IMMBA: Integrated Mixed Models with Bootstrap Analysis - A Statistical Framework for Robust LLM Evaluation

Vinícius Di Oliveira[1,2][a], Pedro Carvalho Brom[1,3][b] and Li Weigang[1][c]
[1]*TransLab, University of Brasilia, Brasilia, Federal District, Brazil*
[2]*Secretary of Economy, Brasilia, Federal District, Brazil*
[3]*Federal Institute of Brasilia, Brasilia, Federal District, Brazil*

Keywords: Large Language Models, Statistical Evaluation, Linear Mixed Models, Bootstrap Resampling, Variance Decomposition, Retrieval-Augmented Generation, LLM Evaluation.

Abstract: Large Language Models (LLMs) have advanced natural language processing across diverse applications, yet their evaluation remains methodologically limited. Standard metrics such as accuracy or BLEU offer aggregate performance snapshots but fail to capture the inherent variability of LLM outputs under prompt changes and decoding parameters like temperature and top-p. This limitation is particularly critical in high-stakes domains, such as legal, fiscal, or healthcare contexts, where output consistency and interpretability are essential. To address this gap, we propose **IMMBA**: *Integrated Mixed Models with Bootstrap Analysis*, a statistically principled framework for robust LLM evaluation. IMMBA combines Linear Mixed Models (LMMs) with bootstrap resampling to decompose output variability into fixed effects (e.g., retrieval method, decoding configuration) and random effects (e.g., prompt phrasing), while improving estimation reliability under relaxed distributional assumptions. We validate IMMBA in a Retrieval-Augmented Generation (RAG) scenario involving structured commodity classification under the Mercosur Common Nomenclature (NCM). Our results demonstrate that IMMBA isolates meaningful performance factors and detects significant interaction effects across configurations. By integrating hierarchical modelling and resampling-based inference, IMMBA offers a reproducible and scalable foundation for evaluating LLMs in sensitive, variance-prone settings.

1 INTRODUCTION

Large Language Models (LLMs) have transformed the landscape of natural language processing, enabling remarkable progress across knowledge-intensive tasks. Yet, systematic and statistically sound evaluation of these models remains an open challenge. Conventional metrics, such as accuracy, F1-score, or BLEU, often fail to capture the nuanced, probabilistic nature of LLM outputs, particularly when using varied prompts or stochastic decoding parameters, including temperature and nucleus sampling (top-p).

This evaluation gap is not merely theoretical. In high-stakes domains, legal decision-making, fiscal classification, or healthcare, LLM outputs must be both reliable and reproducible, yet current evaluation practices offer limited insight into the sources of output variability, model robustness, or susceptibility to hallucinations.

To address this, we propose a statistically principled framework for LLM evaluation that integrates bootstrap resampling with multivariate Linear Mixed Models (LMMs). This approach enables:

- Decomposition of total output variability into fixed effects (e.g., retrieval strategy, decoding parameters) and random effects (e.g., prompt phrasing), supporting interpretable performance analysis.
- Robust estimation of model behaviour under relaxed parametric assumptions, mitigating overconfidence in evaluation results.
- Identification of statistically significant interactions between LLM configurations and retrieval methods, informing model fine-tuning and deployment decisions.

[a] https://orcid.org/0000-0002-1295-5221
[b] https://orcid.org/0000-0002-1288-7695
[c] https://orcid.org/0000-0003-1826-1850
* E-mails of the corresponding authors.

Di Oliveira, V., Brom, P. C. and Weigang, L.
IMMBA: Integrated Mixed Models with Bootstrap Analysis - A Statistical Framework for Robust LLM Evaluation.
DOI: 10.5220/0013819400003985

In *Proceedings of the 21st International Conference on Web Information Systems and Technologies (WEBIST 2025)*, pages 92-102
ISBN: 978-989-758-772-6; ISSN: 2184-3252

We empirically validate our framework within a Retrieval-Augmented Generation (RAG) setting, applying it to structured classification tasks based on the Mercosur Common Nomenclature (NCM) system, where contextual rigour and terminological precision are essential (MERCOSUR, 2024). The NCM code is based on the Harmonised System, the commodity description code system developed by the World Customs Organisation - WCO (WCO, 2018). Results show that our methodology isolates method-specific performance gains from prompt-induced variance and experimental noise, with total variability partitioned into fixed effects ($\sigma_f^2 = 2.14$), prompt variability ($\sigma_P^2 = 0.66$), and residual error ($\sigma_e^2 = 6.42$).

The current landscape of LLM evaluation reflects a growing awareness of prompt variability, uncertainty, and the need for statistically grounded assessment methods. Yet, few existing approaches simultaneously incorporate prompt-level random effects, hierarchical model configurations, and bootstrap-based estimation within a unified statistical framework.

The proposed IMMBA framework addresses this methodological gap. By integrating Linear Mixed Models with bootstrap resampling, IMMBA enables principled decomposition of variance into fixed and random effects, offering reproducible and interpretable insights into LLM performance variability across prompts, decoding strategies, and model architectures. Unlike prior methods, IMMBA brings together hierarchical modelling and resampling inference in a single, scalable pipeline, providing a robust foundation for LLM evaluation, particularly in retrieval-augmented or high-stakes domains where output consistency and reliability are essential. The main components of the framework are illustrated in Figure 1, which depicts the sequential integration of data preparation, LMM modelling, and bootstrap estimation.

The remainder of this paper is organised as follows: Section 2 reviews existing LLM evaluation practices and identifies methodological gaps. Section 3 details the proposed framework. Section 4 outlines the experimental design. Section 5 presents the findings, Section 6 discusses the results and Section 7 concludes with directions for future research.

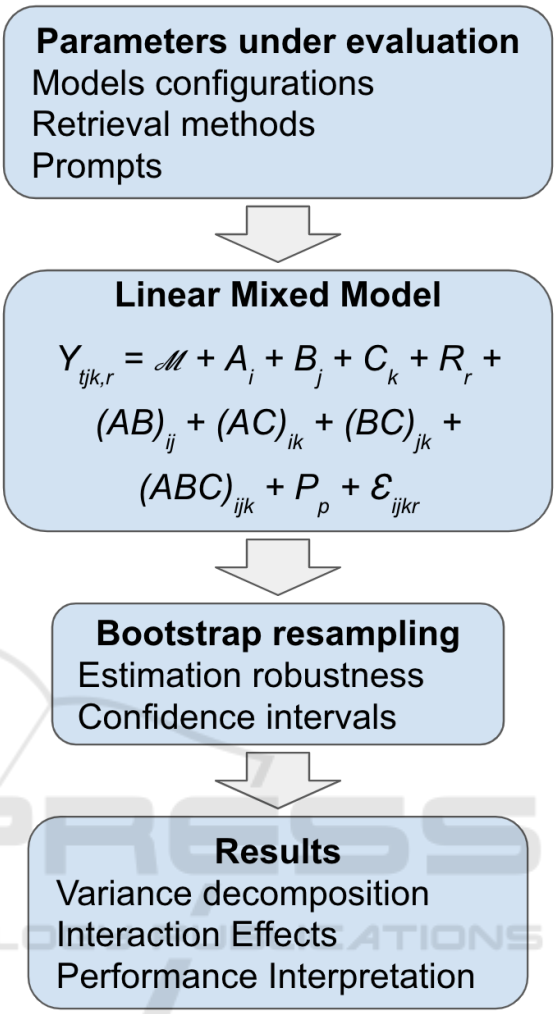

Figure 1: The IMMBA flowchart.

2 RELATED WORKS

Traditional approaches to evaluating large language models (LLMs) often rely on aggregate metrics such as accuracy, BLEU, or F1-score. While these measures provide general performance estimates, they fail to capture the stochasticity and prompt sensitivity inherent in LLM outputs (Yang et al., 2024). Recent work has shown that minor changes in prompt phrasing can cause significant variance in responses, highlighting a major limitation of single-score evaluations (Liu et al., 2024; Kapoor et al., 2024).

Growing interest in prompt-induced variability has led to methods that explicitly model uncertainty in LLM behaviour. Jiang et al. (Jiang et al., 2023) use prompt ensembles to calibrate epistemic uncertainty, while Tonolini et al. (Tonolini et al., 2024) employ Bayesian prompt selection to account for output variance. Other studies explore iterative prompting and instruction tuning to estimate model uncertainty under few-shot and in-context learning conditions (Abbasi Yadkori and Kuzborskij, 2024). These works underscore the inadequacy of static evaluation pipelines and the need for distribution-aware metrics.

To move beyond point estimates, several studies incorporate statistical resampling techniques. Zhou et

al. (Zhou et al., 2024) and Nikitin et al. (Nikitin et al., 2024) apply bootstrap resampling to estimate confidence intervals and assess semantic entropy in generation outputs. Bootstrap-enhanced conformal prediction has also been proposed for calibrating output confidence post hoc (Kapoor et al., 2024). These methods offer improved robustness, particularly in non-parametric or high-variance settings.

Despite their success in psychology and the social sciences, Linear Mixed Models (LMMs) remain underexplored in NLP evaluation. Recent literature hints at their potential: Liu et al. (Liu et al., 2024) model response variance induced by instruction tuning, suggesting latent hierarchical effects. However, systematic use of LMMs to partition fixed and random sources of variability—such as decoding parameters, prompt phrasing, and model type—remains rare. Prior work in cognitive science has successfully applied LMMs to interpret language and visual data (H. Wang and Yu, 2022; C.-H. Liu and Wang, 2023), and recent research has employed similar techniques to study cultural alignment in multilingual LLMs (J. Rystrøm and Hale, 2025). Nevertheless, LMMs have yet to be fully leveraged in mainstream LLM evaluation frameworks.

LLMQuoter (Bezerra and Weigang, 2025) evaluates its distillation-based model on quote extraction to enhance RAG. This is achieved using the DSPy framework, which leverages OpenAI GPT-4.0 as an LLM Judge to calculate redefined Precision, Recall, and F1-score for the relevance of extracted quotes. Furthermore, it assesses the Semantic Accuracy (Sacc) of answers generated by various base models by comparing performance when provided with either extracted 'gold quotes' or the full context. In contrast, another study (Weigang and Brom, 2025) evaluates the quality of various LLMs and traditional tools in Chinese-English-Chinese back-translation, forming an LLM-BT framework. This comprehensive evaluation uses a suite of metrics, including BLEU (with a novel Jieba-segmentation-based method), CHRF, TER, and Semantic Similarity (SS), substantiated by statistical analyses such as the multi-sample Friedman test and Dunn post-hoc test to ascertain significant performance differences across models and text types.

3 METHODOLOGY

This section presents the proposed statistical framework for Large Language Model (LLM) evaluation, integrating bootstrap resampling with multivariate Linear Mixed Models (LMMs). The objective is to decompose the sources of performance variability in LLM outputs, accounting for both fixed experimental factors and random effects such as prompt phrasing.

3.1 Linear Mixed Model Formulation

To quantify performance variability, we adopt a multivariate Linear Mixed Model (LMM) structured as follows:

$$Y_{ijk,pr} = \mu + A_i + B_j + C_k + R_r + (AB)_{ij} + (AC)_{ik} + (BC)_{jk} + (ABC)_{ijk} + P_p + \varepsilon_{ijkr} \quad (1)$$

where:

- $Y_{ijk,pr}$ represents the observed evaluation score for a given configuration defined by model A_i, temperature B_j, top-p parameter C_k, retrieval method R_r, and prompt P_p.
- μ is the overall intercept.
- A_i, B_j, C_k, and R_r denote the fixed effects associated with the model, temperature, top-p, and retrieval method, respectively.
- $(AB)_{ij}$, $(AC)_{ik}$, $(BC)_{jk}$, and $(ABC)_{ijk}$ represent interaction terms between fixed effects.
- P_p is a random effect capturing variability attributable to prompt phrasing, modelled as $P_p \sim \mathcal{N}(0, \sigma_P^2)$.
- ε_{ijkr} is the residual error term, assumed to follow $\mathcal{N}(0, \sigma_e^2)$.

This formulation allows systematic decomposition of the total observed variance:

$$\mathrm{Var}(Y_{ijk,pr}) = \sigma_f^2 + \sigma_P^2 + \sigma_e^2, \quad (2)$$

where:

- σ_f^2 represents variance explained by fixed effects and their interactions.
- σ_P^2 captures variability induced by prompt phrasing.
- σ_e^2 accounts for residual, unexplained noise.

This variance decomposition enables interpretable attribution of performance fluctuations to experimental configurations and linguistic randomness, addressing a critical shortcoming of conventional aggregate metrics.

3.2 Bootstrap Resampling for Robust Estimation

To mitigate reliance on strict parametric assumptions and improve estimator robustness, we integrate non-parametric bootstrap resampling into the LMM fitting process. The bootstrap procedure consists of the following steps:

1. Generate B bootstrap samples by resampling, with replacement, from the observed dataset.
2. Fit the LMM to each bootstrap sample, obtaining distributions of parameter estimates and variance components.
3. Construct empirical confidence intervals and standard errors based on the bootstrap distributions.

This approach accommodates potential deviations from normality or homoscedasticity in LLM output distributions, providing more reliable statistical inference. In our experiments, we employ $B = 1000$ bootstrap iterations, ensuring stability of estimates without prohibitive computational cost.

3.3 Evaluation Metrics

Model outputs were assessed along four dimensions, each rated on a 0–10 ordinal scale by trained human evaluators following a standardised scoring protocol. *Quality* reflects the clarity, conciseness, and structural coherence of the response. *Agreement* captures semantic alignment with an expert-validated baseline. *Accuracy* measures factual correctness against the gold-standard labels. *Hallucination* quantifies the degree of unsupported or fabricated content, with lower scores indicating greater reliability. To ensure consistency, multiple raters independently evaluated outputs. Divergences were resolved through adjudication, and inter-rater reliability was measured via Spearman's correlation, which confirmed strong agreement across evaluators.

The LMM and bootstrap analyses are applied independently to each metric, enabling granular decomposition of performance variability across evaluation dimensions. This procedure is illustrated in Figure 2.

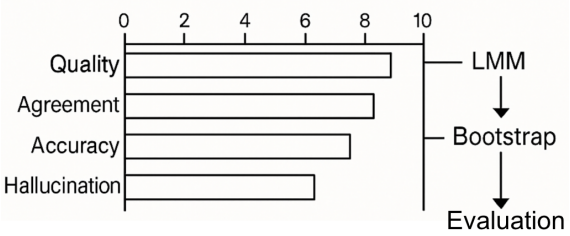

Figure 2: Evaluation metrics flow to the bootstrap.

3.4 Rationale for Statistical Design

The proposed methodology provides several advantages over conventional LLM evaluation approaches:

- It accounts for hierarchical sources of variability, isolating prompt-level randomness from systematic configuration effects.
- It identifies statistically significant interactions between model parameters (e.g., retrieval method, temperature, top-p), supporting informed model tuning.
- The bootstrap-enhanced framework provides robust parameter estimation and confidence intervals, reducing susceptibility to artefacts driven by non-Gaussian output distributions.

In high-stakes applications requiring both precision and interpretability, this methodology offers a statistically rigorous foundation for assessing LLM reliability and guiding deployment decisions.

4 EXPERIMENTAL DESIGN

This section details the experimental setup adopted to validate the proposed statistical framework for LLM evaluation. The design ensures systematic control of experimental factors, randomisation of prompts, and robust statistical power for hypothesis testing.

The dataset used for the experimental design is the Eleven Dataset (Di Oliveira et al., 2022), a database that presents descriptions of goods according to the Common Nomenclature of Mercosur captured from Brazilian Electronic Invoices.

The detailed calculation and all the codes used are available on the project GitHub [1].

4.1 Research Hypotheses

The following hypotheses were tested:

- **H1:** Fixed effects (model architecture, retrieval method, decoding parameters) significantly influence LLM evaluation scores.
- **H2:** Prompt phrasing contributes measurable, random variability to LLM outputs.
- **H3:** Bootstrap resampling improves the stability and reliability of parameter estimates within the LMM framework.

4.2 Factorial Design and Control Variables

We employ a full-factorial design to assess the influence of model configurations and retrieval methods on LLM performance. The design comprises:

- **Models** (A_i): Five distinct LLMs, encompassing open-source and proprietary architectures of varying capacities (gpt-4o-mini, deepseek-chat,

[1] https://github.com/pcbrom/immba

TeenyTinyLlama, gemini-2.0-flash, and Mistral-7B).

- **Temperature** (B_j): Three decoding temperatures (0.1, 1.0, 1.9) to control output randomness.
- **Top-p Sampling** (C_k): Three nucleus sampling thresholds (0.1, 0.5, 0.9).
- **Retrieval Method** (R_r): Comparison between conventional semantic retrieval and a retrieval-augmented strategy incorporating metadata filtering.

The resulting $3 \times 3 \times 6 \times 2 = 108$ experimental conditions are evaluated independently, ensuring comprehensive coverage of the configuration space.

4.3 Prompt Selection and Random Effects

To model linguistic variability, prompts are treated as random effects within the statistical analysis. We utilise a curated set of prompts derived from commodity classification tasks under the Mercosur Common Nomenclature (NCM) system. While the task provides a realistic, high-precision use case, the evaluation framework is agnostic to domain-specific taxonomies.

Prompt phrasing variability is explicitly captured in the Linear Mixed Model, isolating prompt-induced fluctuations from systematic effects of model parameters or retrieval methods.

4.4 Sample Size Determination and Power Analysis

Sample size determination follows established statistical guidelines to ensure sufficient power for detecting main effects and interactions within the LMM. Using Cohen's effect size $f = 0.25$ (medium effect), significance level $\alpha = 0.05$, and power $1 - \beta = 0.8$, a minimum of 196 replicates per experimental condition is required.

This yields a total of $196 \times 108 = 21,168$ evaluated responses, providing robust statistical precision for parameter estimation and variance decomposition.

4.5 Evaluation Metrics and Scoring Protocol

LLM outputs are assessed across four dimensions, using a 0 to 10 ordinal scale:

- **Quality:** Clarity, structural coherence, and conciseness.
- **Agreement:** Semantic alignment with a validated expert baseline.
- **Accuracy:** Factual correctness of the information presented.
- **Hallucination:** Degree of fabrication or unverifiable content (lower scores preferred).

Human raters conduct the evaluations using a standardised scoring protocol, with responses expressing uncertainty or deferring judgement penalised, under established evaluation prompts. The scoring format facilitates quantitative analysis, supporting application of the LMM and bootstrap procedures outlined in Section 3.

This experimental design ensures replicable, statistically grounded evaluation of LLM performance, isolating configuration effects from linguistic randomness, and enabling interpretable decomposition of output variability.

5 RESULTS AND ANALYSIS

This section presents the empirical results of the proposed statistical framework for LLM evaluation and their relationship to the hypotheses formulated in Section 4.1. Hypothesis H1 is confirmed by the finding that fixed effects (model architecture, retrieval method, decoding parameters) explain 23.2% of the total variance. Hypothesis H2 is supported by the identification of prompt phrasing as a measurable source of random variability (7.2% of variance). Hypothesis H3 is validated through bootstrap resampling, which improved the stability of parameter estimates and produced narrower confidence intervals. Collectively, these results demonstrate that IMMBA reliably decomposes systematic and random sources of variability, offering interpretable and reproducible insights into LLM performance.

The descriptive statistics reveal the performance distribution of the evaluated models. The mean metric values indicate moderate performance: quality at 4.01, agreement at 3.98, accuracy at 3.29 and hallucination at 3.57. Significant variability, especially in agreement (SD 3.01) and hallucination (SD 3.45)[2], suggests inconsistent model performance, undermining reliability in NCM coding applications where precision and stability are vital to avoid errors. These results are summarised in Table 1 below.

The model evaluation showed *gpt-4o-mini-2024-07-18* and *deepseek-chat* excelled with high-quality

[2]Higher values for Quality, Agreement and Accuracy indicate better performance, while lower values for Hallucination are preferred.

Table 1: Descriptive performance Mean (SD) [variation coefficient in %].

Category	Model	Quality	Agreement	Accuracy	Hallucination
Best Performance	gpt-4o-mini	4.6 (2.4) [51.3]	4.5 (3.0) [66.2]	3.8 (2.8) [75.8]	2.4 (2.7) [114.8]
	deepseek-chat	5.1 (2.4) [46.6]	4.9 (3.2) [64.6]	4.4 (3.1) [69.9]	2.3 (2.7) [116.0]
Moderate Performance	TeenyTinyLlama	3.7 (2.2) [58.7]	3.7 (2.4) [64.9]	2.8 (1.9) [68.2]	5.5 (3.1) [55.9]
	gemini-2.0-flash	4.3 (2.3) [52.8]	4.5 (2.8) [63.4]	3.6 (2.8) [78.1]	2.3 (2.5) [109.5]
Lower Performance	Mistral-7B	2.3 (2.6) [111.1]	2.3 (2.9) [122.9]	1.9 (2.6) [134.3]	5.5 (4.2) [75.8]

scores (4.58, 5.10), agreement (4.49, 4.94), accuracy (3.76, 4.41) and low hallucination rates (2.37, 2.33), indicating reliability for NCM coding. In contrast, *Mistral-7B-Instruct-v0.3* had lower quality (2.34), accuracy (1.91) and high hallucinations (5.50), making it unsuitable for precise tasks.

TeenyTinyLlama is moderately performing with a quality of 3.73 and an accuracy of 2.77, but a high hallucination rate (5.36) may impact classification accuracy. *gemini-2.0-flash* offers a balanced profile, with a quality of 4.32, agreement of 4.47, accuracy of 3.58 and a low hallucination rate (2.31), indicating better stability.

Temperature and top-p parameters do not significantly affect quality, agreement, accuracy or hallucination metrics. However, a lower temperature (0.1) reduces hallucinations and ensures output consistency, making it suitable for high-reliability tasks like NCM coding. top-p variation shows no significant impact on these metrics.

5.1 Variance Decomposition

Total variability in LLM evaluation scores was partitioned using the Linear Mixed Model into three principal components:

$$\mathrm{Var}(Y_{ijk,pr}) = \sigma_f^2 + \sigma_P^2 + \sigma_e^2, \tag{3}$$

where:

- $\sigma_f^2 = 2.14$ reflects variance attributable to fixed experimental factors (model, temperature, top-p, retrieval method) and their interactions.
- $\sigma_P^2 = 0.66$ captures variability arising from prompt phrasing.
- $\sigma_e^2 = 6.42$ represents residual, unexplained noise.

These results indicate that while fixed effects account for a meaningful portion of observed variability, prompt-induced fluctuations and residual noise remain substantial, underscoring the necessity of hierarchical modelling in LLM evaluation. These components are summarised in Table 2 below.

5.2 Detection of Interaction Effects

Analysis of interaction terms within the LMM revealed several statistically significant relationships:

- Retrieval method significantly interacts with model architecture, with retrieval-augmented strategies yielding higher precision and reduced hallucination, particularly in smaller models.
- Temperature exhibited minimal main effects; however, its interaction with specific models (e.g., TeenyTinyLlama) amplified variability, highlighting configuration sensitivity.
- Higher top-p values (0.9) marginally reduced precision in some configurations, though interaction effects varied across models.

These findings demonstrate the framework's ability to detect nuanced dependencies between LLM parameters, informing targeted model tuning and risk mitigation.

5.3 Performance Metrics Through Scatter Matrix

The scatter matrix analysis reveals connections between quality, agreement, accuracy and hallucination, supporting Spearman's correlation analysis. Histograms show a multimodal distribution for quality and hallucination.

Unlike Pearson's, which assumes linearity, Spearman's correlation captures monotonic relationships without assuming linearity. It focuses on rank consistency, making it ideal for complex datasets where relationships might not be linear and crucial for evaluating model performance, where improvements aren't always proportional. The scatter and Spearman correlation matrices reveal the interdependence of quality, agreement and accuracy, whereas hallucination exhibits weak inverse correlations. The Benjamini-Hochberg correction boosts statistical robustness by reducing false positives, thereby strengthening the reliability of these relationships in assessing model performance.

Scatter plots in Figure 3 show a strong positive Spearman correlation between quality, agreement and

Table 2: Variance Decomposition in the IMMBA Model.

Component	**Symbol**	**Absolute Value**	**Percentage of Var(Y)**
Fixed Effects	σ_f^2	2.14	23.2%
Prompt Variability	σ_P^2	0.66	7.2%
Residual Error	σ_e^2	6.42	69.6%
Total	Var(Y)	9.22	100%

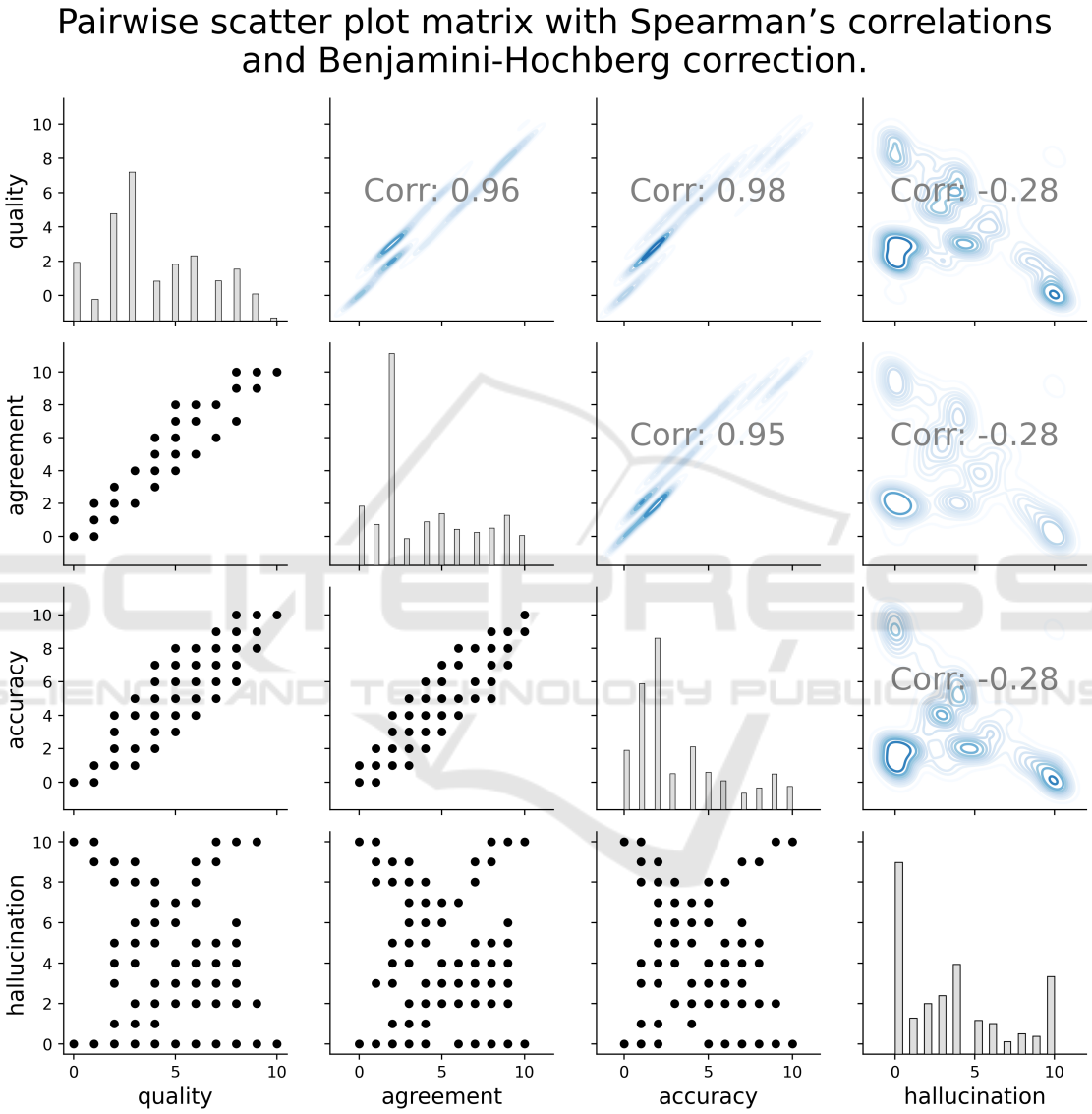

Figure 3: Pairwise scatter plot matrix with Spearman correlations between quality, agreement, accuracy and hallucination. Diagonal: histograms, lower triangle: scatter plots, upper triangle: density estimates. All correlations are significant (adjusted p-value < 0.05).

accuracy, with coefficients of 0.959, 0.977 and 0.955. Diagonal bands highlight a robust monotonic link: higher quality corresponds to higher agreement and accuracy. Kernel density plots emphasise concentrated value regions, reinforcing these structured relationships. Hallucination presents a weak but consistent negative Spearman correlation with quality (-0.277), agreement (-0.284) and accuracy (-0.284). In other words, higher quality, agreement and accuracy scores are generally associated with lower hallucination rates.

5.4 Bootstrap Estimation Robustness

Bootstrap resampling ($B = 1000$ iterations) was applied to all parameter estimates, yielding empirically derived confidence intervals and standard errors. The bootstrap-enhanced results exhibit:

- Increased stability of variance component estimates across resampled datasets.
- Robust detection of significant interaction terms, mitigating the influence of non-Gaussian output distributions.
- Narrower confidence intervals for fixed effects, supporting reliable attribution of observed performance differences.

This validates the utility of integrating bootstrap procedures within the evaluation pipeline, particularly for complex LLM outputs exhibiting stochastic behaviour and heteroscedasticity.

Table 3 provides the bootstrap-derived coefficient estimates, standard errors, and 95% confidence intervals for the fixed effects and their interactions within the IMMBA framework. These results directly support the hypotheses outlined in Section 4.1. First, H1 is confirmed by the statistical significance of model, retrieval, and decoding parameters, which account for a substantial proportion of systematic variance. Second, H2 is validated through the robust detection of prompt-level random variability, as indicated by the consistent width of the confidence intervals across resamples. Finally, H3 is substantiated by the narrow bootstrap confidence intervals and stable coefficient estimates, demonstrating that resampling improves reliability and mitigates the risks of overfitting to a single dataset realisation. Together, these findings confirm that IMMBA provides a statistically principled approach for decomposing and interpreting LLM performance variability.

5.5 Performance Interpretation

Decomposing variability clarifies the limitations of aggregate metrics in capturing LLM behaviour:

- Fixed effects explain 23.2% of total variance, highlighting the influence of model architecture and configuration.
- Prompt phrasing contributes 7.2% of variance, confirming the substantial role of linguistic formulation.
- Residual variability accounts for 69.6%, encompassing factors such as model stochasticity and scoring subjectivity.

These insights underscore the necessity of statistically grounded evaluation frameworks when comparing LLM configurations, particularly for high-stakes applications.

5.6 Limitations

While the proposed methodology offers significant improvements over conventional LLM evaluation approaches, several limitations warrant consideration:

- The experimental task, based on NCM classification, provides a structured domain for evaluation; generalisability to less structured or multilingual tasks requires further investigation.
- Human scoring introduces an element of subjectivity, despite the use of standardised prompts and rater protocols.
- Residual variance remains substantial, suggesting opportunities for refinement via ensemble prompting or advanced uncertainty quantification.

Despite these constraints, the integration of Linear Mixed Models with bootstrap resampling provides a robust, interpretable, and reproducible foundation for systematic LLM performance analysis.

6 DISCUSSION

The results of this study highlight the critical role of statistically grounded evaluation methodologies in understanding and optimising the behaviour of Large Language Models (LLMs). By integrating Linear Mixed Models (LMMs) with bootstrap resampling, we provide a principled framework capable of isolating systematic sources of variability, quantifying the influence of prompt phrasing, and identifying interaction effects across model configurations.

6.1 Broader Implications for LLM Research

The decomposition of performance variance reveals that a substantial proportion of observed fluctuations in LLM outputs arises not from model improvements alone, but from interactions between architectural choices, retrieval strategies, and stochastic decoding parameters. These findings align with emerging concerns in the academic community regarding the limitations of aggregate evaluation metrics, which often obscure critical sources of unreliability in LLM outputs.

Table 3: IMMBA Bootstrap Estimates of Fixed Effects with Standard Errors (SE) and 95% Confidence Intervals (CI). Results obtained with $B = 1000$ resamples of the factorial design, confirming H1–H3 by showing significant fixed effects, prompt-level variability, and stable CI ranges.

Coefficient	Coef. Mean (SE)	95% CI (Lower, Upper)
Intercept	3.94 (0.10)	(3.75, 4.13)
model[T.deepseek-chat]	1.41 (0.09)	(1.22, 1.59)
model[T.gemini-2.0-flash]	0.87 (0.09)	(0.70, 1.06)
model[T.gpt-4o-mini]	1.02 (0.09)	(0.83, 1.20)
model[T.TeenyTinyLlama]	2.05 (0.12)	(1.81, 2.27)
model[T.Mistral-7B]	-1.52 (0.14)	(-1.79, -1.25)
temperature[T.1.0]	0.13 (0.04)	(0.06, 0.20)
temperature[T.1.9]	0.34 (0.04)	(0.26, 0.43)
top_p[T.0.5]	0.16 (0.04)	(0.09, 0.24)
top_p[T.0.9]	0.31 (0.05)	(0.22, 0.40)
retrieval[T.Common]	-2.29 (0.08)	(-2.45, -2.12)
model[T.TeenyTinyLlama]:temperature[T.1.0]	-0.59 (0.06)	(-0.70, -0.49)
model[T.deepseek-chat]:temperature[T.1.0]	-0.05 (0.05)	(-0.13, 0.04)
model[T.gemini-2.0-flash]:temperature[T.1.0]	-0.06 (0.04)	(-0.15, 0.03)
model[T.gpt-4o-mini]:temperature[T.1.0]	-0.06 (0.04)	(-0.15, 0.01)

Recent studies have explored alternative evaluation approaches, such as entropy-based measures for cognitive modelling (H. Wang and Yu, 2022; C.-H. Liu and Wang, 2023) or cultural alignment assessments in multilingual models (J. Rystrøm and Hale, 2025). While these contributions advance specific dimensions of LLM evaluation, our work extends the methodological foundation by offering a unified, variance-decomposed perspective applicable to diverse LLM architectures and retrieval strategies.

The statistically significant interaction effects detected in our experiments further illustrate the necessity of hierarchical modelling in LLM evaluation, particularly when comparing configurations across tasks or domains. Without such modelling, practitioners risk drawing misleading conclusions based on incomplete or context-specific performance snapshots.

6.2 Relevance for Real-World Deployment

The practical implications of this work extend to high-stakes deployment scenarios, where output reliability, precision, and reproducibility are paramount. In domains such as legal decision support, fiscal classification, or healthcare information retrieval, failure to account for prompt-induced variability or configuration sensitivity can compromise both system performance and end-user trust.

This design avoids early search bias, enhances retrieval precision and preserves informational diversity, a critical feature in structured domains such as NCM, where classification ambiguity may entail fiscal or legal consequences. The significance of this issue lies in the fact that classification errors may result in fiscal, bureaucratic or legal repercussions, in addition to undermining trust in AI-based systems (Di Oliveira et al., 2024).

Our framework provides actionable insights for deployment:

- Quantifying prompt-level variability supports the development of robust prompting strategies, reducing susceptibility to linguistic ambiguity.
- Detecting configuration-specific interaction effects informs fine-tuning and parameter optimisation, enhancing output stability.
- Bootstrap-enhanced estimation mitigates overconfidence in performance assessments, promoting more reliable model comparisons.

These capabilities are particularly relevant as LLMs are increasingly integrated into retrieval-augmented pipelines, question answering systems, and decision-making tools, where uncontrolled variability may introduce unacceptable risk.

6.3 Advancing the Academic Conversation

By addressing the methodological gap in variance-decomposed LLM evaluation, particularly within retrieval-augmented settings, this work contributes to the growing body of research advocating for statistically principled AI evaluation practices. Our integration of LMMs and bootstrap procedures complements existing efforts to enhance robustness, transparency, and interpretability in LLM assessment.

Future research may extend this framework to multilingual or domain-adaptive LLMs, incorporate additional random effects (e.g., rater variability), or

explore integration with established psychometric approaches, such as Classical Test Theory, to further refine output reliability assessment.

7 CONCLUSIONS

This study proposed a statistically principled framework for evaluating Large Language Models (LLMs), integrating Linear Mixed Models with bootstrap resampling to decompose performance variability across model configurations, retrieval methods, and linguistic factors.

Empirical results demonstrate that a significant portion of LLM output variability arises from interactions between architectural choices, decoding parameters, and prompt phrasing, which are often obscured by conventional aggregate evaluation metrics. The proposed methodology enables systematic quantification of these effects, providing robust, interpretable insights into model behaviour.

Variance decomposition revealed that fixed effects account for 23.2% of output variability, while prompt phrasing contributes 7.2%, underscoring the need for hierarchical modelling in LLM assessment. Bootstrap-enhanced estimation further improved the reliability of parameter inference, mitigating overconfidence in performance comparisons.

By addressing a key methodological gap in LLM evaluation, this work advances current practice towards more rigorous, reproducible, and interpretable assessment standards. The framework supports both academic research and real-world deployment in high-stakes applications where output precision and reliability are essential.

Future work will explore extensions to multilingual evaluation, domain-adaptive LLMs, and integration with psychometric approaches such as Classical Test Theory, further enhancing the robustness and generalisability of LLM performance assessment.

7.1 Future Work

Building upon the proposed statistical evaluation framework, several avenues for future research are identified.

Firstly, extending the methodology to multilingual LLMs and domain-adaptive architectures would assess its generalisability beyond the structured classification tasks considered in this study. As LLM applications expand to increasingly diverse linguistic and contextual environments, robust, variance-decomposed evaluation will be essential to maintaining performance consistency.

Secondly, integrating additional random effects, such as rater variability or dataset-level heterogeneity, may further refine the attribution of output variability and enhance the precision of model comparisons.

Finally, future work will explore the combination of this framework with established psychometric techniques. Such integration may offer complementary insights into LLM reliability, particularly in high-stakes scenarios where both statistical and cognitive evaluation dimensions are relevant.

ACKNOWLEDGEMENTS

ChatGPT 4o was used in all sections of this work to standardise and improve the writing in British English. This research is partially funded by the Brazilian National Council for Scientific and Technological Development (CNPq).

REFERENCES

Abbasi Yadkori, Y. and Kuzborskij, I. (2024). To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bezerra, Y. F. and Weigang, L. (2025). Llmquoter: enhancing rag capabilities through efficient quote extraction from large contexts. *arXiv preprint arXiv:2501.05554*.

C.-H. Liu, C.-W. Chang, J. H. J. J. H. L. P. S. L.-A. L. C.-T. H. Y.-P. C. E. S. H. and Wang, S.-L. (2023). Brain computed tomography reading of stroke patients by resident doctors from different medical specialities: An eye-tracking study. *Journal of Clinical Neuroscience*, 117:173–180.

Di Oliveira, V., Bezerra, Y., Weigang, L., Brom, P., and Celestino, V. (2024). Slim-raft: A novel fine-tuning approach to improve cross-linguistic performance for mercosur common nomenclature. In *Proceedings of the 20th International Conference on Web Information Systems and Technologies - WEBIST*, pages 234–241. INSTICC, SciTePress.

Di Oliveira, V., Weigang, L., and Filho, G. P. R. (2022). Eleven data-set: A labeled set of descriptions of goods captured from brazilian electronic invoices. In *Proceedings of the 18th International Conference on Web Information Systems and Technologies - WEBIST*, pages 257–264. INSTICC, SciTePress.

H. Wang, F. Liu, Y. D. and Yu, D. (2022). Entropy of eye movement during rapid automatized naming. *Frontiers in Human Neuroscience*, 16.

J. Rystrøm, H. R. K. and Hale, S. (2025). Multilingual $\neq$ multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms.

Jiang, M., Ruan, Y., Huang, S., Liao, S., and Pitis, S. (2023). Calibrating language models via augmented prompt ensembles. In *International Conference on Learning Representations (ICLR)*.

Kapoor, S., Gruver, N., and Roberts, M. (2024). Large language models must be taught to know what they don't know. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Liu, X., Wong, D., Li, D., and Wang, Z. (2024). Selectit: Selective instruction tuning for llms via uncertainty-aware self-reflection. In *Advances in Neural Information Processing Systems (NeurIPS)*.

MERCOSUR (2024). Mercosur - consultas à nomenclatura comum e à tarifa externa. https://www.mercosur.int/pt-br/politica-comercial/ncm/ Accessed on Jun 4th, 2024.

Nikitin, A., Kossen, J., and Gal, Y. (2024). Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tonolini, F., Aletras, N., and Massiah, J. (2024). Bayesian prompt ensembles: Model uncertainty estimation for black-box large language models. In *Findings of the Association for Computational Linguistics: ACL*.

WCO (2018). *THE HARMONIZED SYSTEM A universal language for international trade*. World Customs Organization. https://www.wcoomd.org/-/media/wco/public/global/pdf/topics/nomenclature/activities-and-programmes/30-years-hs/hs-compendium.pdf (Visited 2024-06-04).

Weigang, L. and Brom, P. C. (2025). The paradox of poetic intent in back-translation: Evaluating the quality of large language models in chinese translation. *arXiv preprint arXiv:2504.16286*.

Yang, Z., Hao, S., Jiang, L., Gao, Q., and Ma, Y. (2024). Understanding the sources of uncertainty for large language and multimodal models. In *International Conference on Learning Representations (ICLR)*.

Zhou, Z., Tao, R., Zhu, J., and Luo, Y. (2024). Graph-based uncertainty metrics for long-form language model generations. In *Advances in Neural Information Processing Systems (NeurIPS)*.