Speech Disfluencies and LLM Confidence: Length Bias and Pragmatic Insensitivity in Brazilian Portuguese

Valéria Vieira dos Santos
Federal University of São Carlos (UFSCar)
Postgraduate Program in Linguistics
São Carlos, Brazil
valeriavsantos93@gmail.com

Abstract

Training Large Language Models (LLMs) relies predominantly on written, curated corpora, which may limit their reliability on spontaneous speech. Oral language exhibits real-time planning markers — filled pauses, repetitions, false starts, and vowel lengthenings — that modulate epistemic commitment. This pilot study investigates how such disfluencies affect the alignment between LLM confidence and a discourse-pragmatic uncertainty proxy in a Portuguese model (Llama-3.1-8B-Instruct). Using a benchmark of 344 turns from the *Roda Viva* corpus, we contrast faithful Conversation Analysis transcriptions with sanitized versions and combine binned divergence metrics (ECE, OE) with rank correlation and multivariate regression analyses. We find that model confidence is overwhelmingly driven by a surface feature — turn length ($\beta_{\text{std}} = +14.47$, $p < 0.001$) — rather than by pragmatic markers of uncertainty ($\beta_{\text{oral}} = -3.09$, $\beta_{\text{hedges}} = -0.97$, both non-significant; $R^2 = 0.29$). After controlling for length, residual effects of disfluency markers align in the human-expected direction but are dwarfed by length bias. We argue that this surface-feature dominance subsumes the *pragmatic blindness* phenomenon and explains the substantial divergence observed via ECE (41.95) and OE (4.29) between faithful and sanitized conditions.

1 Introduction

Recent applications of Large Language Models (LLMs) to oral language processing face a reliability challenge: the dissociation between the statistical probability assigned by the model and the actual epistemic commitment expressed in human speech. This dissociation can result in critical calibration errors and systematic hallucinations.

Modern neural networks tend toward probabilistic overconfidence, assigning high probability to incorrect answers (Guo et al., 2017). In linguistic contexts, this may manifest when a model expresses categorical certainty about content that does not adequately reflect the speaker's actual degree of commitment (Mielke et al., 2022).

This problem may be amplified when models trained predominantly on written, standardized corpora (Bender et al., 2021) are applied to spontaneous speech. Unlike edited writing, orality involves real-time production processes that generate filled pauses, reformulations, and vowel lengthenings (Biber, 1988; Shriberg, 1999). In interactional linguistics, such markers are not performance failures — they function as pragmatic signals modulating epistemic commitment, constituting what Marcuschi (2003) calls the *planning syntax of speech*.

A natural hypothesis is therefore that LLMs exhibit systematic limitations in interpreting these signals — a phenomenon we term **pragmatic blindness**. Evaluating this hypothesis, however, requires disentangling two competing explanations for model confidence on oral input: *(a)* sensitivity to pragmatic content (the marker-driven account), and *(b)* reliance on surface properties of the input such as turn length (the surface-feature account). We conduct a pilot experiment that explicitly tests both, comparing faithful Conversation Analysis transcriptions (Jefferson, 1983) with sanitized journalistic versions of the same speech from the *Roda Viva* corpus (Vale et al., 2024).

The study contributes *(i)* a quantitative scale of epistemic commitment based on pragmatic markers of orality; *(ii)* a divergence audit via ECE and OE; and *(iii)* a multivariate decomposition identifying turn length as the dominant predictor of confidence. We show that pragmatic blindness is best understood as a manifestation of broader **surface-feature dominance**: LLMs prioritize length over pragmatic content.[1]

[1] Code: https://github.com/ValeriaVSantos/uncertainty-signature-audit

Proceedings of the 2nd Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2026), pages 24–28

2 Background and Evaluation Metrics

To evaluate the impact of orality on LLM reliability, this study combines the mapping of interactional speech phenomena with the analysis of probabilistic deviations in language models. This relationship is operationalized through a linguistic proxy of epistemic commitment — a numerical scale that allows comparison between human discursive signals and the probabilistic confidence estimated by models. The metric does not assess the factual truthfulness of statements, but the degree of epistemic commitment linguistically signaled by the speaker. Its construction draws on two foundations: interactional metadiscourse, in which lexical hedges indicate modulation of propositional commitment (Hyland, 2005), and the planning syntax of speech, in which hesitation markers reflect real-time cognitive formulation processes (Marcuschi, 2003).

2.1 From Calibration to Pragmatic Alignment

In classical Machine Learning, calibration measures the correspondence between a model's predicted probability and the *empirical frequency of correct answers* (Guo et al., 2017). In this study, the reference quantity is not factual correctness but a *discourse-pragmatic proxy of epistemic commitment* derived from linguistic annotation (Section 3). We therefore repurpose Expected Calibration Error (ECE) and Overconfidence Error (OE) as *divergence measures* between model confidence and this pragmatic proxy, rather than as a probabilistic calibration audit in the strict sense. Under this reading, ECE and OE quantify the systematic gap between the epistemic certainty linguistically signaled by the speaker and the confidence assigned by the model: lower values indicate stronger alignment with marked epistemic caution, while persistent positive deviations expose the model's insensitivity to pragmatic uncertainty cues.

We formulate the task as Natural Language Inference (NLI) and apply the Softmax function to the target token logit ("YES"), obtaining the probability associated with the prediction. Predictions are grouped into confidence bins, where the mean algorithmic confidence is contrasted with the pragmatic proxy. We complement ECE and OE with Spearman rank correlation, which measures monotonic alignment without binning artifacts, and with a multivariate linear regression that decomposes model confidence into pragmatic-marker effects and a surface-feature control (turn length), enabling us to test whether observed misalignment reflects insensitivity to pragmatic content or reliance on superficial input properties.

3 Methodology

This study adopts a quantitative experimental design structured in three stages: *(i)* curation and annotation of a contrastive parallel corpus; *(ii)* extraction of model confidence via logit-level softmax; and *(iii)* a divergence audit complemented by multivariate decomposition of confidence into pragmatic and surface-feature components.

The dataset derives from three *Roda Viva* interviews (Heloísa Starling, Marco Aurélio Mello, and Galvão Bueno), segmented into 344 turns of approximately 40 seconds, covering different domains and speaker profiles.

Each turn was structured as a contrastive pair: **Layer A** (faithful version) preserves disfluency markers following Jeffersonian conventions (Jefferson, 1983) — micropauses `(.)`, lengthenings `(::)`, truncations, and filled pauses; **Layer B** (sanitized version) suppresses these markers, approximating written standard norms.

The epistemic commitment proxy was built using a deductive heuristic protocol applied to Layer A, with a maximum score of 100 subject to hierarchical deductions based on pragmatic impact, as shown in Table 1.

3.1 Logit Extraction and Algorithmic Confidence

We used the Meta-Llama-3.1-8B-Instruct model (Dubey et al., 2024), loaded with `bitsandbytes` 4-bit NF4 quantization and double quantization (`bfloat16` compute dtype). The model was prompted with a binary instruction: *"Based strictly on the text, does the speaker express absolute conviction about the information? Answer only YES or NO."* In Layer A, Jeffersonian markers were preserved in the input. Model confidence was obtained by applying softmax over the full vocabulary logits at the final position of the prompt and reading the probability mass on the YES token; no sampling or temperature scaling was applied. The procedure was run independently on Layer A and Layer B inputs for each of the 344 turns.

3.2 Divergence and Statistical Analysis

Confidences were grouped into M=10 bins for ECE and OE computation, and the Wilcoxon

Category	Marker	Example	Pts	Theoretical Basis
Epistemic Hedges	Lexical hedges	*maybe, I think, seems*	−15	Epistemic retreat (Hyland, 2005)
Reformulations	False starts	*has to end... rewrite*	−10	Syntactic abandonment (Marcuschi, 2003)
Filled Pauses	Hesitation voc.	*uh..., um...*	−5	Macrostructural planning marker
Lengthenings	Vowel prolongation	*veryyy, forrrr*	−5	Lexical selection in progress
Repetitions	Term repetition	*that that, but but*	−5	Rhythmic hesitation

Table 1: Epistemic Commitment Annotation Matrix.

signed-rank test (Dror et al., 2018) was used to compare conditions. We further quantified, per turn, the frequency of each marker category (hedges, filled pauses, lengthenings, repetitions, false starts) and the turn length in words, computing Spearman rank correlations between these predictors and model confidence. To disentangle pragmatic from surface effects, we fitted an ordinary least squares regression of model confidence on three predictors — turn length, total oral disfluency markers, and lexical hedges — reporting standardized coefficients to allow magnitude comparison.

4 Results and Discussion

We report three nested levels of analysis: global divergence (§4.1), bivariate correlations and their length confound (§4.2), and a multivariate decomposition that identifies the dominant predictor of model confidence (§4.3).

4.1 Confidence–Proxy Divergence

Table 2 shows global ECE and OE for both layers. Divergence is large in both conditions (ECE > 40), with the faithful layer slightly worse. The Wilcoxon signed-rank test confirms a statistically significant difference between conditions (W=10988.50, p=0.0023). Figure 1 visualizes the gap, showing systematic deviations from the diagonal across mid-confidence bins.

Input Condition	ECE	OE
Layer A (Faithful)	41.95	4.29
Layer B (Sanitized)	41.14	3.31

Table 2: Divergence between model confidence and the discourse-pragmatic proxy under both transcription layers (N=344). Lower values indicate stronger alignment.

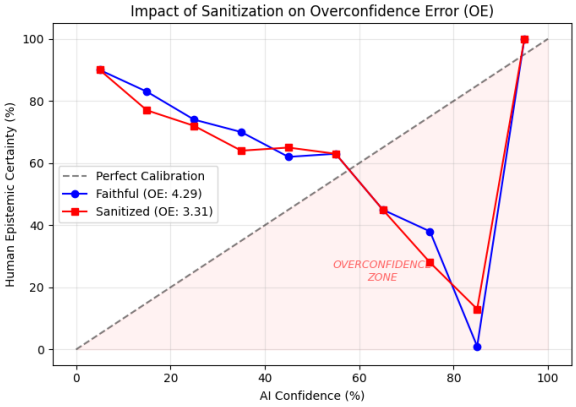

Figure 1: Reliability Diagram for Layer A (Faithful) and Layer B (Sanitized). The dashed diagonal represents perfect alignment with the proxy. The upper bin (80–100%) is sparsely populated and should be interpreted with caution.

4.2 Length Confounds Bivariate Correlations

The full proxy correlates negatively with model confidence in both conditions ($\rho_{\text{Faithful}} = -0.49$, $\rho_{\text{Sanitized}} = -0.43$, $p < 0.001$). Read directly, this might suggest a paradoxical positive association between hesitation cues and model confidence. Two factors complicate this reading. First, the proxy is constructed by subtracting points for marker presence, so turns with many markers are by definition assigned lower scores. Second, turns with more markers are systematically longer, and model confidence shows a strong positive association with raw turn length ($\rho_{\text{words}} = +0.58$, $p < 0.001$). The bivariate result therefore conflates pragmatic and length effects. Figure 2 displays the bivariate Spearman correlations together with the confidence distribution shift between conditions.

4.3 Surface-Feature Dominance

To disentangle pragmatic content from input length, we regressed model confidence (Layer A) on three

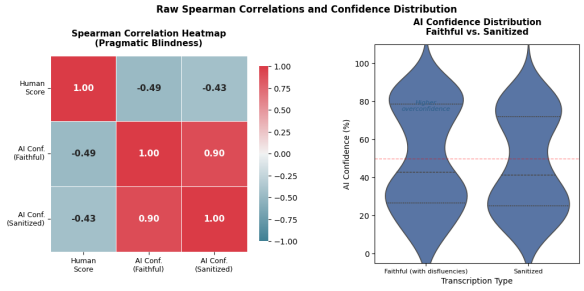

Figure 2: Statistical alignment analysis. Left: Spearman correlation heatmap between the pragmatic proxy and model confidence under both conditions (raw correlations; see §4.2 for the length-controlled analysis). Right: confidence distribution shift between faithful and sanitized conditions.

predictors: turn length, total oral disfluency markers, and lexical hedges. Table 3 reports the standardized coefficients.

Predictor	β_{std}	**p**
Turn length (`n_words`)	$+14.47$	< 0.001
Oral disfluency markers	-3.09	0.115
Lexical hedges	-0.97	0.434

Table 3: OLS regression of model confidence (Layer A, N=344) on surface and pragmatic predictors. standardized coefficients; R^2=0.285, $F(3,340)$=45.23, $p < 0.001$.

Turn length is the only significant predictor and dwarfs both pragmatic categories: a one-standard-deviation increase in length raises predicted confidence by $\sim$14.5 points, while marker-based predictors are an order of magnitude smaller and not statistically distinguishable from zero. Partial Spearman correlations controlling for length recover small but human-aligned effects ($\rho_{\text{oral|length}} = -0.21$, $\rho_{\text{hedges|length}} = -0.15$, both $p < 0.01$): once length is accounted for, the model does decrease confidence in the presence of hesitation markers, but the effect is too weak to surface in the unconditional analysis.

This finding reframes the pragmatic blindness phenomenon hypothesis. Disfluency markers are not invisible to the model — they exert a small effect in the human-expected direction. They are, however, overwhelmed by a surface predictor with no semantic motivation, a phenomenon we term **surface-feature dominance**. Table 4 illustrates the resulting regime: short, marker-laden turns receive high confidence (ID 190), and confidence remains high even where the speaker explicitly retracts (ID 8). Structural completeness of the turn outweighs the epistemic-caution signals it contains.

ID	**Layer A (excerpt)**	**Score**	**AI Conf.**	**Gap**
190	*"uh uh uh I kind of thought it was my moment but then I said wait..."*	0	80.85%	+80.85
133	*"that wasn't because of Brazil's elimination (.) went crazy...it was something that had never happened..."*	25	69.53%	+44.53
8	*"first of all I did not criticize colleague Sérgio Moro (.) not not in any way so- it was misunderstood..."*	45	72.26%	+27.26

Table 4: Excerpts from turns illustrating the confidence–proxy gap. Excerpts are abbreviated for space; full turns contained additional disfluency markers and hedges contributing to the displayed scores. Gap = AI Conf. − Score; positive values indicate overconfidence relative to the proxy.

5 Conclusion

This pilot study examined how disfluency markers shape Llama-3.1-8B confidence on faithful versus sanitized transcriptions from the *Roda Viva* corpus. Multivariate decomposition shows that model confidence on spontaneous speech is dominated by a surface feature — turn length — rather than by pragmatic content: length alone accounts for an order-of-magnitude larger effect than all disfluency markers combined. After controlling for length, residual marker effects are small and in the human-expected direction, indicating that pragmatic blindness is best understood as a manifestation of broader **surface-feature dominance** rather than categorical insensitivity to pragmatic content.

The implications are practical: applications routing spontaneous speech through LLMs risk inheriting confidence estimates driven by superficial input properties. Promising mitigations include explicit semantic annotation of pragmatic markers, length-normalized calibration, and discourse-aware benchmarks that decouple structural and pragmatic predictors — productive directions for linguistically grounded LLM evaluation in under-resourced varieties such as Brazilian Portuguese.

Limitations

This study has inherent limitations given its exploratory scope. The dataset size (344 turns) and focus on a single discursive domain (journalistic interview) restrict immediate generalization to other sociolinguistic varieties or speech genres. Experiments were conducted on a single autoregressive model (Llama-3.1-8B-Instruct), and whether surface-feature dominance generalizes across architectures, model sizes, tokenization schemes, and Portuguese-specific models such as the Sabiá family remains an open question. The epistemic commitment proxy was developed deductively from linguistic theory and annotated by a single researcher; while the marker-type ablation partially validates its construction, multi-annotator human ratings of epistemic commitment would strengthen the reference signal. Our regression model is restricted to three predictors and assumes linear additive effects; alternative surface features (e.g., vocabulary complexity, syntactic depth, named-entity density), non-linear interactions, and a parallel analysis on the sanitized condition (Layer B) remain to be examined. Finally, the regression accounts for $\sim$29% of confidence variance, leaving room for additional surface and content predictors not captured here.

Ethical Considerations

This research aligns with Responsible AI principles, aiming to mitigate overconfidence hallucinations in language technologies. All data derive from journalistic interviews of public figures broadcast on open television (*Roda Viva* corpus), constituting public domain material. The study involves no sensitive data or personally identifiable information.

Acknowledgments

The author thanks the three anonymous reviewers of CODI-CRAC 2026, whose feedback motivated the extended multivariate analysis presented in Section 4.3. Generative AI tools (Claude) were used for assistance with English translation and language revision of this manuscript; all methodology, analysis, and conclusions are solely the author's own.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, New York, NY. Association for Computing Machinery.

Douglas Biber. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge, UK.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1383–1392.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and 1 others. 2024. The Llama 3 herd of models. *Computing Research Repository*, arXiv:2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR.

Ken Hyland. 2005. *Metadiscourse: Exploring Interaction in Writing*. Continuum, London.

Gail Jefferson. 1983. Issues in the transcription of naturally occurring talk: Caricature versus capturing pronunciational particulars. *Tilburg Papers in Language and Literature*, 34:1–12.

Luiz Antônio Marcuschi. 2003. *Análise da Conversação*, 5 edition. Ática, São Paulo.

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.

Elizabeth Shriberg. 1999. To "errrr" is human: Ecology and acoustics of speech disfluencies. In *Proceedings of the International Congress of Phonetic Sciences*.

Oto Araújo Vale, Arnaldo Candido Junior, Amanda Rassi, and Jorge Baptista. 2024. Roda Viva corpus: An audiovisual Brazilian Portuguese interview corpus. In *Proceedings of the 16th International Conference on the Computational Processing of Portuguese*, pages 204–210.