Tail Smoothing and Cross-Lingual Volatility: Evaluating Estimative Uncertainty in Large Language Models for Brazilian Portuguese

Renato O. Miyaji[1], **Pedro L. P. Corrêa**[1]

[1]Escola Politécnica – Universidade de São Paulo (USP)

{re.miyaji, pedro.correa}@usp.br

Abstract. *This study evaluates how LLMs interpret Words of Estimative Probability (WEPs) in Brazilian Portuguese compared to English. We translated an English benchmark and compared multilingual models (GPT-5.1, Gemini 3 Flash) against a region-specific model (Sabiá 4). Our findings reveal a "tail smoothing" phenomenon, where models systematically compress extreme probabilities. Notably, while Gemini 3 Flash demonstrated remarkable cross-lingual stability, GPT-5.1 exhibited significant calibration degradation. Counterintuitively, when measured against the English human baseline, Sabiá 4 displayed the most aggressive distribution compression. These results suggest a complex dynamic: either linguistic fine-tuning does not ensure alignment, or it actively captures a culturally specific pragmatic ambiguity in Brazilian Portuguese. This exposes the vulnerabilities and nuances of deploying deploying LLMs in nuanced semantic tasks in Brazilian Portuguese.*

1. Introduction

Words of estimative probability (WEPs), such as "maybe", "likely," or "probably not" are fundamental components of natural language used to communicate varying degrees of uncertainty [Erev and Cohen 1990, Juanchich and Sirota 2020]. In human communication, these linguistic markers possess complex probabilistic semantics that rely heavily on context, culture, and shared cognitive understanding [Vlasyan et al. 2018]. As artificial intelligence systems increasingly mediate critical decision-making processes, the calibration of these expressions with precise numerical estimates has become a vital area of study to ensure safe and reliable human-computer interaction [Tang et al. 2026].

The proliferation of Large Language Models (LLMs) has amplified the need for robust uncertainty communication in natural language processing. While modern LLMs generate fluent and seemingly confident text, their internal statistical representations often diverge from human cognitive interpretations of uncertainty [Schneider 2016]. Understanding whether these models assign the same probabilistic weight to WEPs as human speakers do is essential for preventing miscommunication, particularly in high-stakes domains where the precise likelihood of an event is crucial.

Recent research has highlighted significant gaps in how LLMs interpret and deploy WEPs compared to human baselines. For instance, [Tang et al. 2026] demonstrated that while established LLMs show some alignment with human estimates for specific WEPs in English, their performance degrades when confronted with cross-linguistic

prompts or complex contextual scenarios. Their study emphasized the limits of statistical language models in reproducing the subtleties of communication under uncertainty, pointing to a pressing need for broader evaluations across diverse linguistic landscapes.

Despite these advances, the vast majority of studies on estimative uncertainty in LLMs remain heavily skewed toward English [Shorinwa et al. 2025] and, to a lesser extent, languages like Mandarin Chinese [Tang et al. 2026]. Portuguese, specifically Brazilian Portuguese, remains largely underexplored in this context, despite being a widely spoken language with its own rich set of pragmatic and semantic nuances regarding uncertainty [Freitag et al. 2022]. This gap poses a critical challenge for the equitable and accurate deployment of LLMs for Portuguese speakers, as cultural and linguistic variations can significantly alter the interpretation of probabilistic language [Schuck et al. 2025].

To address these challenges, this study is guided by two primary Research Questions: (RQ1) *"How do generalized commercial LLMs compare to language-specific models in aligning with human calibrations of WEPs in Brazilian Portuguese?"* and (RQ2) *"What are the downstream vulnerabilities introduced by misaligned uncertainty representations in this language?"*. Our main contributions include the adaptation of an English WEP benchmark into Brazilian Portuguese and a comparative analysis of general-purpose models, such as GPT-5.1 [OpenAI 2026] and Gemini 3 Flash [Google DeepMind 2026], and Sabiá 4 [Laitz et al. 2026]. By evaluating these systems, we highlight the practical limitations and inherent risks of deploying large commercial models for nuanced linguistic tasks in Portuguese. These findings are particularly critical for high-stakes applications, such as decision-making pipelines and LLM-as-a-Judge frameworks, where the subtle misinterpretation of probabilistic language can lead to skewed evaluations and unsafe outcomes. It is crucial to clarify that this study does not treat the English human baseline as the "ground truth" for Brazilian Portuguese cognition. Given that estimative uncertainty is inherently tied to cultural pragmatics, any divergence from this baseline by the models in Portuguese is analyzed through a dual lens: it may represent a breakdown in the model's cross-lingual semantic alignment, or it may indicate that the model is actively capturing a culturally distinct probabilistic mapping.

2. Related Works

Recent research on Uncertainty Quantification (UQ) in LLMs has expanded beyond traditional parametric metrics to address the unique complexities of natural language generation. While modern UQ frameworks categorize uncertainty into multiple dimensions a critical area of investigation is how models linguistically communicate this uncertainty to users [Shorinwa et al. 2025]. Studies have demonstrated that despite possessing internal statistical uncertainty, LLMs frequently generate highly decisive and persuasive responses, failing to faithfully hedge their assertions even when explicitly prompted to express doubt. This discrepancy between a model's intrinsic confidence and its linguistic decisiveness highlights a significant vulnerability in human-AI interaction, as overly confident language can easily lead to unwarranted user reliance on inaccurate outputs [Yona et al. 2024].

To systematically evaluate this communicative gap, researchers have focused on how models interpret WEPs. Empirical evaluations comparing LLM probability distributions to human baselines have revealed significant misalignments, particularly for mid-

range, context-dependent WEPs, whereas extreme terms like “almost certain” tend to show better alignment. Furthermore, a model’s statistical interpretation of these words is highly sensitive to the prompt’s context, with measurable distribution shifts occurring when translating between languages or when introducing gender-specific narratives [Tang et al. 2026]. Building upon these foundational findings, our work extends the evaluation of WEPs into the underexplored landscape of Brazilian Portuguese to assess whether localized fine-tuning can effectively mitigate these cross-lingual and communicative misalignments.

3. Methodology

This study’s experimental design is structured to replicate and extend the foundational framework established by [Tang et al. 2026] for evaluating WEPs. Our primary objective is to systematically assess how recent advancements in LLMs impact the calibration of estimative uncertainty, particularly when transitioning from English to Brazilian Portuguese. To achieve this, we developed a bilingual experimental pipeline that mirrors the original study’s conditions while introducing state-of-the-art models to the evaluation matrix to test their pragmatic alignment.

To establish a rigorous evaluation framework, our methodology leverages the dataset constructed by [Tang et al. 2026]. This dataset is built upon the empirical foundation of the WEPs originally utilized in the survey by Fagen-Ulmschneider [Fagen-Ulmschneider 2019]. Specifically, it assesses a standardized set of seventeen expressions that span the probability spectrum: *almost certain, highly likely, very good chance, probable, likely, we believe, probably, better than even, about even, we doubt, improbable, unlikely, probably not, little chance, almost no chance, highly unlikely*, and *chances are slight*. By employing these linguistic triggers, our study ensures that the evaluation captures a comprehensive range of estimative uncertainty, mirroring established baselines of human cognitive interpretation. Furthermore, the dataset rigorously tests contextual resilience through various types of structured statements designed to frame these WEPs. These include Concise Narrative Contexts, which present simple sentences to provide a direct background for the WEPs, and Extended Narrative Contexts, which introduce more prolonged scenarios with increased contextual information and clauses [Tang et al. 2026]. By adapting these diverse statement structures into Brazilian Portuguese, our study can investigate how varying levels of contextual complexity and specific narrative backgrounds influence the probabilistic estimations generated by the models in a cross-lingual setting.

The first critical phase of our methodology involved the adaptation of the dataset. To ensure rigor, we conducted a systematic translation of the original English statements and their associated WEPs into Brazilian Portuguese. This process went beyond word-for-word translation; it required careful consideration of regional nuances to preserve the ordinal probabilistic weight and contextual subtleties of each expression. Specifically, the translation was performed by two native Brazilian Portuguese speakers with proficiency in English and backgrounds in natural language processing and linguistics. The procedure involved an initial independent forward-translation phase, followed by a consensus meeting to resolve discrepancies.

For the human baseline comparison, our study utilizes the extensive em-

pirical data gathered in the original English survey by Fagen-Ulmschneider [Fagen-Ulmschneider 2019]. We explicitly note that we do not treat these English human probability judgments as the cognitive "ground truth" for Brazilian Portuguese speakers, as probabilistic intuition is heavily influenced by cultural and linguistic factors. Instead, we employ the English survey data as a reference point for cross-lingual stability. This approach allows us to measure whether models, which are predominantly trained on English corpora, maintain their internal uncertainty calibration when prompts are translated into Portuguese. This choice is necessitated by the current scarcity of established research on estimative uncertainty specifically within the Brazilian Portuguese context. While conducting a primary human survey in Portuguese is beyond the scope of this work, our methodology rigorously evaluates the semantic degradation and transferability of WEPs across languages.

To capture the current landscape of artificial intelligence, we updated the model selection to include three LLMs: GPT-5.1 [OpenAI 2026], Gemini 3 Flash [Google DeepMind 2026], and Sabiá 4 [Laitz et al. 2026]. GPT-5.1 and Gemini 3 Flash represent the frontier of generalized models, boasting massive but predominantly English-centric multilingual training corpora. In contrast, Sabiá 4 was selected as a baseline; it is a model explicitly designed and fine-tuned for Brazilian Portuguese, possessing a deeper focus on the cultural and linguistic patterns of Brazilian language. The inclusion of Sabiá 4 is central to addressing our research questions regarding regional alignment. By comparing this region-specific model against generalized ones, we aimed to evaluate whether localized training paradigms yield a higher degree of calibration with human probability estimates in Portuguese. We specifically test the hypothesis that Sabiá 4 will demonstrate less variance and better distributional accuracy when processing portuguese WEPs, effectively minimizing the semantic degradation often observed when generalized models process non-English probabilistic language.

Our prompting strategy rigorously reproduced all experimental scenarios from the original study across both languages. Every experiment was executed in both English and Portuguese. This dual-language execution allows us to isolate the specific effect of language on model performance and to determine if the commercial models exhibit altered confidence intervals or pragmatic failures when operating outside their primary linguistic domain.

During the inference phase, the models were prompted to assign precise numerical probability values to the provided statements containing the WEPs. To ensure consistency and prevent the models from generating evasive or overly verbose responses, we utilized standardized zero-shot prompting with strict output constraints. The models were instructed to return specific percentage values, which were then systematically extracted, cleaned, and aggregated to construct probability distributions for each combination of WEP, model, and language. Finally, the quantitative evaluation of these distributions focuses on measuring the statistical divergence between the LLM-generated estimates and the human survey baseline. By analyzing metrics such as median values, interquartile ranges, and overall distribution shapes, we can objectively quantify the degree of misalignment.

This experimental design not only tests the raw capabilities of modern LLMs but also rigorously exposes the potential vulnerabilities that arise when these systems are deployed for nuanced probabilistic reasoning in Brazilian Portuguese. To rigorously eva-

luate the degree of alignment between the estimates generated by the language models and human perception, we adopted an analytical approach based on robust statistical metrics, focusing on both the shape of the distributions and stochastic equality, following [Tang et al. 2026]. The first stage of our quantitative analysis utilized the Kullback-Leibler (KL) Divergence to measure the dissimilarity between the probability distributions generated by the LLMs and the baseline reference distribution obtained from the Fagen-Ulmschneider human survey [Fagen-Ulmschneider 2019]. By calculating the KL divergence for each WEP in both languages, we were able to objectively rank the models' performance. Lower KL values indicate that the model more faithfully captured the variance and concentration of human estimates, allowing us to directly compare the semantic degradation between the English tests and their counterparts translated into Brazilian Portuguese.

However, since the probability estimates provided by both humans and language models frequently exhibit asymmetrical and non-normal distributions, traditional parametric tests for comparing means would be inadequate [Tang et al. 2026]. To overcome this limitation and achieve greater statistical rigor, we employed the Brunner-Munzel test as our primary inference tool. The Brunner-Munzel test is a non-parametric method designed to compare samples from two independent populations without assuming equal variances or normal distributions, making it ideal for the stochastic nature of WEP data.

The application of the Brunner-Munzel test allowed us to evaluate the hypothesis of stochastic equality, in other words, to verify whether the probabilities assigned by an LLM tend to be systematically higher or lower than those assigned by humans. By extracting the p-values from this test, we could determine with statistical precision which models presented significant divergences from the human baseline. This cross-analysis between the KL divergence and the Brunner-Munzel test provided the empirical foundation necessary to conclude whether Sabiá 4 indeed outperforms generalist commercial models in the fine-grained calibration of uncertainties in the Portuguese language.

4. Results

Figure 1 summarizes the experimental results.

4.1. Experiments in English

The comparative analysis between the human survey baseline and the language models operating in English reveals a persistent phenomenon of "tail smoothing" across the evaluated LLMs, as presented in Table 1. When analyzing the generated probability distributions against the Fagen-Ulmschneider survey data [Fagen-Ulmschneider 2019], it becomes evident that while the models generally respect the ordinal ranking of WEPs, they often struggle to capture the extreme values of human intuition. Both KL divergence scores and Brunner-Munzel stochastic equality tests indicate that the models tend to compress their estimates toward the center, resulting in medians that are less extreme than those provided by human respondents for both very low and very high-probability expressions.

An examination of GPT-5.1's performance highlights significant misalignment, particularly in the maximum certainty zone. For high-probability WEPs such as "almost certain," human respondents anchored their estimates at a median of 95%, whereas GPT-5.1 provided a statistically distinct median of 90%. The Brunner-Munzel test confirms this divergence as highly significant ($p < 0.001$), accompanied by a pronounced

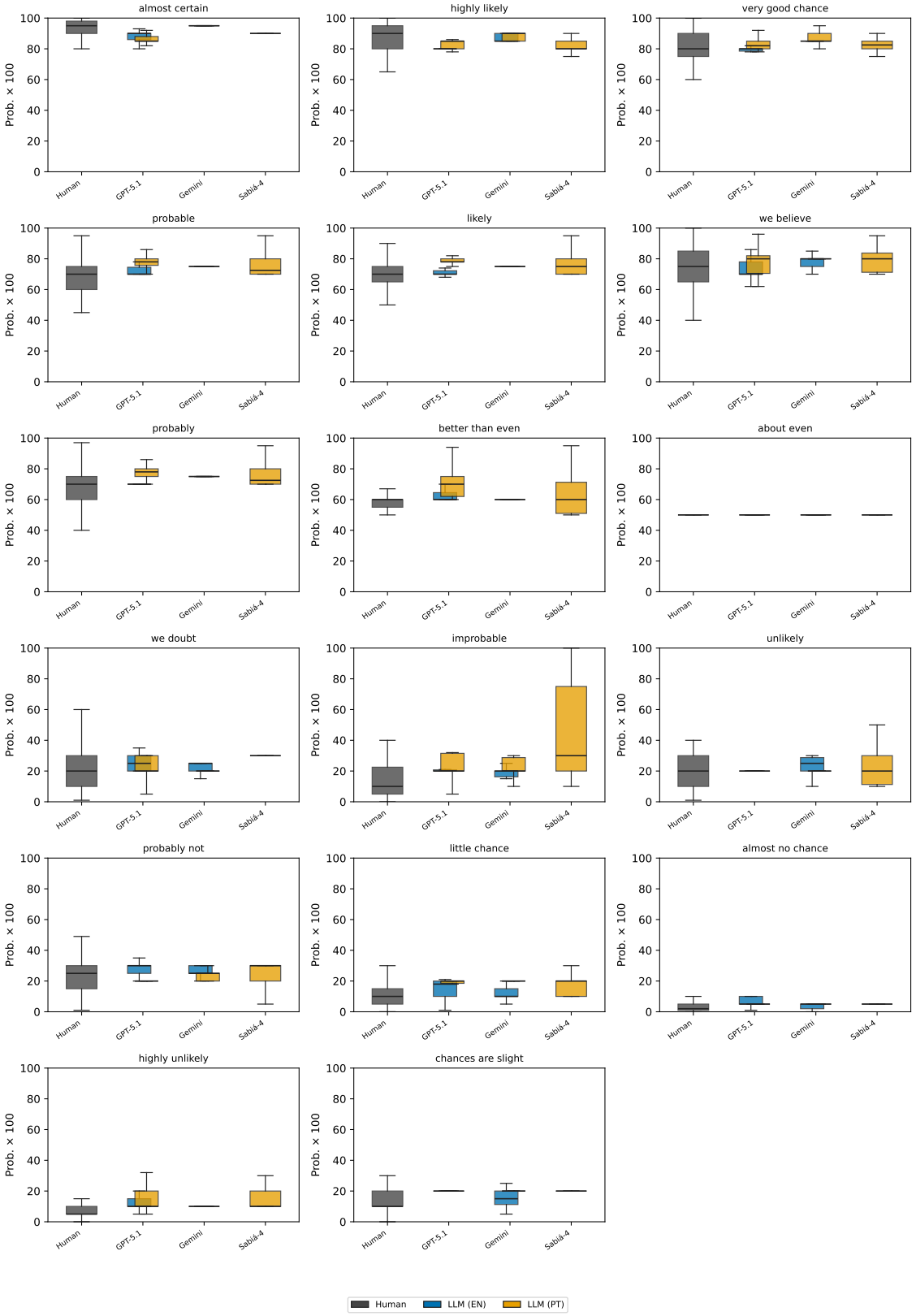

Figura 1. Comparative boxplots illustrating the probability distributions assigned to seventeen words of estimative probability. The charts contrast the Human baseline with the estimates generated by GPT-5.1, Gemini, and Sabiá-4. Model responses are distinguished by language, with LLM (EN) indicating English prompts and LLM (PT) indicating Portuguese prompts. This visualization highlights cross-lingual distribution shifts and demonstrates the ”tail smoothing”effect, where model estimations frequently compress toward the center compared to the broader variance of human intuition.

Tabela 1. English baseline: human medians vs. GPT-5.1 and Gemini 3 Flash. p_{BM}**: Brunner–Munzel test (human vs. model); KL:** $\mathrm{KL}(P_{\mathrm{human}} \| P_{\mathrm{model}})$**; AMD: absolute median difference (percentage points).**

Expression	M_{hum}	M_{GPT}	M_{Gem}	$p_{\mathrm{BM}}^{\mathrm{GPT}}$	$p_{\mathrm{BM}}^{\mathrm{Gem}}$	$\mathrm{KL}^{\mathrm{GPT}}$	$\mathrm{KL}^{\mathrm{Gem}}$	$\mathrm{AMD}^{\mathrm{GPT}}$	$\mathrm{AMD}^{\mathrm{Gem}}$
Almost certain	95	90	95	< 0.001	0.537	14.98	6.65	5	0
Highly likely	90	80	85	< 0.001	0.061	12.90	11.59	10	5
Very good chance	80	80	85	0.014	0.024	2.31	5.21	0	5
Probable	70	70	75	0.300	< 0.001	4.81	10.05	0	5
Likely	70	70	75	0.247	0.012	3.46	7.52	0	5
We believe	75	70	80	0.014	0.116	6.27	3.30	5	5
Probably	70	70	75	0.789	0.002	6.28	10.11	0	5
Better than even	60	60	60	< 0.001	0.002	8.74	10.25	0	0
About even	50	50	50	0.971	0.154	1.19	1.67	0	0
We doubt	20	25	25	0.002	0.342	9.76	9.59	5	5
Improbable	10	20	20	< 0.001	0.002	4.70	8.00	10	10
Unlikely	20	20	25	0.906	0.010	8.99	5.78	0	5
Probably not	25	30	25	0.105	0.596	5.09	7.00	5	0
Little chance	10	18	10	< 0.001	0.037	5.73	2.92	8	0
Almost no chance	2	5	5	< 0.001	0.003	1.76	3.26	3	3
Highly unlikely	5	10	10	< 0.001	< 0.001	1.80	3.64	5	5
Chances are slight	10	20	15	< 0.001	< 0.001	14.33	3.72	10	5

KL divergence score. This underestimation trend continues with the term "highly likely," where the human median of 90% is sharply contrasted by the model's median of 80% ($p < 0.001$). These metrics strongly suggest that GPT-5.1 systematically hesitates to assign the near-absolute probabilistic weight that English speakers naturally attribute to these specific terms.

In contrast, Gemini 3 Flash demonstrates a higher degree of stability and closer alignment with human baselines at the upper extremes of the probability spectrum. For the term "almost certain," Gemini successfully matched the human median of 95%. Furthermore, the Brunner-Munzel test yielded a p-value of 0.537, indicating no statistically significant difference between the human and model distributions, which is also reflected in a substantially lower KL divergence score compared to GPT-5.1. Gemini also showed borderline alignment for "highly likely," achieving a median of 85% with a non-significant divergence ($p = 0.06$), suggesting that its internal representation of high-confidence English WEPs is more robustly calibrated to human expectations.

Despite Gemini's success at the upper tail, both models exhibited shared vulnerabilities when processing low-probability WEPs. For terms like "we doubt" and "improbable," both GPT-5.1 and Gemini 3 Flash generated estimates that were systematically higher than human baselines. Humans assigned medians of 20% and 15% to these terms, respectively. In contrast, both LLMs shifted these estimates upwards, returning medians of 25% for "we doubt" and 20% for "improbable." The Brunner-Munzel tests confirm that these upward shifts are statistically significant across both models. This indicates a shared architectural or training-induced reluctance to generate extremely low probability percentages, effectively truncating the lower tail of the uncertainty distribution.

This baseline English analysis establishes that AMD alone are insufficient for capturing the complex behavior of LLMs. The combined application of KL divergence and

non-parametric testing proves that even when models closely approximate human medians, the underlying distributional shapes can fundamentally differ. While models like Gemini 3 Flash show promising calibration at the high end of the spectrum, the persistent tail-smoothing effects across commercial models in English.

4.2. Experiments in Brazilian Portuguese

The transition from English to Brazilian Portuguese reveals complex shifts in how Large Language Models handle estimative uncertainty, demonstrating that the "tail smoothing" phenomenon persists but manifests differently depending on the model, as presented in Table 2. When comparing the Portuguese WEP estimates to the human survey baseline, it becomes evident that the models still tend to compress probabilities toward the center of the distribution. However, the degree of semantic degradation and the alignment with human expectations vary among the evaluated models, exposing distinct vulnerabilities.

Tabela 2. Portuguese: human respondent medians vs. GPT-5.1, Gemini 3 Flash, and Sabiá 4. Subscripts: G = GPT-5.1, Gm = Gemini, S = Sabia 4.

Expression	M_{hum}	M_{G}	M_{Gm}	M_{S}	p_{G}	p_{Gm}	p_{S}	KL_{G}	$\mathrm{KL}_{\mathrm{Gm}}$	KL_{S}	$\mathrm{AMD}_{\mathrm{G}}$	$\mathrm{AMD}_{\mathrm{Gm}}$	$\mathrm{AMD}_{\mathrm{S}}$
Quase certo	95	85	95	90	< 0.001	0.274	< 0.001	2.54	2.96	1.06	10	0	5
Muito provável	90	85	90	80	< 0.001	0.455	< 0.001	12.07	11.56	2.19	5	0	10
Grande chance	80	82	85	82.5	0.629	< 0.001	0.573	1.54	5.83	1.77	2	5	2.5
Bem provável	70	78	75	75	< 0.001	< 0.001	0.028	4.38	11.60	6.36	8	5	5
Provável	70	78	75	72.5	< 0.001	< 0.001	0.004	4.47	13.97	8.19	8	5	2.5
Acreditamos	75	80	80	80	0.299	0.002	0.380	3.21	6.15	4.67	5	5	5
Provavelmente	70	78	75	72.5	< 0.001	< 0.001	0.005	5.92	14.01	8.33	8	5	2.5
Mais que provável	60	70	60	60	< 0.001	< 0.001	0.859	9.12	9.88	6.05	10	0	0
Cerca de 50/50	50	50	50	50	0.956	0.056	0.334	1.35	1.62	1.64	0	0	0
Duvidamos	20	20	20	30	0.015	0.402	< 0.001	6.84	9.33	10.13	0	0	10
Improvável	10	20	20	30	< 0.001	< 0.001	< 0.001	8.41	7.67	11.00	10	10	20
Pouco provável	20	20	20	20	0.397	0.550	0.351	8.62	13.05	8.04	0	0	0
Provavelmente não	25	20	25	30	0.146	0.379	0.021	8.04	4.92	6.83	5	0	5
Pouca chance	10	20	20	20	< 0.001	< 0.001	< 0.001	3.73	6.45	7.43	10	10	10
Quase nenhuma chance	2	5	5	5	< 0.001	< 0.001	< 0.001	2.14	3.29	13.06	3	3	3
Muito improvável	5	10	10	10	< 0.001	< 0.001	< 0.001	6.09	3.97	13.59	5	5	5
Chances remotas	10	20	20	20	< 0.001	< 0.001	< 0.001	11.88	3.07	6.80	10	10	10

An analysis of GPT-5.1 reveals a degradation in calibration at the high-certainty spectrum when operating in Portuguese. For the Portuguese translation of "almost certain," human respondents anchored their estimates at a median of 95%. In contrast, GPT-5.1 dropped to a median of 85%, resulting in an AMD of 10.0 and a highly significant Brunner-Munzel p-value ($p < 0.001$). A similar significant divergence occurred for "highly likely" ($p < 0.001$). This indicates a cross-lingual vulnerability; while GPT-5.1 hesitated in English, its reluctance to assign high probabilistic weight is amplified in Portuguese. Conversely, Gemini 3 Flash demonstrates exceptional cross-lingual robustness and stability, acting as the most reliable proxy for human probabilistic intuition in this language. For high-probability expressions like "almost certain" and "highly likely," Gemini perfectly matched the human medians of 95% and 90%, respectively. The Brunner-Munzel tests yielded non-significant p-values, paired with low KL divergence scores. While Gemini still exhibits minor tail-smoothing at the absolute lower end—shifting the median of "improbable" from the human 15% to 20% ($p < 0.001$)—it avoids the severe central compression seen in other models.

The performance of Sabiá 4 presents a counter-intuitive finding. Despite being explicitly fine-tuned for Brazilian Portuguese, Sabiá 4 exhibited the most aggressive tail-

smoothing. It failed to capture the extreme distributions of human intuition, assigning a median of 90% to "almost certain" and compressing low-probability terms. For instance, Sabiá 4 assigned a median of 30% to the Portuguese equivalent of "we doubt". The Brunner-Munzel tests confirmed these misalignments as significant, indicating that localized linguistic training did not translate into accurate probabilistic calibration for this model. The analysis underscores that linguistic localization does not automatically guarantee adherence to established English baselines for estimative uncertainty.

The contrast between Gemini 3 Flash and Sabiá 4 highlights the complex nature of uncertainty representation in LLMs, raising the question of whether compressed distributions reflect calibration failures or localized linguistic nuances. Regardless, for high-stakes applications in Brazilian Portuguese, these results expose the practical risks of assuming semantic parity across cultures and commercial LLM outputs.

4.3. Mutilingual Comparison

A direct comparative analysis of model performance across the language barrier reveals distinct sensitivities to linguistic framing, as presented in Table 3. By isolating the language variable and comparing the English and Portuguese output distributions directly, we can assess whether these models possess a language-agnostic internal representation of uncertainty, or if their probabilistic calibrations are tightly bound to the specific language of the prompt. The data supports the latter, demonstrating that translation induces measurable shifts in probability estimation.

Tabela 3. English vs. Portuguese comparison in responses from the *same* model (GPT-5.1 and Gemini 3 Flash). p_{BM}: Brunner–Munzel test between EN and PT distributions; KL: $\mathrm{KL}(P_{\mathrm{EN}} \| P_{\mathrm{PT}})$ (`kl_en_vs_pt` in the data); AMD: absolute EN–PT median difference (percentage points).

Expression	$\mathrm{KL}^{\mathrm{GPT}}$	$\mathrm{KL}^{\mathrm{Gem}}$	$p_{\mathrm{BM}}^{\mathrm{GPT}}$	$p_{\mathrm{BM}}^{\mathrm{Gem}}$	$\mathrm{AMD}^{\mathrm{GPT}}$	$\mathrm{AMD}^{\mathrm{Gem}}$
Almost certain	2.36	0.04	0.002	0.337	5	0
Highly likely	1.73	0.10	< 0.001	0.062	5	5
Very good chance	1.85	0.70	0.001	0.001	2	0
Probable	2.61	1.77	< 0.001	0.055	8	0
Likely	4.21	0.98	< 0.001	0.025	8	0
We believe	2.88	0.65	< 0.001	0.036	10	0
Probably	2.87	1.72	< 0.001	0.188	8	0
Better than even	1.88	1.32	< 0.001	0.443	10	0
About even	3.42	0.40	0.937	1.000	0	0
We doubt	3.16	10.02	0.055	0.034	5	5
Improbable	2.14	1.50	0.454	< 0.001	0	0
Unlikely	4.01	14.43	0.715	0.002	0	5
Probably not	3.68	1.81	< 0.001	0.004	10	0
Little chance	1.89	3.05	0.039	< 0.001	2	10
Almost no chance	0.92	0.61	0.431	0.029	0	0
Highly unlikely	1.47	0.91	0.044	0.283	0	0
Chances are slight	1.37	2.31	0.348	< 0.001	0	5

GPT-5.1 exhibits significant cross-lingual instability, demonstrating statistically significant shifts in its probability distributions when processing Portuguese compared to English. For instance, in the high-certainty spectrum, the transition from "almost certain" to its Portuguese equivalent resulted in an Absolute Median Difference (AMD) of 5.0. More importantly, the Brunner-Munzel test confirms this shift is not mere noise.

A similar degradation occurs with "highly likely," which also displays an AMD of 5.0 and a significant p-value between its English and Portuguese inferences. This indicates that GPT-5.1 does not map these expressions symmetrically; its probabilistic confidence inherently drops when the semantic environment switches to Portuguese, exposing a vulnerability in its multilingual alignment.

In contrast, Gemini 3 Flash demonstrates remarkable cross-lingual stability, acting as a robust multilingual reasoner. For the vast majority of WEPs, Gemini's English and Portuguese distributions are statistically indistinguishable. Looking at "almost certain," Gemini maintained a perfect cross-lingual median alignment, with a Brunner-Munzel p-value of 0.336, indicating no significant difference between how it handles the concept in either language. Mid-range terms like "about even" were also perfectly mirrored. While minor discrepancies appeared in the lower tail, Gemini largely preserved the semantic equivalence of estimative uncertainty across the linguistic divide.

These comparisons are crucial. They prove that simply translating a prompt does not guarantee equivalent cognitive or statistical processing by an LLM. The pronounced linguistic effects observed in GPT-5.1 suggest that its English-centric training data dictates its probabilistic confidence, which fails to cleanly project onto Brazilian Portuguese. Conversely, Gemini's performance provides empirical evidence that it is feasible to build multimodal systems capable of maintaining stable and language-agnostic representations of complex pragmatic concepts like uncertainty.

For practitioners and researchers, particularly those deploying LLMs for Brazilian Portuguese, these findings carry practical implications. It becomes evident that prompt engineering and probability calibration are not perfectly transferable across languages. Developers relying on generalized models like GPT-5.1 cannot assume that an English prompt carefully tuned for a specific probability threshold will yield the same probabilistic reasoning when localized. This cross-lingual volatility underscores the necessity of language-specific validation phases, especially in high-stakes applications like automated decision-making.

5. Conclusions

This study demonstrates that while modern LLMs exhibit high linguistic fluency, their internal representations of estimative uncertainty remain significantly misaligned with human cognitive intuition. Our comparative analysis reveals a persistent "tail smoothing" phenomenon, where models systematically avoid extreme probabilistic values. Notably, while Gemini 3 Flash showed cross-lingual stability, the region-specific Sabiá 4 exhibited aggressive distribution compression. This suggests that localized fine-tuning may capture distinct cultural pragmatics rather than simply failing to align with English-centric probabilistic baselines.

Future research should prioritize the development of a Brazilian Portuguese human baseline to ensure that evaluation benchmarks are more culturally grounded than current English-centric models. Furthermore, it is essential to explore how alternative prompting strategies and calibration techniques might mitigate the cross-lingual volatility. Extending this methodology to a broader range of markers and diverse linguistic contexts will be crucial for developing language-agnostic and reliable AI systems.

Limitations

Our study presents important limitations regarding its experimental design and evaluation framework. Primarily, our baseline for human cognitive interpretation relies on empirical data gathered in an English survey [Fagen-Ulmschneider 2019]. We recognize that utilizing English data as a baseline for Portuguese inferences is a theoretical limitation, as it does not account for the native pragmatic interpretation of Brazilian speakers. We adopt this approach to measure the cross-lingual stability of the models rather than making absolute claims about human-model alignment in Brazil. Without a localized human baseline, it remains theoretically challenging to definitively disentangle whether the distributional shifts observed (such as Sabiá 4's aggressive tail-smoothing) represent a mathematical failure in model calibration or an accurate capture of culturally distinct Brazilian Portuguese pragmatics. Conducting a primary human survey in Portuguese that matches the rigorous quality standards and demographic scale of the existing baseline involves prohibitive operational costs at this stage. Consequently, developing a native Brazilian Portuguese human baseline remains an urgent priority for future work to ensure that evaluation benchmarks are truly culturally grounded rather than English-centric.

Ethics Statement

From an ethical perspective, this research highlights vulnerabilities in human-AI interaction, particularly concerning the equitable and accurate deployment of LLMs for non-English speakers. The vast majority of studies on estimative uncertainty remain heavily skewed toward English, leaving widely spoken languages like Brazilian Portuguese largely underexplored. This gap poses a significant challenge, as cultural and linguistic variations can significantly alter the interpretation of probabilistic language. As LLM-based systems increasingly mediate critical decision-making processes, ensuring the safe and reliable communication of uncertainty is vital. Our research emphasizes that overly confident language generated by models can easily lead to unwarranted user reliance on inaccurate outputs.

Generative AI Usage in Paper Preparation

The use of Generative AI tools in the preparation of this manuscript was strictly limited to language editing and refining the academic writing.

Referências

Erev, I. and Cohen, B. L. (1990). Verbal versus numerical probabilities: efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes*, 45:1–18.

Fagen-Ulmschneider, W. (2019). Perception of probability words. University of Illinois at Urbana-Champaign.

Freitag, R. M. K., Cardoso, P. B., and Tejada, J. (2022). Linguistic and paralinguistic constraints on the function of (eu) acho que as discourse markers in brazilian portuguese: A multilevel approach. *Pragmatics & Cognition*, 29(2):324–346.

Google DeepMind (2026). Gemini 3 flash. Technical report. Multimodal large language model developed by Google DeepMind.

Juanchich, M. and Sirota, M. (2020). Do people really prefer verbal probabilities? *Psychological Research*, 84:2325–2338.

Laitz, T., Almeida, T. S., Abonizio, H., Malaquias Junior, R., Bonás, G. K., Piau, M., Larcher, C., Pires, R., and Nogueira, R. (2026). Sabiá-4 technical report. *arXiv preprint arXiv:2603.10213*.

OpenAI (2026). Gpt-5.1. Technical report. Large language model developed by OpenAI.

Schneider, S. (2016). Communicating uncertainty: A challenge for science communication. In Drake, J., Kontar, Y., Eichelberger, J., Rupp, T., and Taylor, K., editors, *Communicating Climate-Change and Natural Hazard Risk and Cultivating Resilience*, volume 45 of *Advances in Natural and Technological Hazards Research*. Springer, Cham.

Schuck, A. d. F., Garcia, G. L., Manesco, J. R. R., Paiola, P. H., et al. (2025). Evaluating large language models for brazilian portuguese sentiment analysis: A comparative study of multilingual state-of-the-art vs. brazilian portuguese fine-tuned llms. *Journal of the Brazilian Computer Society*, 31(1):885–917.

Shorinwa, O., Mei, Z., Lidard, J., Ren, A. Z., and Majumdar, A. (2025). A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3).

Tang, Z., Shen, K., and Kejriwal, M. (2026). An evaluation of estimative uncertainty in large language models. *npj Complex*, 3:8.

Vlasyan, G. R. et al. (2018). Linguistic hedging in the light of politeness theory. In *European Proceedings of Social and Behavioural Sciences*. Future Academy.

Yona, G., Aharoni, R., and Geva, M. (2024). Can large language models faithfully express their intrinsic uncertainty in words? In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764, Miami, Florida, USA. Association for Computational Linguistics.