Cultural Calibration of Verbalised Uncertainty in Multilingual LLMs

Anonymous ACL submission

Abstract

Epistemic markers, such as *maybe* or *definitely*, signal a speaker's certainty or uncertainty about communicated information, and their use varies across languages and cultures. As Large Language Models (LLMs) are increasingly deployed in multilingual and multicultural settings, it is important to understand whether they adapt their expression of certainty to cultural conventions. We investigate whether LLMs generate culturally appropriate epistemic expressions by prompting seven multilingual models on a multiple-choice Q&A task across nine countries spanning three languages (German, English, Spanish), using documented contrastive-pragmatic patterns and Hofstede's UAI as complementary proxies for locale-level conventions. We evaluate model behaviour in terms of consistency, accuracy, and abstention calibration, and assess the certainty level of generated expressions using human-annotated classifiers.

1 Introduction

Epistemic markers are used in language to signal whether the speaker is certain or uncertain of the information being communicated. Their production and interpretation vary systematically across languages and cultures (Shuper et al., 2004; Phillips and Wright, 1977).

As LLMs are deployed globally, their expressions of certainty are increasingly consumed by users with divergent epistemic conventions. A culturally appropriate model would take these conventions into account, adjusting how it conveys the same internal confidence to different audiences. Prior work has shown that LLMs are systematically overconfident (Mielke et al., 2022; Xiong et al., 2024; Zhou et al., 2024), that this overconfidence persists across languages (Rathi et al., 2025; Krause et al., 2023), and that verbalised confidence rarely reflects intrinsic uncertainty (Yona et al., 2024; Liu et al., 2025). These studies, however, treat the mapping from internal confidence to surface expression as a property of language alone, not of culture.

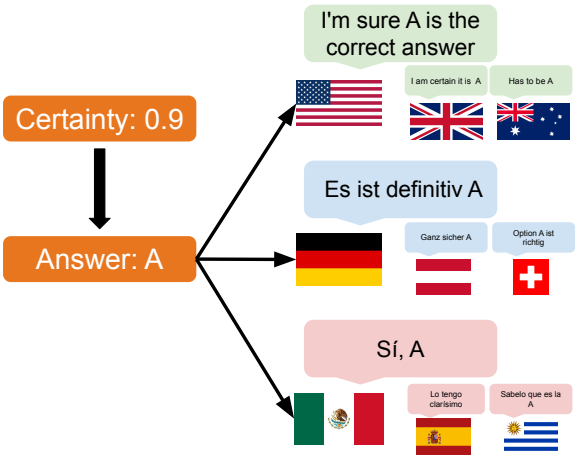

Figure 1: Differences in how knowledge confidence is communicated as expressed certainty, based on language and cultural background (as approximated by country).

In this work we address the research question: *To what extent do multilingual LLMs adapt their verbalised uncertainty to the linguistic and cultural conventions of the target speech community, and does this adaptation faithfully reflect their internal confidence?* We prompt nine multilingual LLMs on a multiple-choice Q&A task across nine countries spanning three languages (German, English, Spanish), using Hofstede's Uncertainty Avoidance Index (UAI) (Hofstede, 2001) as a cultural proxy. Our hypotheses are shown in Table 1.

Contributions (i) We provide the first systematic evaluation of *cultural*, as distinct from purely linguistic, calibration of verbalised uncertainty in multilingual LLMs. (ii) We release a locale-grounded inventory of natural certainty and uncertainty expressions elicited from native speakers across nine country-language combinations. (iii) We release a human-annotated corpus of model-generated epis-

Hypothesis	Prediction
H1: Language-family calibration. Models adapt epistemic expression to broad linguistic conventions at the language-family level.	Epistemic expression varies systematically across English, German, and Spanish, tracking documented cross-linguistic patterns: more lexical hedging in English, more high-probability markers in German, and more variable expression in Spanish with grammatical mood as an additional resource.
H2: Locale-level linguistic calibration. Models adapt to fine-grained linguistic conventions that vary across varieties of the same language.	Within each language, locale-level differences track documented contrastive-pragmatic patterns: GBR hedges more than USA and AUS; URY uses more downgrading than ESP and shows more tentativeness overall; AUT uses more upgrading and hearer-orientation than DEU; CHE prefers consensus-oriented indirectness.
H3: Locale-level cultural calibration. Models adapt to cultural epistemic norms beyond what language family and linguistic variety predict.	Within each language, locale-level abstention and certainty expression track UAI. Tested in two directions: **H3a** (uncertainty-aversion) — high UAI → less abstention, less hedging; **H3b** (epistemic-honesty) — high UAI → more abstention, more hedging.
H4: No adaptation. Models produce uniform epistemic expression across all conditions.	No significant variation in abstention or certainty expression across locales or languages.

Table 1: Hypotheses and predictions for epistemic expression behaviour across locales, organised by level of granularity: language-family (H1), locale-level linguistic (H2), locale-level cultural (H3), and no adaptation (H4).

temic expressions rated for perceived certainty by native annotators, providing a benchmark for future work on faithful calibration in multilingual settings. (iv) Empirically, we find that models adapt their epistemic expressions across locales along language lines, but not cultural ones, that no significant correlation emerges between UAI and any behavioural or expressive metric, and that models cannot reliably distinguish uncertainty from error. All code and data will be release upon acceptance.

2 Related work

In the following subsections we revise the literature related to three topics: calibration and verbalised uncertainty, multilingual uncertainty expression, and culture in NLP.

2.1 Calibration and Verbalised Uncertainty

Confidence calibration in LLMs has typically been pursued from a factual perspective, aligning a model's confidence with its empirical accuracy (Guo et al., 2017; Kamath et al., 2020; Jiang et al., 2021). White-box methods exploit token probabilities, entropy, or hidden-state probes (Duan et al., 2024; Azaria and Mitchell, 2023; Burns et al., 2024); black-box methods infer confidence from outputs via semantic consistency (Kuhn et al., 2023), self-consistency (Wang et al., 2023), self-evaluation (Kadavath et al., 2022), or auxiliary predictors trained on inputs and outputs alone (Ulmer et al., 2024). A related line targets selective prediction and abstention (Kamath et al., 2020; Yin et al., 2023; Zhang et al., 2024), which we utilise through an explicit "I do not know" option in our task. For natural-language interfaces, calibration must extend to surface form. Lin et al. (2022) and Tian et al. (2023) show that LLMs can verbalise calibrated confidence through fine-tuning or prompting, and Mielke et al. (2022) introduce linguistic calibration via controllable generation. Xiong et al. (2024) benchmark elicitation strategies and find that verbalised confidences concentrate in narrow, overconfident ranges, while Zhou et al. (2023, 2024) document a systematic reluctance to express uncertainty, motivating our few-shot elicitation design. A distinct strand targets faithful calibration: the alignment between intrinsic and verbalised uncertainty. Yona et al. (2024) formalise this as an example-level metric, and Liu et al. (2025) extend it into a systematic benchmark, showing that standard prompting yields only marginal gains and that factuality-based methods can actively harm faithfulness. Both treat the mapping from intrinsic confidence to surface form as language-independent. The behavioural stakes are non-trivial: Kim et al. (2024) find that first-person hedges reduce overreliance, Xu et al. (2025) show that moderate verbalised uncertainty maximises user trust, and Ulmer et al. (2025) demonstrate that verbalised uncertainty shifts with conversational context, arguing that calibration should be evaluated against human communicative norms.

2.2 Multilingual Uncertainty Expression

Cross-lingual evaluations of verbalised uncertainty remain scarce (Ulmer et al., 2025). Krause et al.

(2023) provide an early multilingual benchmark on a multiple-choice QA task across five languages, finding systematic miscalibration and cross-lingual differences in both production and accuracy. Rathi et al. (2025) extend this to further languages and find that LLMs are uniformly more overconfident in non-English settings while tracking documented linguistic norms (more weakeners in Japanese, more strengtheners in German and Mandarin). Crucially, they show that human reliance on epistemic markers also varies cross-linguistically, in ways that can cancel out the apparent benefit of more frequent hedging. We adopt their few-shot prompting approach and Likert-scale annotation protocol (see Section 4). Where Rathi et al. attribute differences to language, we ask whether cultural variation within a shared language plays an analogous role. Cross-cultural work has shown that probabilistic thinking itself varies with cultural background (Phillips and Wright, 1977; Shuper et al., 2004), and that identical verbal probability expressions map to systematically different numerical thresholds across U.S./German (Doupnik and Richter, 2003, 2004) and German/Spanish (Bocklisch et al., 2013) speakers, with the language-culture effect more pervasive than translation effects alone. This phenomenon has not, to our knowledge, been examined in multilingual LLMs.

2.3 Culture in NLP

Cross-cultural NLP raises both methodological and conceptual challenges (Hershcovich et al., 2022). Culture is most commonly operationalised through country as a demographic proxy (Adilazuarda et al., 2024), a simplification we adopt with full awareness of its limits. Studies probing whether LLMs reproduce cross-national survey patterns consistently find limited alignment with human responses: across the VSM and WVS (Arora et al., 2023; Cao et al., 2023; Kharchenko et al., 2025; Bulté and Rigouts Terryn, 2025), Hofstede-based benchmarks (Karinshak et al., 2024), and persona-based simulations (Kwok et al., 2024; AlKhamissi et al., 2024). Khan et al. (2025) caution that observed cultural variation may reflect prompt sensitivity rather than stable representation. We build on this line of work but redirect it from values to communicative norms, asking whether multilingual LLMs adapt their epistemic expressions to the cultural conventions of the target speech community.

3 Cross-Linguistic and Cross-Cultural Patterns of Epistemic Expression

Two frameworks motivate our locale-level design. Contrastive pragmatics documents how English, Spanish, and German encode and deploy epistemic markers (Kranich, 2011; Martín, 2008; Vold, 2006; House, 2006). Hofstede's Uncertainty Avoidance Index (UAI) captures broader cultural orientations towards ambiguity (Hofstede, 2001). The latter has methodological limitations (Baskerville, 2003), but remains useful for predicting broad patterns of communicative preference (Reinecke and Bernstein, 2011). Notably, linguistic hedging frequency and cultural uncertainty avoidance are not necessarily aligned: some high-UAI cultures hedge less rather than more, preferring precision or grammatical mood alternation over lexical mitigation (Kranich, 2011; House, 2006).

Three patterns inform our analysis. First, British English hedges the most despite low UAI, because hedging in English serves face-management (Brown and Levinson, 1987; Hyland, 1998). American English uses moderate lexical hedges, and Australian English sits between British indirectness and American directness, though direct comparisons of hedging frequency across the three varieties remain scarce. Second, Spanish encodes epistemic commitment grammatically through the subjunctive/indicative alternation (Faulkner, 2022; Mejias-Bikandi, 1998), supplementing lexical hedging. Within Spanish, country variants diverge: Peninsular Spanish is direct with low hedging, Mexican Spanish layers extensive politeness mitigation (Félix-Brasdefer, 2005, 2006), and Uruguayan Spanish is the most tentative, with elaborate downgrading devices (Márquez Reiter, 2002). Third, German hedges infrequently and prefers high-probability markers ("wahrscheinlich," "vermutlich") over mere-possibility markers (Kranich, 2011). German speakers also assign lower numerical thresholds to identical verbal probability expressions than English speakers do (Doupnik and Richter, 2003, 2004). Austrian German diverges from German German pragmatically: Muhr (1994) shows Austrian apologies adopt a hearer perspective more than German German apologies, are more explicit, and use more upgrading. This is consistent with a moderately higher UAI (70 vs 65). Swiss German is consensus-oriented and relationally indirect (Werlen, 1998).

The full descriptive overview is in Appendix A.1.

A model that calibrates uncertainty expression at the language level alone will produce outputs that are pragmatically marked or inappropriate for users in specific cultural contexts.

4 Experimental set up

Languages and countries We selected 3 languages: German (DE), English (EN) and Spanish (ES). For each language we selected 3 countries that maximise the variation in the Hofstede scale: Austria (AUT), Switzerland (CHE), Germany (DEU), Australia (AUS), the United Kingdom (GBR), the United States (USA), Spain (ESP), Mexico (MEX), and Uruguay (URY). Figure 2 shows the UAI scores for the 9 countries included in this study, revealing that EN-speaking countries have the lowest UAI overall, while ES-speaking countries have the highest UAI. Additionally, we generate a default condition for each language to investigate language-level behaviour independently of cultural-level behaviour. The set of 9 countries and 3 default language conditions are referred to as *locales*.

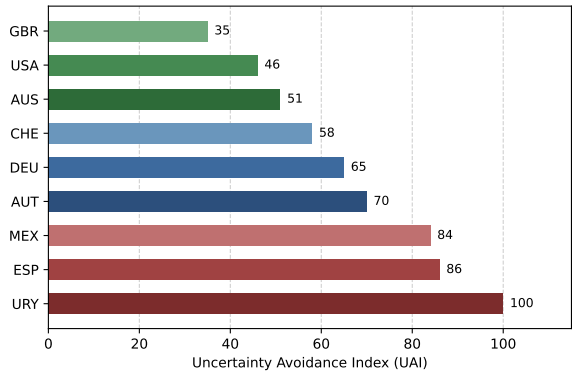

Figure 2: Hofstede values for Uncertainty Avoidance (UAI). Colors indicate groups by language family (blues = DE, greens = EN, reds = ES).

Models We selected nine instruct-tuned LLMs of comparable size (7–9B parameters) that support the three chosen languages: Apertus 8B[1], Aya Expanse 8B[2], Bloomz 7B[3], EuroLLM 9B[4], Gemma 2 9B[5], Granite 3.3 8B[6], Llama 3.1 8B[7], Ministral 3 8B[8], and Qwen2.5 7B[9]. After initial investigations into their performance, Aya, Bloomz, and EuroLLM were not considered for reported results due to their inability to produce epistemic markers as instructed.

Dataset We used the MMLU-ProX-lite, a multiple-choice Q&A dataset (Xuan et al., 2025). Each question has 10 options to choose from (A–J). To encourage explicit abstention, we added an 11th option (K): "I do not know". All questions are identical across languages and locales, being translations of the same items, ensuring that any observed differences in model behaviour across locales reflect linguistic and cultural factors rather than differences in task difficulty. We further restrict the question set to STEM domains (Biology, Chemistry, Physics, Mathematics, Computer Science, and Engineering) to minimise the influence of culturally loaded question content. The validation set consists of 30 items, used for few-shot prompting and model selection, while the test set consists of 293 items, used for model behaviour analysis.

Prompt We follow a similar approach to Rathi et al., with few-shot prompting and explicitly eliciting epistemic markers[10]. For the few-shot examples, we randomly selected 10 items from the validation set and added a *comment* to the answers to include epistemic expression. We assigned five cases to express certainty about the comment, four cases to express uncertainty, and one case for abstention. To ground the few-shot exemplars in authentic local usage, we elicited certainty and uncertainty expressions from native speakers in each of the nine country-language combinations. Three native speakers per locale provided five expressions of certainty and five of uncertainty that they would naturally use when speaking to a fellow national. The most frequent expressions per locale were selected as the epistemic markers for the few-shot examples. The prompt was constructed in the target language and included the following text as instructions: "The comment should be written in natural [COUNTRY] [LANGUAGE]." An example prompt is shown in the Appendix.

Implementation We used the "chat template" approach from HuggingFace and investigated the best

[1] swiss-ai/Apertus-8B-Instruct-2509
[2] CohereLabs/aya-expanse-8b
[3] bigscience/bloomz-7b1
[4] utter-project/EuroLLM-9B-Instruct
[5] google/gemma-2-9b-it
[6] ibm-granite/granite-3.3-8b-instruct
[7] meta-llama/Llama-3.1-8B-Instruct
[8] mistralai/Ministral-3-8B-Instruct-2512
[9] Qwen/Qwen2.5-7B-Instruct
[10] As investigated by Zhou et al., LLMs must be explicitly encouraged to incorporate a hedge or a strengthener to signify their certainty level.

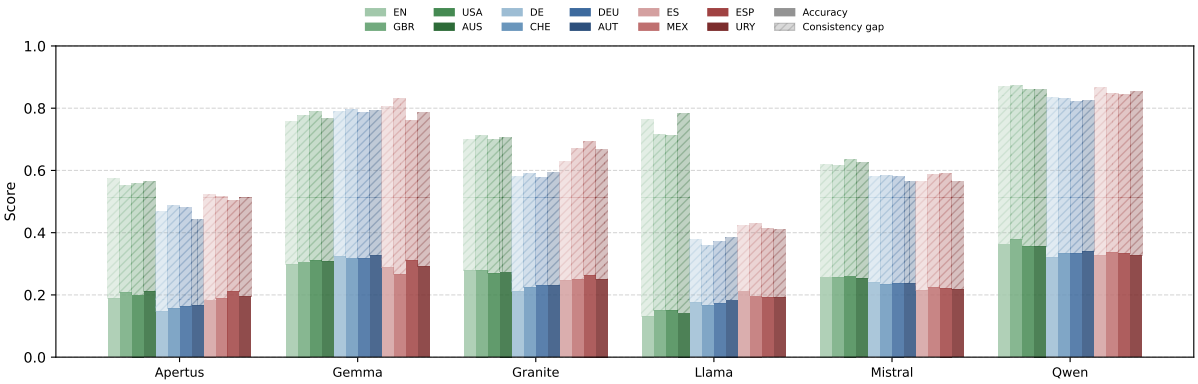

Figure 3: Accuracy and consistency per model and locale.

setup for adjusting the system prompt. These ablation studies can be found in Appendix A.2. We run each locale 15 times with temperature and top-p = 0.9. We had access to 16 NVIDIA GeForce RTX 4090.

5 Evaluation

We analysed the model outcomes along three dimensions: behavioural, mechanistic, and expressed confidence.

Behavioural: accuracy and explicit abstention We computed accuracy (the proportion of runs in which the model chose the correct answer) as well as the abstention rate (the proportion of runs in which the model chose option K).

Mechanistic: consistency and log-likelihood We computed consistency (the proportion of runs in which the model chose the same answer). We also extracted the log-likelihood (ll) of the token corresponding to the selected answer.

Expressed confidence: epistemic markers We extracted epistemic expressions from the generated answer comments, filtering out invalid or overly long (>15 words) comments and normalising references to specific answer options (e.g., "I'm pretty sure it's G." → "I'm pretty sure it's [X]."). Unique expressions were then rated by native-speaker annotators recruited via Prolific (three per country) on a 7-point Likert scale from 1 (very uncertain) to 7 (very certain), following Rathi et al. Annotators rated only how certain the speaker sounded, independent of answer correctness. We applied attention checks and item duplication for quality control, and refined the protocol after a pilot round on validation-set generations. Full details of recruitment, compensation, quality control, pilot procedure, and inter-/intra-annotator agreement are reported in Appendix A.3. Annotation coverage exceeded 90% of generations across all models and locales.

6 Results

6.1 Trends across models, languages, countries

Figure 3 shows model consistency and accuracy across locales. Qwen and Gemma lead on both metrics (accuracy: 0.342 and 0.306; consistency: 0.849 and 0.787). Yet, they also show the largest gap between the metrics, indicating they repeat the same wrong answer rather than being confident in a correct one. Apertus and Llama are better calibrated in this respect, though Llama stands out for its distinct behaviour in English compared to other languages.

Across languages, English has the highest accuracy and consistency (0.258 and 0.712), followed by Spanish (0.248 and 0.637) and German (0.241 and 0.604). The accuracy–consistency gap, however, inverts this ranking. German locales are the best calibrated (gap: 0.353–0.370), Spanish locales fall in the middle (0.378–0.403), and English locales show the largest gap (0.446–0.461), meaning models are more likely to confidently repeat incorrect answers in English. Within each language, default locales tend to have a slightly larger gap than their country-specific counterparts.

6.2 Thresholds for low/moderate/high expressions of certainty per language

Cross-cultural work on probability expressions has shown that identical verbal terms map to different numerical thresholds depending on the listener's cultural background (Doupnik and Richter, 2003,

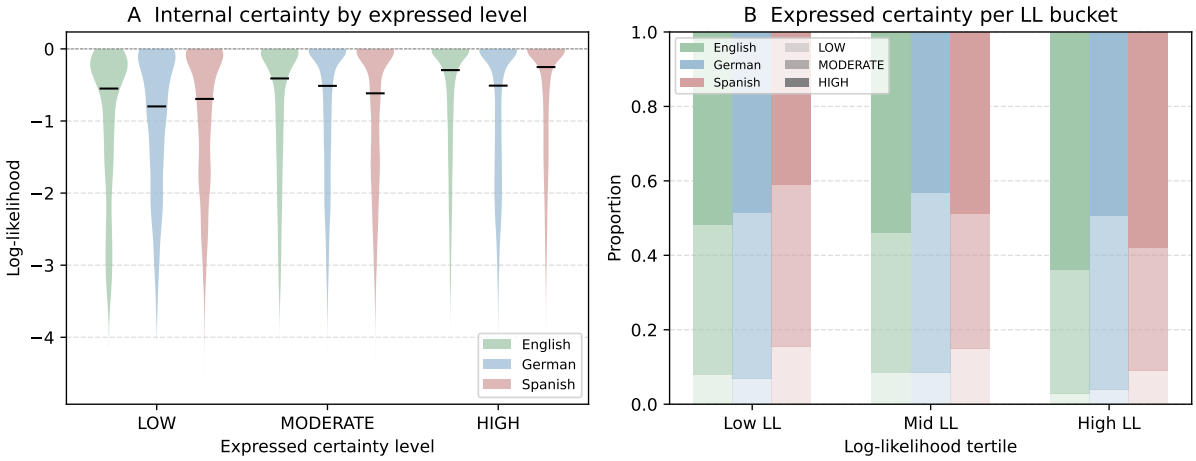

Figure 4: Left: log-likelihood distributions per expressed certainty level and language. Right: distribution of expressed certainty levels per log-likelihood tertile, by language.

2004; Bocklisch et al., 2013). We examine whether LLMs reproduce this pattern by relating the internal certainty of the model (log-likelihood and consistency across runs) to the level of certainty expressed in the generated comment.

Table 2 summarises the results. Spearman correlations between log-likelihoods and expressed certainty levels are statistically significant but small, as are those between consistency and expressed certainty. Mean log-likelihoods increase monotonically across certainty levels, but the spread is large. Consistency also increases monotonically but has a smaller spread. When broken down per language, mean log-likelihoods are also monotonic for English and Spanish, but not so for German. Mean consistency follows the monotonic trend for all languages, but is steeper for English and Spanish than for German.

	Level	$\bar{\ell}$	SD ℓ	$\bar{c}$	SD c	$\rho(\ell)$	$\rho(c)$
All	LOW	-1.00	0.97	0.65	0.25		
	MOD	-0.88	0.95	0.69	0.27	$.098^{***}$	$.083^{***}$
	HIGH	-0.76	0.90	0.72	0.27		
EN	LOW	-1.01	1.02	0.68	0.22		
	MOD	-0.84	0.94	0.71	0.25	$.116^{***}$	$.087^{***}$
	HIGH	-0.70	0.87	0.74	0.25		
DE	LOW	-1.03	0.94	0.63	0.26		
	MOD	-0.88	0.96	0.69	0.28	$.023^{***}$	.002
	HIGH	-0.90	0.96	0.67	0.29		
ES	LOW	-0.98	0.94	0.64	0.26		
	MOD	-0.93	0.96	0.67	0.28	$.157^{***}$	$.158^{***}$
	HIGH	-0.68	0.85	0.75	0.26		

Table 2: Mean and SD of log-likelihood ($\bar{\ell}$, SD ℓ) and consistency ($\bar{c}$, SD c), and Spearman correlations between internal certainty and expressed certainty level ($\rho(\ell)$, $\rho(c)$), overall and by language. $^{***}p<.001$.

As seen in Figure 4, bucketing log-likelihood into thirds (low: ≤ -1.08; mid: -1.08 - -0.11; high: > -0.11) reveals that HIGH-certainty expressions always dominate. The distribution of expressed certainty barely shifts across LL buckets, confirming that models default to high-certainty language regardless of internal confidence.

6.3 How often do models abstain?

English varieties show the lowest abstention rates (4.3–4.4%), followed by German (4.9–6.5%) and Spanish (5.3–10.0%). Spanish spreads from ESP, which is close to the German range, to MEX , which reaches the highest abstention rate at 10%. Default locales tend to abstain more than their country-specific counterparts, suggesting that cultural framing reduces abstention.

6.4 When do models abstain?

Across all locales, correct answers are the result of higher LL (mean = -0.499) than incorrect ones (mean = -1.123). Yet, abstentions lay close to wrong answers (means = -1.090), suggesting that, overall, models cannot reliably distinguish uncertainty from error.

This pattern is not uniform across locales, as shown in Figure 5. For English locales, abstentions occur at even lower LL than wrong answers (δ = -0.09 to -0.33), indicating that abstention follows uncertainty. German locales show the opposite trend where abstentions occur at higher LL than wrong answers (δ = +0.05 to +0.16), meaning models abstain even when relatively confident. Spanish is mixed. ESP abstains conservatively (δ = -0.17), while MEX and the default show the largest positive gaps (+0.32 and +0.35). The LL gap between

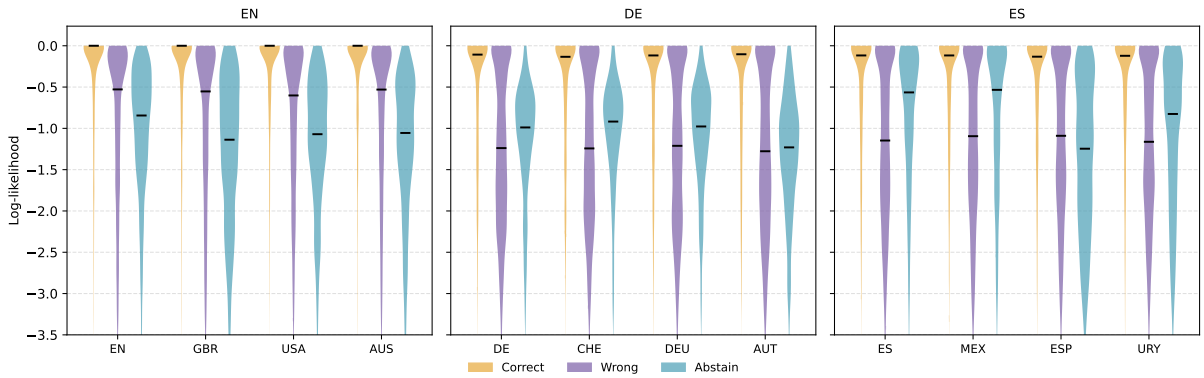

Figure 5: Log-likelihood distributions by chosen answer and locale, grouped by language.

correct and wrong answers is widest for German locales (0.70), moderate for Spanish (0.65), and smallest for English (0.51), indicating that LL is a stronger signal in German-language contexts.

6.5 How do models express themselves when abstaining?

When models abstain, they use explicit uncertainty expressions such as "I don't know", "Ich weiß es nicht", and "No lo sé", rated LOW or MODERATE certainty (mean = 1.46). When models answer correctly, the top expressions are assertive: "Absolutely certain", "Es ist definitiv B", "Estoy seguro" (mean = 2.52). The same high-certainty expressions are strongly present in wrong answers, with "I'm sure that E is the correct answer", "Es ist definitiv A", and "Estoy seguro" among the most frequent (mean = 2.45). This shows that high-certainty language is equally common whether the answer is right or wrong.

The patterns vary across languages. Spanish locales express the strongest uncertainty when abstaining (mean = 1.21), followed by English (mean = 1.47) and German (mean = 1.76). English locales express the most certainty when answering, whether correct (mean = 2.60) or wrong (mean = 2.50).

6.6 How do models express themselves overall?

Firgure 6 shows the trends in expressed certainty per country. In English locales, AUS expresses HIGH certainty in 63.1% of cases, while GBR and USA use it at lower rates (47.7% and 46.5% respectively). Between the latter two, the main difference is that USA rarely uses LOW certainty markers (1.7%). In German locales, CHE and AUT are dominated by MODERATE certainty (48.2% on both cases), while DEU behaves closer to AUS with mostly expressing HIGH certainty (60%). Overall, German locales rarely express low certainty AUT 5.3%, CHE 4.3%, DEU 4.6%. In Spanish locales, MEX and ESP show the most LOW certainty overall (20.6% and 11.0%). Between them, they are differentiated by ESP having the most HIGH certainty expressions across all locales (64.9%). URY stands apart with a different behaviour, using MODERATE the most across all locales (58.7%).

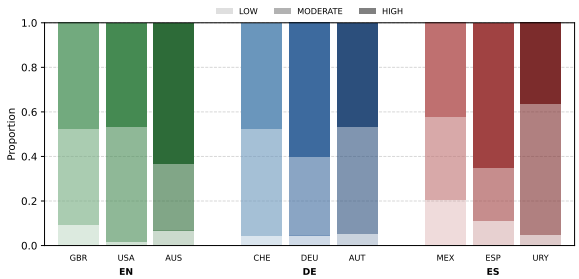

Figure 6: Distribution of expressed certainty per locale, grouped by language, averaged across models.

6.7 Correlations with UAI

No significant correlation between UAI and any metric emerges when aggregated across all locales (all $|\rho| \leq 0.17$, $p > 0.18$). Within Spanish locales, UAI positively correlates with certainty when abstaining ($\rho = 0.62$, $p < 0.01$, $n = 19$), but this pattern does not replicate within German or English groups. URY, the highest-UAI locale, behaves similarly to low-UAI German varieties.

To assess whether the null result is specific to UAI or reflects a broader absence of cultural signal, we computed Spearman correlations for all five Hofstede dimensions against all metrics, aggregated and grouped by language. The dimensions are: Uncertainty Avoidance (UAI), Power Distance

(PDI), Indulgence vs Restraint (IVR), Masculinity vs Femininity (MAS), and Individualism vs. Collectivism (IDV). No dimension showed consistent significant correlations. Full results are in Appendix A.4.

7 Discussion

Given our results, we evaluate the hypotheses from Table 1.

Language-family calibration H1 receives the strongest support, with the three languages showing structurally distinct calibration patterns. English combines highest accuracy and consistency with the largest accuracy–consistency gap (0.446–0.461), indicating confident repetition of wrong answers. German shows the most distinctive pattern: the widest LL gap between correct and wrong answers, the lowest LOW-certainty rate across the study (4.3–5.3% per locale), and abstention at higher LL than wrong answers (+0.05 to +0.16), the opposite of the English pattern. Spanish sits between on the core accuracy and consistency metrics, while showing the highest abstention rates of the three languages. This pattern suggests that the models reproduce language conventions from training data rather than engaging in locale-sensitive epistemic reasoning.

Locale-level linguistic calibration H2 is supported where contrastive-pragmatic predictions track gross distributional differences, most clearly within Spanish (URY tentativeness, MEX mitigation, ESP directness), and fails for finer-grained contrasts within English (the predicted GBR > USA, AUS hedging ordering does not appear) and German (the predicted AUT > DEU upgrading does not appear). The URY result in particular should be treated cautiously, as Uruguayan Spanish is less web-represented than ESP or MEX, and its MODERATE-heavy pattern may reflect an underspecified default rather than genuine calibration.

Locale-level cultural calibration H3 tests two competing UAI-driven patterns. H3a predicts that high-UAI locales would abstain and hedge less, preferring decisive responses. We find the opposite: Spanish locales, which have the highest UAI scores, show the highest abstention rates and the most LOW certainty when abstaining. H3b predicts that high-UAI locales abstain and hedge more, preferring epistemic honesty. This is consistent with the language-level abstention pattern (ES > DE > EN), but does not hold within language groups, where UAI ordering does not predict abstention or certainty expression. Therefore, we find no support for either UAI-driven patterns at the locale level.

No adaptation H4 is rejected at the language level, but partially consistent within each language, where locale-level differences are weak or absent.

Faithfulness of expressed uncertainty Independent of locale, expressed certainty barely tracks internal confidence: log-likelihood correlates only weakly with expressed certainty level, and HIGH-certainty language dominates across all log-likelihood buckets. Models cannot reliably distinguish uncertainty from error, with abstentions clustering at log-likelihoods close to wrong rather than correct answers. This means that even where surface adaptation occurs, it does not reflect faithful calibration of internal confidence.

8 Conclusion

This paper investigated whether multilingual LLMs calibrate their epistemic expressions to cultural conventions as approximated by Hofstede's Uncertainty Avoidance Index. Across six models and nine locales spanning German, English, and Spanish, we find that models do adapt their certainty expressions across locales, but in ways that reflect language level linguistic conventions rather than cultural calibration per locale. We find only partial adaptation to documented locale-linguistic conventions and no significant correlation between UAI and any behavioural or expressive metric. Mechanistically, models cannot reliably distinguish uncertainty from error: abstentions and wrong answers cluster at near-identical log-likelihoods, and high-certainty language is equally common whether the model is right or wrong. Abstention calibration is strongest in English locales and largely absent in German and Spanish. These findings suggest that current multilingual LLMs lack cultural epistemic calibration that can adapt to different locales using the same language, and that observed locale differences are driven by the linguistic properties of the target language rather than culturally grounded uncertainty conventions. We release both the locale-grounded expression inventory and the human-annotated corpus of model-generated epistemic expressions to support future work on multilingual faithful calibration.

Limitations

Our study has several limitations. First, we operationalise culture through country, following common practice in cross-cultural NLP (Adilazuarda et al., 2024), but country is a coarse proxy that obscures within-country variation along regional, generational, and socio-economic lines. Hofstede's UAI inherits the same limitation and has been critiqued on methodological grounds (Baskerville, 2003); our null result on UAI does not rule out cultural calibration under finer-grained operationalisations. Second, we restrict the study to three languages and nine locales, which constrains the generalisability of our findings. The selected language families share a Western European heritage, and patterns may differ for typologically more distant languages or for locales with weaker representation in training corpora. URY's anomalous behaviour suggests that locale representation in pretraining data is a confound we cannot directly measure. Third, we evaluate six instruct-tuned models in the 7–9B parameter range. Larger models, models trained with different post-training pipelines, or models targeting specific locales may exhibit different calibration behaviour. Three of the nine initially selected models were excluded for failing to follow the epistemic-marker instruction, which itself indicates that our findings depend on a model's ability to verbalise certainty on demand. Fourth, the task is restricted to STEM multiple-choice questions to minimise culturally loaded content. This restriction limits the applicability of our findings to domains where cultural framing of the question itself, rather than only of the answer, may shape epistemic expression. Fifth, our annotation protocol relies on three native speakers per locale recruited via Prolific. Inter-annotator agreement is moderate, reflecting the inherent subjectivity of perceived certainty ratings. We mitigate this through attention checks, item duplication, and discarding annotators who fail quality controls, but residual noise in the certainty labels may attenuate observed correlations.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Francisco Alonso Alonso Almeida and María Luisa Carrió Pastor. 2017. Variation and function of modals in linguistics and engineering research papers in english. In *Evidentiality and modality in european languages: discourse-pragmatic perspectives*, pages 277–310. Peter Lang.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Rachel F Baskerville. 2003. Hofstede never studied culture. *Accounting, organizations and society*, 28(1):1–14.

Shoshana Blum-Kulka and Elite Olshtain. 1984. Requests and apologies: A cross-cultural study of speech act realization patterns (ccsarp). *Applied linguistics*, 5(3):196–213.

Franziska Bocklisch, Anne Georg, Steffen Bocklisch, and Josef Krems. 2013. Do you mean what you say? the effect of uncertainty avoidance on the interpretation of probability expressions-a comparative study between spanish and german. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Bram Bulté and Ayla Rigouts Terryn. 2025. Llms and cultural values: The impact of prompt language and explicit cultural framing. *Computational Linguistics*, pages 1–85.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2024. Discovering latent knowledge in language models without supervision. *Preprint*, arXiv:2212.03827.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of*

the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Timothy S. Doupnik and Martin Richter. 2003. Interpretation of uncertainty expressions: a cross-national study. *Accounting, Organizations and Society*, 28(1):15–35.

Timothy S. Doupnik and Martin Richter. 2004. The impact of culture on the interpretation of "in context" verbal probability expressions. *Journal of International Accounting Research*, 3(1):1–20.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Tris Faulkner. 2022. The two spanish subjunctives: the required and default subjunctives. *Borealis–An International Journal of Hispanic Linguistics*, 11(1):70–100.

J César Félix-Brasdefer. 2005. Indirectness and politeness in mexican requests. In *Selected proceedings of the 7th Hispanic linguistics symposium*, volume 6678. Cascadilla Proceedings Project Somerville, MA.

J César Félix-Brasdefer. 2006. Linguistic politeness in mexico: Refusal strategies among male speakers of mexican spanish. *Journal of pragmatics*, 38(12):2158–2187.

Cliff Goddard and Rahel Cramer. 2016. "laid back" and "irreverent": An ethnopragmatic analysis of two cultural themes in australian english communication. *Handbook of Communication in Cross-Cultural Perspective, London: Routledge*, pages 89–103.

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

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Mark Hoff. 2020. Pragmatic alternations in rioplatense spanish. *Hispanic Linguistics: Current issues and new directions*, 26:161.

Geert Hofstede. 1991. Cultures and organizations: Software of the mind. *McGraw Hills*.

Geert Hofstede. 2001. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications.

Juliane House. 2006. Communicative styles in english and german. *European journal of English studies*, 10(3):249–267.

Ken Hyland. 1998. Hedging in scientific research articles. *Pragmatics & beyond. New series*, 54.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. Llm-globe: A benchmark evaluating the cultural values embedded in llm output. *Preprint*, arXiv:2411.06032.

Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, not representation: The unreliability of evaluating cultural alignment in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 2151–2165, New York, NY, USA. Association for Computing Machinery.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2025. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *Preprint*, arXiv:2406.14805.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "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*, FAccT '24, page 822–835, New York, NY, USA. Association for Computing Machinery.

Svenja Kranich. 2011. To hedge or not to hedge: the use of epistemic modal expressions in popular science in english texts, english–german translations, and german original texts. *Text & Talk*, 31(1).

Lea Krause, Wondimagegnhue Tsegaye Tufa, Selene Báez Santamaría, Angel Daza, Urja Khurana, and Piek Vossen. 2023. Confidently Wrong: Exploring the Calibration and Expression of (Un)Certainty of Large Language Models in a Multilingual Setting. In *Proceedings of the 1st Workshop on Muiltimodal, Multilingual NLG (MM-NLG)*.

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*.

Louis Kwok, Michal Bravansky, and Lewis Griffin. 2024. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. In *First Conference on Language Modeling*.

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

Gabrielle Kaili-May Liu, Gal Yona, Avi Caciularu, Idan Szpektor, Tim G. J. Rudner, and Arman Cohan. 2025. MetaFaith: Faithful natural language uncertainty expression in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29612–29656, Suzhou, China. Association for Computational Linguistics.

Rosina Márquez Reiter. 2002. A contrastive study of conventional indirectness in spanish: Evidence from peninsular and uruguayan spanish. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 12(2):135–151.

Pedro Angel Martín Martín. 2008. The mitigation of scientific claims in research papers: A comparative study. *IJES: international journal of English studies*, 8(2):133–152.

Errapel Mejias-Bikandi. 1998. Pragmatic presupposition and old information in the use of the subjunctive mood in spanish. *Hispania*, 81(4):941–948.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872. Place: Cambridge, MA.

Rudolf Muhr. 1994. "Entschuldigen Sie Frau Kollegin...": Sprechaktrealisierungsunterschiede an Universitäten in Österreich und Deutschland. In G. Bachleitner-Held, editor, *Verbale Interaktion*, pages 126–143. Dr. Kovac, Hamburg.

Lawrence D. Phillips and C. N. Wright. 1977. *Cultural Differences in Viewing Uncertainty and Assessing Probabilities*, pages 507–519. Springer Netherlands, Dordrecht.

Neil Rathi, Dan Jurafsky, and Kaitlyn Zhou. 2025. Humans overrely on overconfident language models, across languages. In *Second Conference on Language Modeling*.

Katharina Reinecke and Abraham Bernstein. 2011. Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces. *ACM Trans. Comput.-Hum. Interact.*, 18(2).

Klaus Peter Schneider and Anne Barron. 2008. *Variational pragmatics: A focus on regional varieties in pluricentric languages*. John Benjamins Publishing.

Paul A. Shuper, Richard M. Sorrentino, Yukiko Otsubo, Gordon Hodson, and A. M. Walker. 2004. A theory of uncertainty orientation: Implications for the study of individual differences within and across cultures. *Journal of Cross-Cultural Psychology*, 35(4):460–480.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. 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*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. 2024. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15440–15459, Bangkok, Thailand. Association for Computational Linguistics.

Dennis Ulmer, Alexandra Lorson, Ivan Titov, and Christian Hardmeier. 2025. Anthropomimetic uncertainty: What verbalized uncertainty in language models is missing. *arXiv preprint arXiv:2507.10587*.

Ignacio Vázquez Orta. 2010. A contrastive analysis of the use of modal verbs in the expression of epistemic stance in business management research articles in english and spanish. *Ibérica, Revista de la Asociación Europea de Lenguas para Fines Específicos*, (19):77–95.

Eva Thue Vold. 2006. Epistemic modality markers in research articles: a cross-linguistic and cross-disciplinary study. *International journal of applied linguistics*, 16(1):61–87.

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*.

Erika Werlen. 1998. *Sprache, Kommunikationskultur und Mentalität*. De Gruyter, Berlin, Boston.

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. In *The Twelfth International Conference on Learning Representations*.

Zhengtao Xu, Tianqi Song, and Yi-Chieh Lee. 2025. Confronting verbalized uncertainty: Understanding how llm's verbalized uncertainty influences users in ai-assisted decision-making. *International Journal of Human-Computer Studies*, 197:103455.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

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

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Kaitlyn Zhou, Jena D. 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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Extended Cross-Linguistic and Cross-Cultural Overview

Frameworks Two frameworks inform our locale-level analysis. Contrastive pragmatics documents how English, Spanish, and German encode and deploy epistemic markers at the textual level (Kranich, 2011; Martín, 2008; Vold, 2006). Hofstede's Uncertainty Avoidance Index (UAI) captures broader cultural orientations towards ambiguity (Hofstede, 2001). Hofstede states that high uncertainty avoidance "leads to a reduction of ambiguity. Uncertainty avoiding cultures shun ambiguous situations. People in such cultures look for structure in their organisations, institutions and relationships, which makes events clearly interpretable and predictable." (Hofstede, 1991). While the Hofstede framework has been critiqued for its methodological limitations (Baskerville, 2003), it remains a useful heuristic for predicting broad patterns of communicative and cultural preference (Reinecke and Bernstein, 2011).

A finding from this literature is that hedging frequency and cultural uncertainty avoidance are not necessarily aligned. Cultures that score high on uncertainty avoidance do not necessarily hedge more. In many cases they hedge less, preferring instead to communicate with precision or direct admission of uncertainty over lexical mitigation (House, 2006; Kranich, 2011).

English

General English uses a rich system of epistemic hedging. English texts employ epistemic modal markers at higher rates than German or Spanish texts performing equivalent functions (Kranich, 2011; Martín, 2008). English speakers prefer markers of moderate certainty or mere possibility ("might," "perhaps," "I think," "could") to manage face-threats and signal epistemic caution (Hyland, 1998). Hedging in English serves both politeness and knowledge-management functions, letting speakers present claims as revisable while acknowledging alternatives (Brown and Levinson, 1987).

United Kingdom (GBR) British English hedges more than the other two English varieties in our study. British communicative norms favour negative politeness, which minimises imposition and preserves face through indirectness (Brown and Levinson, 1987). Blum-Kulka and Olshtain (1984)

found that British speakers prefer conventionally indirect strategies and use a wide range of internal mitigating devices. Downtoners ("rather," "quite," "a bit") and tentative constructions ("I should think," "one might suggest") are common, alongside understatement and litotes. Intensifier use also differs from American English: "quite" works as a downtoner in British usage but as an amplifier in American usage. Despite all this hedging, the UK has a low UAI (35), so British hedging is likely driven by politeness rather than by anxiety about ambiguity.

United States (USA) American English has moderate UAI (46) and favours positive politeness over the negative politeness more typical of British English (Brown and Levinson, 1987). Uncertainty is expressed through moderate markers ("I think," "I believe," "probably") rather than through extensive indirectness. Hedging is functional rather than ritualised. The relatively low UAI is consistent with a tolerance for ambiguity that allows direct expression of both confidence and doubt.

Australia (AUS) Australian English sits between British indirectness and American directness. Australian speakers prefer conventional indirectness as their dominant strategy, but they switch directness levels less across social contexts than speakers of Hebrew, German, French, or Spanish (Blum-Kulka and Olshtain, 1984). Communication norms value egalitarianism and informality, and for uncertainty this translated to expressions that would read as casual by other varieties ("could be," "reckon," "not sure") rather than through elaborate hedging (Goddard and Cramer, 2016). UAI is moderate (51).

German

General German combines moderate-to-high uncertainty avoidance with a communicative preference for directness. German texts use epistemic modal markers less frequently than English texts (Kranich, 2011). The two languages also differ in marker type: English favours markers of mere possibility ("maybe," "might"), while German uses markers of high probability ("wahrscheinlich," "vermutlich") more often. Kranich (2011) attributes this to a contrast between German content-orientation and English addressee-orientation. German speakers also interpret probability expressions differently, assigning approximately 66% likelihood to "probable" compared to higher thresholds among American speakers (Doupnik and Richter, 2003, 2004). Rathi et al. (2025) confirm these norms at scale, finding that LLMs generate more certainty markers in German than in other languages.

Switzerland — German-speaking (CHE) Swiss German communication is consensus-oriented, with a preference for indirectness when formulating directives or criticism (Werlen, 1998). The relational dimension of interaction is more prominent than in German German. Switzerland's UAI (58) is lower than Germany's, and uncertainty is often framed collaboratively rather than asserted one-sidedly.

Germany (DEU) German communicative culture values Sachlichkeit (objectivity, matter-of-factness) and directness. House (2006) identifies five dimensions on which German diverges from English, including directness over indirectness and content-orientation over addressee-orientation. Within the German frame, direct strategies are not impolite, instead they are read as a mark of respect and transparency. The UAI score is moderately high (65), but Germans cope with uncertainty through planning and precision rather than through linguistic hedging. When uncertain, speakers either use high-probability markers or acknowledge ignorance directly ("Ich weiß es nicht"). Vague hedging without informational content is read as imprecise rather than polite.

Austria (AUT) Austrian German diverges from German German pragmatically. Muhr (1994) shows that Austrian apologies adopt a hearer perspective more than German German apologies, are more explicit, and use more upgrading. Face-threatening explanations are less frequent in Austrian German than in German German. The UAI (70) is somewhat higher than Germany's (65), which is consistent with a culture that combines uncertainty avoidance with greater attention to relational management. Direct comparative work on epistemic marker use between Austrian and German German is limited (Schneider and Barron, 2008).

Spanish

General Spanish has a structural resource for uncertainty expression that English and German lack: the subjunctive mood. The subjunctive is systematically deployed in complement clauses governed by verbs of doubt, uncertainty, and non-assertion

(Faulkner, 2022; Mejias-Bikandi, 1998). Verbs of uncertainty take either subjunctive or indicative complements, with the indicative signalling higher speaker commitment. This alternation works as a grammatical epistemic signal that operates below the level of lexical hedging.

English uses hedging devices at higher frequencies than Spanish across academic disciplines (Martín, 2008). In business management research articles, native English writers use modal verbs at 7.0 per thousand words, compared to 5.3 for Spanish writers using English as a second language, and the two groups also differ in their hedge and booster patterns (Vázquez Orta, 2010). Spanish academic culture relies more on grammatical mood and less on lexical hedging to manage epistemic commitment (Alonso Almeida and Carrió Pastor, 2017).

Mexico (MEX) Mexican Spanish emphasises indirectness, deference, and the avoidance of offence. Félix-Brasdefer (2005) showed that conventional indirectness is the most common requesting strategy in situations involving power asymmetry or social distance. In refusal interactions, face needs are oriented toward the group rather than the individual (Félix-Brasdefer, 2006). Uncertainty expression is layered with diminutives, conditional verb forms ("podría ser," "sería posible"), and politeness frames, and expressing a flat negative is culturally marked. Despite a high UAI (82), Mexican Spanish sits at the negative politeness end of the spectrum.

Spain (ESP) Peninsular Spanish combines a direct communicative style with very high uncertainty avoidance (UAI 86). The variety favours explicit, unambiguous expressions, including unmitigated imperatives, and limits communication to the message itself without much ritual softening (House, 2006). Uncertainty is expressed directly: "no lo sé," "no estoy seguro," "puede ser," relying on grammatical mood rather than elaborate hedging. The apparent paradox of high UAI and low lexical hedging resolves once Spanish uncertainty avoidance is understood as a preference for clarity rather than for indirectness. Spaniards cope with uncertainty by seeking definitive answers, not by hedging.

Uruguay (URY) Uruguay sits highest on the UAI scale (100). Márquez Reiter (2002) compared Uruguayan and Peninsular Spanish requests directly, and found Uruguayan requests significantly more tentative. The tentativeness is achieved through more frequent external modifications (presequences, justifications, disarmers) and a higher rate of internal downgrading devices (conditional verb forms, diminutives, softening particles). Rioplatense Spanish also has non-standard mood alternation: in future-framed adverbial clauses, the indicative can occur where prescriptive grammar predicts the subjunctive, with the indicative conveying certainty and the subjunctive conveying doubt (Hoff, 2020) Uruguayan speakers thus have a fine-grained grammatical gradient of epistemic commitment. The combination of high UAI, elaborate tentativeness, and gradient mood alternation makes Uruguayan Spanish the most mitigation-heavy variety in our study.

A.2 Prompt design ablation study

We conducted a prompt design ablation on the validation split to select the best prompting configuration before running full test-set experiments. The study compared four conditions across all nine models.

- **A** (baseline): original instructions, model's default system prompt unchanged.
- **B**: updated instructions + empty system prompt (suppresses the model's default where possible).
- **C**: updated instructions + model's default system prompt retained.
- **D**: updated instructions moved into the system turn; user turn contains only the few-shot examples and the question.

The updated prompt (conditions B–D) adds two sentences to every locale's instructions relative to the baseline: (1) *"Output only the letter and the comment,"* to reduce verbose, off-task generation; and (2) *"The comment should express certainty or uncertainty about the answer, as seen in the examples,"* to better target the expected comment style.

Models differ in their system-prompt support, as summarised in Table 3. Gemma does not support the system role and therefore falls back to condition B when condition D is applied; Bloomz has no chat template and is only evaluated under conditions A and B.

Results are summarised in Figure 7. The updated prompt (B–D) increased comment generation rates

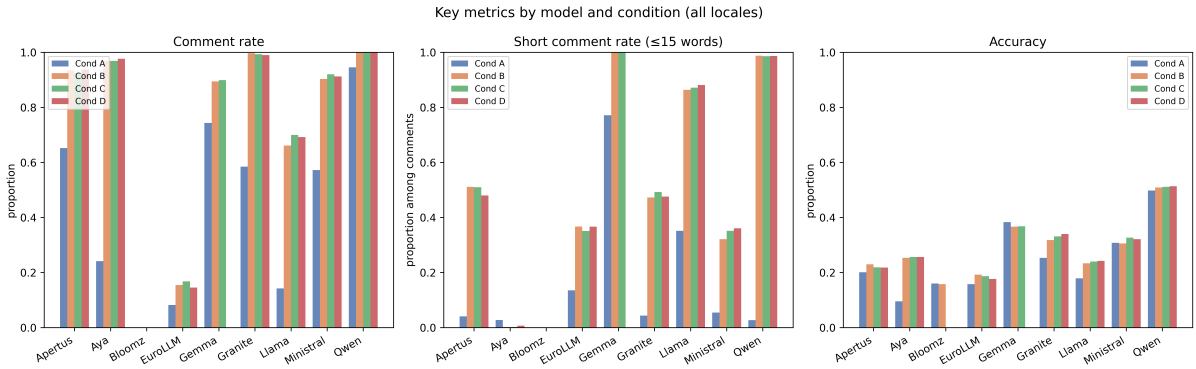

Figure 7: Results for ablation study comparing prompting design. Metrics shown are comment rate, short-comment rate, and accuracy.

Model	Default	Suppress?	Sys. role?
Qwen	✓	✓	✓
Apertus	✓	✓	✓
Ministral	✓	✓	✓
Granite	✓	✓	✓
Llama	date (hardcoded)	×	✓
EuroLLM	—	—	✓
Aya	—	—	✓
Gemma	—	—	×
Bloomz	*no chat template*		

Table 3: System prompt handling per model. **Default**: model has a default system prompt. **Suppress?**: passing an empty string removes it. **Sys. role?**: system role is supported by the chat template.

across all models as well as accuracy, though more moderately so. Differences among B, C, and D were small (typically <2%), with no single condition dominating across all models.

A.3 Annotation of epistemic markers

We extracted the epistemic expressions from the generated answer comments. As quality control, we filtered out invalid comments (empty, or non alphabetical) as well as comments longer than 15 words. We also normalize the expressions to change references to specific answer options: "I'm pretty sure it's G." → "I'm pretty sure it's [X]." We collected all unique expressions produced per locale across runs and models. We then set up an annotation task to label each expression.

Annotation task We asked annotators to rate how certain or uncertain a speaker sounded based only on how they express themselves, not on whether they thought their answer was correct. Each expression was valued on a 7-point Likert scale ranging from 1 (very uncertain) to 7 (very certain), similar to Rathi et al..

Annotators We recruited 3 annotators per country via Prolific. The selection criteria required participants to be native speakers of the target language, hold the nationality of the target country, and have spent most of their life before the age of 18 in that country. Annotators were compensated at £9 per hour.

Annotation quality We included two quality control mechanisms. First, we added attention checks consisting of one clearly certain expression ("It's definitely, 100% B.", expected range $[5, 7]$) and one clearly uncertain expression ("I have absolutely no idea, maybe B?", expected range $[1, 3]$). As per Prolific guidelines, we discarded annotators who failed 2 or more attention checks. Second, some questions were duplicated which allowed us to measure intra-annotator consistency.

Annotation pilot and refinement Using the 20 items left from the validation set, we first performed an exploratory phase to investigate the generation epistemic expressions by the nine initially selected models. Llama performed best in generating the desired epistemic markers, and thus was selected for an initial round of human annotation. We selected the top-300 expressions per country for a pilot annotation. Given the results of the pilot annotation, we adapted the annotation task parameters and applied them to the model generations from the test set. Attention checks were updated from every 40 questions to every 50 questions. The duplication rate went down from 10% of questions in the pilot and 5% of questions in the refined annotation task. Finally, the top-450 expressions across models per country were selected for annotation.

Annotation outcomes The resulting inter and intra-annotator agreement scores are shown in Ta-

ble 4. The coverage of both pilot and refined annotations compared to the total generations is shown in Figure 8. All model generations were covered to at least 90%.

Country	N	Intra-κ	Inter-κ
English			
AUS	3/3	0.444	0.285
GBR	3/4	0.424	0.225
USA	3/3	0.414	0.066
Spanish			
ESP	3/5	0.546	0.359
MEX	3/6	0.339	0.156
URY	2/5	0.374	0.175
German			
AUT	3/7	0.515	0.216
CHE	2/6	0.418	0.270
DEU	3/4	0.410	0.273

Table 4: Inter- and average intra-annotator agreement (Fleiss' κ). N = passed/total annotators.

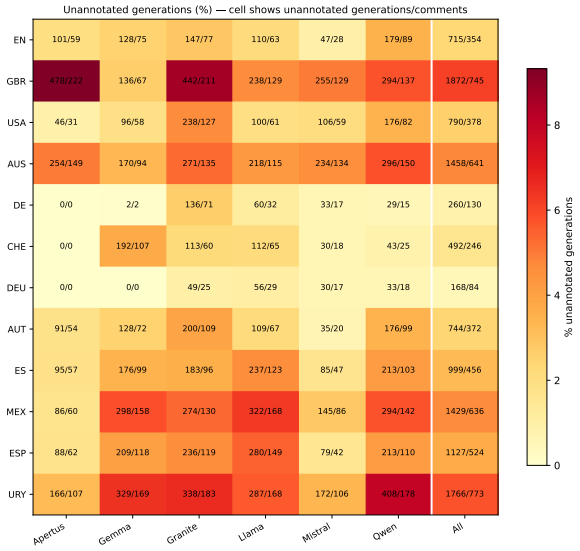

Figure 8: Percentage of model generations whose epistemic expression was not covered by the human annotation survey, per model (columns) and locale (rows). Cell text shows unannotated generations / unique unannotated expressions.

A.4 Correlations across all Hofstede dimensions

Figure 9 shows the mean values per metric against locales ordered by UAI. No clear monotonic pattern is shown across locales.

Furthermore, Table 5 reports Spearman correlations between Hofstede dimensions and all metrics.

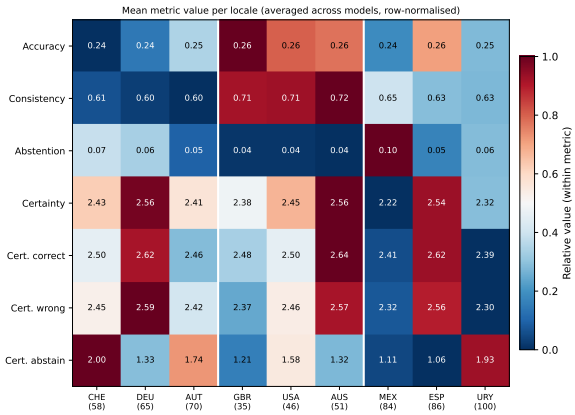

Figure 9: Mean value per metric and locale, averaged across models. Each row is independently normalized to highlight within-metric variation across locales.

The Bonferroni-corrected significance threshold for 140 comparisons is $p < 0.00036$. No result approaches this threshold.

		UAI		PDI		IVR		MAS		IDV	
Metric	**Lang**	ρ	p	ρ	p	ρ	p	ρ	p	ρ	p
Accuracy	All	-0.04	0.77	+0.07	0.57	+0.07	0.61	-0.07	0.58	-0.09	0.46
	DE	+0.14	0.55	-0.05	0.84	-0.09	0.69	+0.05	0.84	-0.11	0.64
	EN	-0.01	0.97	+0.01	0.97	-0.02	0.93	+0.01	0.97	-0.01	0.97
	ES	-0.09	0.69	+0.01	0.97	+0.01	0.97	+0.09	0.69	-0.01	0.97
Consistency	All	-0.13	0.32	+0.09	0.47	+0.17	0.17	-0.09	0.47	-0.14	0.29
	DE	+0.02	0.93	-0.07	0.77	+0.05	0.84	+0.07	0.77	-0.05	0.83
	EN	+0.03	0.88	+0.01	0.97	+0.02	0.92	-0.03	0.88	-0.01	0.97
	ES	-0.10	0.68	+0.08	0.74	+0.08	0.74	+0.10	0.68	-0.08	0.74
Abstention	All	+0.17	0.19	+0.13	0.30	+0.01	0.92	+0.01	0.96	-0.11	0.39
	DE	+0.02	0.93	-0.02	0.92	+0.00	0.98	+0.02	0.92	-0.03	0.91
	EN	-0.04	0.87	-0.05	0.84	+0.01	0.97	+0.04	0.87	+0.05	0.84
	ES	-0.16	0.48	+0.15	0.50	+0.15	0.50	+0.16	0.48	-0.15	0.50
Expressed certainty	All	-0.14	0.28	-0.20	0.12	-0.12	0.33	-0.01	0.96	+0.23	.07
	DE	+0.03	0.90	+0.14	0.53	-0.17	0.45	-0.14	0.53	+0.07	0.77
	EN	+0.16	0.48	-0.05	0.84	+0.21	0.36	-0.16	0.48	+0.05	0.84
	ES	+0.04	0.87	-0.39	.08	-0.39	.08	-0.04	0.87	+0.39	.08
Cert. correct	All	-0.10	0.42	-0.13	0.30	-0.05	0.70	+0.02	0.85	+0.17	0.18
	DE	+0.02	0.93	+0.08	0.74	-0.10	0.68	-0.08	0.74	+0.03	0.89
	EN	+0.23	0.31	-0.06	0.80	+0.29	0.20	-0.23	0.31	+0.06	0.80
	ES	-0.14	0.53	-0.26	0.25	-0.26	0.25	+0.14	0.53	+0.26	0.25
Cert. wrong	All	-0.11	0.39	-0.17	0.17	-0.13	0.32	+0.02	0.87	+0.22	.08
	DE	+0.04	0.87	+0.15	0.50	-0.19	0.40	-0.15	0.50	+0.07	0.77
	EN	+0.18	0.43	-0.01	0.97	+0.19	0.40	-0.18	0.43	+0.01	0.97
	ES	-0.09	0.71	-0.32	0.16	-0.32	0.16	+0.09	0.71	+0.32	0.16
Cert. abstain	All	+0.02	0.88	**-0.30**	**.04**	-0.19	0.19	+0.11	0.47	+0.24	.10
	DE	-0.21	0.47	-0.11	0.70	+0.35	0.22	+0.11	0.70	+0.06	0.85
	EN	+0.04	0.89	+0.31	0.26	-0.33	0.23	-0.04	0.89	-0.31	0.26
	ES	**+0.62**	**.00**	+0.11	0.65	+0.11	0.65	**-0.62**	**.00**	-0.11	0.65

Table 5: Spearman ρ and p-values for all five Hofstede dimensions against all metrics, pooled across models and shown for all locales (All) and per language group (DE, EN, ES). Bold marks nominally significant results ($p < 0.05$); none survive Bonferroni correction ($\alpha_{\text{corr}} = 0.00036$ for 140 comparisons).