270m	-0.008	-0.002	-0.020	-0.015	-0.017	-0.015	-0.017
gemma-3-1b-pt	-0.033	-0.026	-0.053	-0.037	-0.034	-0.029	-0.038
gemma-3-4b-pt	-0.034	-0.014	-0.090	-0.098	-0.095	-0.075	-0.069
gemma-3-12b-pt	-0.040	-0.018	-0.103	-0.102	-0.092	-0.065	-0.075
gemma-3-270m-it	-0.011	-0.013	-0.040	-0.041	-0.038	-0.037	-0.033
gemma-3-1b-it	-0.033	-0.016	-0.066	-0.035	-0.033	-0.029	-0.021
gemma-3-4b-it	-0.044	-0.011	-0.103	-0.085	-0.069	-0.058	-0.034
gemma-3-12b-it	-0.035	-0.006	-0.131	-0.103	-0.075	-0.042	-0.035
Average	-0.027	-0.014	-0.056	-0.054	-0.046	-0.039	-0.036
	possibly	certainly	0%	25%	50%	75%	100%

Figure 11: Difference in Accuracy between the non-injected and injected templates. Positive values show an improvement over the non-injected version.

is roberta-base, where accuracy increased by 0.2 percentage points for the injection of "certainly". We also observe that larger models are more sensitive to these injections, particularly to numerical confidence expressions.