Two-Step RAG for Metadata Filtering and Statistical LLM Evaluation

Vinícius Di Oliveira, Pedro Carvalho Brom, and Li Weigang, *IEEE, Senior Member*

***Abstract*—This study addresses a limitation in Retrieval-Augmented Generation (RAG) systems: poor retrieval accuracy when vague prompts or metadata are missing. We propose the *Two-Step RAG* method to overcome this. The first step performs a broad semantic search. The second uses an LLM to extract structured metadata to refine results through contextual filtering. This structure balances coverage and precision, proving effective in well-structured domains such as the Mercosur Common Nomenclature (NCM). The method is evaluated using a bootstrap-based multivariate linear mixed model, accounting for variability in temperature, top-p and prompt formulation. Two-Step RAG improves quality by a factor of 1.94, agreement by 2.31 and accuracy by 2.51 on average, while reducing hallucination to 0.82× compared to conventional RAG. It also shows reduced output variability in high-performing models, with coefficients of variation in quality dropping to 30–33% for *gpt-4o-mini* and *deepseek-chat*. These models achieve the best results, with accuracy gains exceeding 3× and hallucination reduced to 36–55% of the baseline. The method is robust across configurations and offers practical value for applications requiring high retrieval precision.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/9793

***Index Terms*—Retrieval-Augmented Generation, Metadata Filtering, Large Language Models, NCM Classification, Information Retrieval.**

I. Introduction

THE integration of language models with information retrieval, known as *Retrieval-Augmented Generation* (RAG), has proven effective for generating contextually grounded responses [6]. Nevertheless, conventional RAG systems struggle with underspecified prompts, as their reliance on semantic similarity often yields imprecise results [1], [3], [6], a limitation observed in both general-domain and domain-specific retrieval contexts. This is particularly problematic in domains requiring terminological accuracy, such as the classification of goods under the *Mercosur Common Nomenclature - NCM*[1], a regional taxonomy derived from the Harmonised System[2] maintained by the World Customs Organization[3].

The associate editor coordinating the review of this manuscript and approving it for publication was Bruno Henrique Groenner Barbosa (*Corresponding author: Vinícius Di Oliveira*).

Vinícius Di Oliveira, P. C. Brom, and Li Weigang are with the TransLab, University of Brasília, Brasília, Brazil (e-mails: vinicius-oliveira.vo@aluno.unb.br, pedro.brom@aluno.unb.br, and weigang@unb.br).

[1]https://www.mercosur.int/pt-br/politica-comercial/ncm/

[2]https://www.wcoomd.org/-/media/wco/public/global/pdf/topics/nomenclature/activities-and-programmes/30-years-hs/hs-compendium.pdf

[3]https://www.wcoomd.org/en/topics/nomenclature/overview/list-of-contracting-parties-to-the-hs-convention-and-countries-using-the-hs.aspx

Recent enhancements incorporate metadata to guide retrieval. For example, Multi-Meta-RAG utilises multiple metadata-enriched queries [1], and BlendFilter applies heuristic filters during query generation [3]. While promising, these approaches often depend on predefined schemas. This reliance limits both flexibility and scalability.

We introduce the *Two-Step RAG*, a structured framework with two distinct retrieval phases to address these limitations. Initially, a broad semantic search ensures wide coverage without relying on metadata. Subsequently, structured metadata is extracted from the prompt via a language model and applied as a filter to refine results. This dynamic, schema-agnostic strategy improves precision without restricting initial retrieval scope, a critical advantage in domains like NCM, where ambiguity may lead to fiscal or legal implications.

Empirical evaluation confirms that Two-Step RAG enhances output quality and reduces hallucinations compared to traditional RAG. Its performance was analysed using a bootstrap-based multivariate linear mixed model, decomposing variability across 54 configurations varying temperature, top-p, and RAG type. This statistical approach supports rigorous, replicable assessment of LLMs, offering a robust framework for high-stakes applications demanding both precision and adaptability.

This paper is organised as follows. Section II reviews related work on metadata-guided retrieval and language model evaluation. Section III details the proposed Two-Step RAG methodology. Experimental design in Section IV, results and statistical analyses are presented in Section V. Finally, Section VI concludes the paper and outlines directions for future research.

II. Related Works and Dataset

The evolution of Retrieval-Augmented Generation (RAG) systems has led to various strategies to improve retrieval effectiveness, particularly under vague or metadata-sparse prompts. This section surveys contributions to metadata-aware semantic retrieval and language model evaluation, particularly statistical methods and dataset design.

A. Metadata-Aware Retrieval in RAG Systems

Recent advances in RAG have increasingly acknowledged the value of metadata in enhancing contextual accuracy. Traditional RAG systems rely primarily on prompt–document semantic similarity. This often results in suboptimal retrieval when the context is domain-specific or ambiguous. To address this, Multi-Meta-RAG [1] employs parallel queries augmented

with metadata variations, while BlendFilter [3] integrates heuristic filters into the query generation pipeline. Meta Knowledge RAG [4] similarly leverages metadata signals to guide retrieval from summarised document clusters. Other research has explored complementary approaches such as adaptive re-ranking using dense–sparse hybrid retrieval [14], query decomposition for complex information needs [16], and hybrid symbolic–neural retrieval frameworks [15], which, while not strictly metadata-aware, share the aim of improving retrieval relevance in specialised domains. These approaches often rely on rigid schemas or heuristics, limiting domain generalisation. Beyond metadata filtering, methods such as GraphRAG [13] use graph-structured knowledge to enhance retrieval in contexts needing terminological precision. Though not metadata-centric, they face similar challenges and illustrate the breadth of ongoing efforts to improve retrieval quality.

The Two-Step RAG proposed herein introduces a modular retrieval pipeline consisting of semantic search (R1), language model-driven metadata extraction (M), and post-hoc filtering (R2). This structure mitigates premature narrowing of the retrieval space while enabling schema-agnostic filtering. By dynamically extracting structured information from natural language prompts, the method supports broader applicability across heterogeneous taxonomies such as the Mercosur Common Nomenclature (NCM).

B. Statistical Evaluation of Language Models

Although benchmark scores remain widespread in LLM assessment, recent literature has proposed statistically principled alternatives to capture model behavior under varying conditions. Linear mixed-effects models (LMMs) provide a robust framework by decomposing fixed and random sources of performance variability. While some studies have applied entropy-based measures to examine cognitive processing during language and visual tasks [7], [8], others have explored model alignment with cultural dimensions in multilingual contexts [9]. These approaches support a more nuanced understanding of LLM outputs beyond aggregate performance scores.

While LMMs account for prompt-level variance, Bootstrap methods have been increasingly adopted to improve robustness against violations of parametric assumptions and to construct confidence intervals [10]. Nevertheless, few studies have explicitly integrated Bootstrap analysis in RAG systems, leaving a methodological gap in retrieval-focused LLM evaluation.

C. NCM Data Set

The challenge of retrieving accurate results from vague prompts is particularly evident in tasks requiring precise classification, such as NCM coding. Errors in classification may entail fiscal, bureaucratic, or legal risks, underscoring the necessity of both contextual sensitivity and retrieval accuracy [11]. We employed the Eleven Dataset [12], a structured corpus of commodity descriptions extracted from Brazilian Electronic Invoices, to evaluate the proposed methodology. The dataset reflects real-world usage of the Mercosur Common Nomenclature[4] and provides a robust foundation for assessing LLMs under metadata-scarce scenarios.

III. Extracting Metadata for Filtering with 2-Steps RAG

The 2-Steps RAG framework was designed to address key limitations in traditional RAG systems, which often yield irrelevant outputs when faced with underspecified prompts. While models such as Multi-Meta-RAG [1] apply metadata-based filtering via large language models (LLMs), they typically lack adaptability across domains due to rigid schemas. Our method enhances retrieval accuracy by structuring the process into two sequential stages, illustrated in Fig. 1.

1) **Common Context Retrieval:** The first phase performs an unconstrained semantic search, ensuring broad retrieval coverage even when metadata is absent or noisy.
2) **Metadata Extraction and Application:** In the second phase, a language model extracts structured metadata, such as NCM code, label, item, and product, from the prompt. This metadata is then applied to filter the initially retrieved documents, refining the context. Unlike systems such as DAQu [2], which rely on relational databases, this approach combines semantic embeddings with adaptive, LLM-based filtering.

In contrast to models such as BlendFilter [3] and Meta Knowledge RAG [4], which depend on predefined rules or summarisation heuristics, 2-Steps RAG introduces a dynamic filtering stage based on prompt-derived metadata. This allows the system to maintain high recall in the initial retrieval while improving precision through context-sensitive post-filtering. By deferring constraints until after a broad semantic search, this approach achieves a flexible balance between coverage and specificity, which is particularly advantageous in structured domains such as NCM classification.

A. Mathematical Formulation of 2-Steps RAG

The 2-Steps RAG approach is a two-phase structured information retrieval process. First, a semantic embedding-based model conducts a broad document search to return initial candidates. Next, a language model refines the retrieval by extracting structured metadata from the prompt to create a precise augmented query. Algorithm 1 shows this process, which is detailed below.

Let $\mathcal{D} = \{d_1, d_2, ..., d_N\}$ be a set of documents, where each document d_i contains textual information along with associated metadata. Now we have a sequence of tokens representing a given Original Prompt:

$$P = (w_1, w_2, ..., w_m), \tag{1}$$

where w_j belongs to a vocabulary $\mathcal{V}$. To enable retrieval, we define an embedding function φ that maps a prompt to a d-dimensional vector space:

$$\varphi : P \to \mathbb{R}^d. \tag{2}$$

[4]https://www.mercosur.int/en/about-mercosur/mercosur-countries/

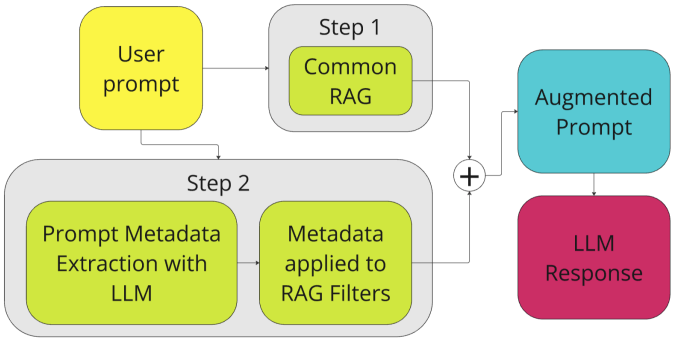

Fig. 1. Illustration of the 2-Steps RAG methodology: The retrieval process is structured into two sequential stages

The knowledge base is indexed using document embeddings, forming a structured set:

$$\mathcal{K} = \{(\varphi(d_1), d_1), (\varphi(d_2), d_2), ..., (\varphi(d_N), d_N)\}. \quad (3)$$

1) Step 1: Common Retrieval (R_1): The first stage, called **common retrieval** (R_1), searches for relevant documents based on the cosine similarity between the prompt embedding $\varphi(P)$ and the document embeddings stored in $\mathcal{K}$. The following equation defines the retrieval process:

$$R_1(P, \mathcal{K}) = \underset{d_i \in \mathcal{D}}{\text{argmax}} \ \cos(\varphi(P), \varphi(d_i)). \quad (4)$$

This function returns a subset of relevant documents, denoted as $D_1 \subseteq \mathcal{D}$, based on semantic similarity.

2) Step 2: Metadata Extraction (M) and Metadata-Driven Filtering (R_2): A language model extracts structured attributes from the prompt in the second step, known as **metadata extraction** (M). The following equation defines the extraction function:

$$M : P \rightarrow \{(k_1, v_1), (k_2, v_2), ..., (k_n, v_n)\}, \quad (5)$$

where each pair (k_i, v_i) represents a metadata key and its corresponding value. The process is executed by a Large Language Model (LLM) as follows:

$$M(P) = \text{LLM}_{\text{metadata}}(P). \quad (6)$$

Algorithm 1 2-Steps RAG with Metadata Filtering

Require: Prompt P, Knowledge Base $\mathcal{K} = \{(\varphi(d_i), d_i)\}_{i=1}^{N}$
Ensure: Augmented Prompt $A(P, D_1, D_2)$
Compute prompt embedding: $\varphi(P)$
Step 1: Common Retrieval
Retrieve initial candidates by semantic similarity
$D_1 \leftarrow R_1(P, \mathcal{K}) = \arg\max_{d_i \in \mathcal{D}} \cos(\varphi(P), \varphi(d_i))$
Step 2: Metadata Extraction
Extract metadata from the prompt
$\mathcal{M} \leftarrow M(P) = \text{LLM}_{\text{metadata}}(P)$
if $\mathcal{M} \neq \emptyset$ **then**
    Step 2: Metadata-Driven Filtering
    Filter documents in D_1 using metadata $\mathcal{M}$
    $D_2 \leftarrow R_2(D_1, \mathcal{M}) = \{d_i \in D_1 \mid \forall (k, v) \in \mathcal{M},\ d_i[k] = v\}$
else
    $D_2 \leftarrow D_1$
end if
Step 3: Prompt Augmentation
Combine P with filtered context
$A(P, D_1, D_2) \leftarrow \text{Concat}(P, \text{Context}(D_1, D_2))$
return Augmented Prompt $A(P, D_1, D_2)$

If no metadata is extracted, the system uses the D_1 dataset without modification. In practice, accuracy is reinforced by validating extracted attributes against the NCM vocabulary. When extraction fails, the system defaults to the Step 1 results (common RAG). In our experiments, this fallback occurred in 6.8% of prompts.

The internal step, called metadata-driven filtering (R_2), refines the results from D_1 by applying the extracted metadata. The following equation defines the filtered document set D_2 as:

$$R_2(D_1, \mathcal{M}) = \{d_i \in D_1 \mid \forall (k, v) \in \mathcal{M}, d_i[k] = v\}. \quad (7)$$

Here, $d_i[k]$ represents the metadata value k associated with document d_i. This filtering step ensures that only the most relevant documents are included according to the extracted metadata.

3) Augmented Prompt: Finally, the augmented prompt (A) combines the original prompt with the refined document set D_1, D_2 to provide a more precise context. The augmentation function is given by:

$$A(P, D_1, D_2) = \text{Concat}(P, \text{Context}(D_1, D_2)), \quad (8)$$

where $\text{Context}(D_1, D_2)$ represents the structured formatting of the filtered documents. This model formalizes the 2-Steps

TABLE I
COMPARISON BETWEEN 2-STEPS RAG AND RELATED METHODS BY CRITERIA AND CONTRIBUTIONS

Criterion	**2-Steps RAG**	**Multi-Meta-RAG** [1]	**BlendFilter** [3]	**DAQu** [2]	**2-Steps RAG Contributions**
Retrieval Steps	Two steps: broad (unfiltered) + refined with extracted metadata	Multiple parallel searches with different metadata	Query with embedded filters	Single SQL query	Avoids early bias by applying filters only after broad context retrieval
Metadata Extraction	Yes – dynamically via LLM	Yes – generally predefined	Implicit in the prompt	Not available	Flexible, does not rely on a fixed schema
Metadata Application	Post-retrieval, as filter	Pre-retrieval, defines multiple searches	During query generation	Pre-retrieval via SQL	Maintains high coverage before filtering, reducing loss of useful documents
Schema/Ontology Dependency	Low – automatic inference	Medium – requires defined attributes	High – rigid filters	High – requires relational mapping	Does not require specific data structure, enabling easier adaptation
Formal Model	Modular: R_1, M, R_2	Parallel: $\bigcup R_i$	Composite: $Q(P) \rightarrow R(Q)$	Structured: $Q_{SQL} \rightarrow R(Q)$	Clear and extensible structure, eases analysis and implementation
Domain Adaptability	High	Medium	Low to medium	Low	Works across domains with minimal configuration
Robustness to Ambiguous Prompts	High – broad retrieval ensures context	Medium – multiple variations help partially	Medium – depends on well-formed prompts	Low – structure requires well-defined query	Performs well even with vague or incomplete prompts
Primary Objective	Balance recall and precision with post-retrieval semantic filtering	Maximize recall via diverse queries	Reduce noise using integrated filters	Precise queries on structured databases	Ensures precision without compromising initial coverage

RAG mechanism, highlighting its ability to balance broad initial retrieval with refined filtering based on dynamically extracted metadata.

This approach improves retrieval precision without prematurely restricting the search space, allowing for an adaptive trade-off between generality and specificity in the retrieval process.

B. 2-Steps RAG Comparative Advantages and Case Study

The 2-Steps RAG framework differs fundamentally from Multi-Meta-RAG [1], BlendFilter [3], and DAQu [2]. While the former two rely on parallelised queries or embedded filters, they often impose constraints during initial retrieval, potentially excluding relevant documents. By contrast, 2-Steps RAG separates broad semantic search (R_1) from targeted filtering (R_2), with metadata dynamically extracted via a language model (M), thus maintaining high recall without premature narrowing of the search space.

Unlike DAQu, which is schema-dependent and operates on structured databases, 2-Steps RAG is ontology-agnostic, inferring metadata directly from unstructured prompts. This enhances generalisability to heterogeneous taxonomies such as the Mercosur Common Nomenclature. The modular design—comprising $\varphi(P)$, $R_1(P, \mathcal{K})$, $M(P)$, and $R_2(D_1, \mathcal{M})$—offers interpretability, flexibility, and robustness. Table I outlines key differences, showing 2-Steps RAG's superior robustness and schema flexibility.

To illustrate practical applicability, we present a case study involving NCM code classification. The system is tasked with answering: *What is the official description of the NCM code for the goods with the following description:* {product description}*?* This scenario evaluates the model's ability to retrieve and generate accurate, context-aware responses.

While conventional RAG fails to locate the correct classification despite explicit cues, the Two-Step RAG successfully retrieves and articulates the appropriate information with high precision via broad retrieval and prompt-driven metadata filtering.

C. Applying Mathematical Layers

Building on the formal definitions in Section III.A, the retrieval and filtering process can be illustrated through a simplified walkthrough.

a) Step 1: Common Retrieval (R_1): The retrieval function R_1 returns a candidate set D_1 based solely on semantic similarity between the prompt and the indexed knowledge base. While effective for coverage, this step may retrieve documents that are related yet not specific to the target product, showing the limits of R_1 alone.

b) Step 2: Metadata Extraction (M) and Filtering (R_2): The metadata function M extracts structured attributes from the prompt—such as NCM code or product label—which are applied in R_2 to filter D_1. Extracted values are validated against the NCM vocabulary; if no valid metadata is found, the system defaults to D_1. In practice, this often narrows results to a single precise match in D_2.

c) Augmented Prompt (A): The final augmented prompt A merges the original input with context from D_1 and D_2, supplying the LLM with refined yet comprehensive evidence for generating a more accurate and contextually grounded response.

IV. Experimental Design

This section introduces the statistical framework adopted to analyse the experimental results, considering multiple control variables and fixed parameters to ensure consistency and robustness.

Sample size determination followed Cohen's power analysis guidelines [5], targeting an F-test in ANOVA and Linear Mixed Models (LMMs) with prompts as random effects. Effect size $f = 0.25$, $\alpha = 0.05$, and power = 0.8 were used. Based on a $3 \times 3 \times 6$ factorial design—three temperatures (0.1, 1.0, 1.9), three top-p values (0.1, 0.5, 0.9), and six LLMs—54 experimental conditions were replicated 196 times each. The models were selected to ensure diversity in architecture, licence type, and computational footprint. This included: (i) High-capacity proprietary models (GPT-4o-mini, Gemini-2.0-flash) to establish upper-bound performance; (ii) Mid-range open-source models (Mistral-7B) for accessible deployment; (iii) Small-footprint models (TeenyTinyLlama) for low-resource environments; (iv) Models trained with different multilingual corpora to assess cross-linguistic robustness. This diversity enables assessment of the Two-Step RAG's adaptability across different model classes and operational contexts. The full derivation is available at the project GitHub[5].

To isolate prompt-induced variability, we employed an LMM treating prompts as random effects:

$$Y_{ijk,pr} = \mu + A_i + B_j + C_k + R_r + (AB)_{ij} + (AC)_{ik} + (BC)_{jk} + (ABC)_{ijk} + P_p + \epsilon_{ijkr}, \quad (9)$$

where $Y_{ijk,pr}$ is the output for model A_i, temperature B_j, top-p C_k, retrieval method R_r, and prompt P_p. The prompt effect is $P_p \sim N(0, \sigma_P^2)$, and residual error $\epsilon_{ijkr} \sim N(0, \sigma^2)$.

Bootstrap resampling was applied to estimate parameters and mitigate assumptions like normality and homoscedasticity [10]. This strengthens inference reliability and enhances generalisability.

Total variance decomposition:

$$\text{Var}(Y_{ijk,pr}) = \sigma_f^2 + \sigma_P^2 + \sigma_e^2, \quad (10)$$

allocates variability across fixed effects (σ_f^2), prompt phrasing (σ_P^2), and residual error (σ_e^2). The observed total variance was 9.2202, underscoring the significance of modelling random effects.

This decomposition enables a precise attribution of performance variability to retrieval strategy, model settings, and prompt phrasing. Comparing 2-Step RAG with baseline RAG, the framework isolates method efficacy from prompt-related fluctuation and experimental noise, avoiding spurious conclusions and preserving the validity of inference.

A. Research Hypotheses

To assess the effectiveness of the Two-Step RAG and the influence of experimental factors, we propose the following hypotheses:

[5]https://github.com/pcbrom/2-Steps-RAG

1) Performance and Retrieval Efficiency:

- **H1:** Two-Step RAG yields higher retrieval precision than conventional RAG.
- **H2:** Metadata filtering in Step 2 reduces noise without reducing diversity.
- **H3:** Performance improves in semantically structured domains (e.g., NCM, Zip Code).

2) Statistical Interactions and Model Tuning:

- **H4:** ANOVA and LMMs detect significant interactions between retrieval method, temperature and top-p, enabling fine-tuning of LLM performance.

B. Evaluation Prompt

Model responses are evaluated against a baseline using a 0–10 scale across four criteria:

- **Quality**: Clarity, conciseness and structure of the response.
- **Agreement**: Semantic alignment with the baseline.
- **Accuracy**: Factual correctness of the response.
- **Hallucination**: Degree of fabrication or unverifiable content.

This design ensures consistent and quantifiable evaluation across models and settings by: (i) Applying identical evaluation prompts across all experimental conditions; (ii) Using a fixed baseline response as a common reference point; (iii) Capturing evaluation outputs in JSON format with fixed keys (quality, agreement, accuracy, hallucination), enabling automated parsing and statistical aggregation; (iv) Implementing the same penalty criteria for incomplete or evasive responses across all configurations.

Here is the **Proposed prompt**.

Portuguese:

> *Avalie as respostas abaixo de 0 a 10, em que 0 representa discordância total e 10 concordância total, considerando os seguintes critérios:* Baseline Response: {*row['baseline']*}
> Model Response: {*row['results']*} *Atribua uma nota de 0 a 10 para a comparação entre as respostas (Baseline e Modelo), justificando brevemente sua avaliação. Respostas do modelo que indicam falta de acesso a informações específicas e recomendam consultar fontes externas (ex: "Desculpe, mas não tenho acesso a informações específicas...") devem ser penalizadas, pois não atingiram o objetivo. Retorne a avaliação no formato JSON, com as chaves quality, agreement, accuracy, hallucination e justification. Não penalize a avaliação em caso de repetições no texto.*

English:

> Rate the responses below from 0 to 10, where 0 represents total disagreement and 10 total agreement, considering the following criteria: Baseline Response: {*row['baseline']*}
> Model Response: {*row['results']*} Give a score from 0 to 10 for comparing the responses (Baseline and Model), briefly justifying your evaluation. Model

TABLE II
TESTED HYPOTHESES AND CORRESPONDING RESULTS WITH THE INTERPRETATION EXPLAINING

Hypothesis	Obtained Result	Interpretation
H1: The 2-Step RAG improves document retrieval precision compared to conventional RAG.	Confirmed – The 2-Step RAG achieved higher average quality (5.30 vs. 2.73), better precision (4.70 vs. 2.12), and a lower hallucination rate (3.93 vs. 4.67).	Refining the search with metadata enhances contextual retrieval.
H2: Applying metadata reduces retrieval noise without compromising the diversity of results.	Confirmed – The variability of responses was lower with the 2-Step RAG, with no significant loss in diversity.	The model maintains retrieval breadth while improving answer specificity.
H3: The performance of the 2-Step RAG varies depending on the document domain.	Partially Confirmed – There was a significant improvement in domains with well-structured metadata (e.g., NCM), but the difference was less pronounced in domains lacking clear metadata.	The method's effectiveness depends on the quality and structure of the extracted metadata.
H4: Mixed Linear Models can identify significant interactions between temperature, top-p and the retrieval method.	Confirmed – The effect of temperature was minimal, while a higher top-p (0.9) slightly reduced precision in some models.	The statistical robustness of the model allowed the detection of these subtle variations.

responses that indicate a lack of access to specific information and recommend consulting external sources (e.g. "Sorry, but I don't have access to specific information...") should be penalised, as they did not achieve the objective. Return the evaluation in JSON format, with the keys quality, agreement, accuracy, hallucination and justification. Do not penalise the evaluation for text repetitions.

Justification for the Evaluation Criteria: The evaluation prompt ensures consistency through objective, quantifiable metrics across configurations. Each criterion addresses a key dimension of response assessment:

- **Quality**: Measures clarity, structure and conciseness.
- **Agreement**: Assesses contextual alignment with the baseline.
- **Accuracy**: Verifies factual consistency.
- **Hallucination**: Penalises fabricated or unverifiable content.

A standardised JSON format supports automated evaluation. Responses expressing uncertainty or deferring judgment are penalised to favour complete, self-contained outputs—enhancing assessment reliability and comparability.

V. RESULTS AND DISCUSSION

Bootstrap-based evaluation ($n = 1000$) confirms hypotheses H1 and shows that the Two-Step RAG significantly outperforms Common RAG across all metrics: quality (5.30 vs. 2.73), agreement (5.56 vs. 2.41), accuracy (4.70 vs. 2.12) and hallucination (3.93 vs. 4.67). Table II summarises the hypothesis tests supporting these findings.

A. Exploratory Performance and Model Variability

Descriptive analysis shows moderate mean scores—quality (4.01), agreement (3.98), accuracy (3.29)—but high variability, especially in hallucination (SD 3.45)[6]. The 2-Step RAG offers consistent gains across models. Performance improvements range from $\times 1.02$ to $\times 4.54$, with GPT-4o-mini and deepseek-chat achieving the most reliable outcomes.

Low temperature settings reduce hallucinations, while top-p variations have limited impact. Table III lists accuracy and hallucination for the main models.

[6]Higher scores in Quality, Agreement and Accuracy are better; lower scores in Hallucination are preferred.

B. Correlation Analysis and Scatter Matrix

Scatter and correlation analyses (Fig. 2) show strong positive associations: quality–accuracy ($\rho = 0.955$), agreement–accuracy ($\rho = 0.977$), while hallucination is negatively correlated with all three (e.g., $\rho = -0.284$ for agreement). Distributions further highlight the stability of 2-Step RAG.

C. Bootstrap Linear Mixed Model Results

Bootstrap-based MLMM results (Table V) confirm GPT-4o-mini and deepseek-chat offer the most consistent gains. Common RAG negatively affects performance. Interaction terms reveal TeenyTinyLlama's sensitivity to temperature and top-p.

D. Variance Decomposition

Model performance variance splits as follows: fixed effects ($\sigma_f^2 = 2.14$), prompt variability ($\sigma_P^2 = 0.66$), and residual noise ($\sigma_e^2 = 6.42$). These findings underscore prompt phrasing's influence and the utility of bootstrap for non-parametric estimation.

Implications: 2-Step RAG paired with robust models like GPT-4o-mini improves NCM classification accuracy and stability, making it suitable for high-stakes domains (e.g., legal, fiscal). Models like Mistral-7B, due to high hallucinations and poor accuracy, should be avoided.

E. Threats to Validity

Although the study employed rigorous statistical controls, some limitations remain. Semi-subjective evaluation criteria may introduce bias, mitigated here through a fixed JSON

TABLE III
MODELS DESCRIPTIVE PERFORMANCE CATEGORIZED: MEAN (SD) [COEFFICIENT OF VARIATION 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]

TABLE IV
METRICS PERFORMANCE RESULTS BY MODEL AND RAG TYPE: MEAN (SD) [CV%] AND 2-STEP VS. COMMON RATIO

Model	Group	Quality	Agreement	Accuracy	Hallucination
Mistral-7B	2-Step	3.64 (2.81) [77.2]	3.73 (3.23) [86.6]	3.13 (2.99) [95.5]	4.44 (3.75) [84.5]
Mistral-7B	Common	1.03 (1.47) [142.7]	0.95 (1.47) [154.7]	0.69 (1.13) [163.8]	6.56 (4.30) [65.6]
Mistral-7B	Ratio	×3.53	×3.93	×4.54	×0.67
TeenyTinyLlama	2-Step	3.77 (2.21) [58.6]	3.74 (2.40) [64.2]	2.80 (1.90)	5.38 (3.08) [57.2]
TeenyTinyLlama	Common	3.68 (2.17) [59.0]	3.65 (2.39) [65.5]	2.73 (1.87) [68.5]	5.33 (3.10) [58.2]
TeenyTinyLlama	Ratio	×1.02	×1.02	×1.03	×1.01
deepseek-chat	2-Step	6.76 (2.07) [30.6]	7.20 (2.69) [37.4]	6.45 (2.90) [67.9]	3.01 (3.20) [106.3]
deepseek-chat	Common	3.43 (1.20) [35.0]	2.68 (1.70) [63.4]	2.38 (1.52) [63.9]	1.65 (1.86) [112.7]
deepseek-chat	Ratio	×1.97	×2.69	×2.71	×1.82
gemini-2.0-flash	2-Step	5.95 (2.01) [33.8]	6.48 (2.51) [38.7]	5.44 (2.71) [49.8]	3.33 (2.67) [80.2]
gemini-2.0-flash	Common	2.69 (1.01) [37.6]	2.45 (1.28) [53.2]	1.73 (1.22) [70.5]	1.29 (1.89) [146.5]
gemini-2.0-flash	Ratio	×2.21	×2.64	×3.14	×2.58
GPT-4o-mini	2-Step	6.35 (1.93) [30.4]	6.67 (2.54) [36.7]	5.69 (2.74) [48.2]	3.49 (2.88) [82.5]
GPT-4o-mini	Common	2.81 (1.03) [36.7]	2.31 (1.30) [56.3]	1.83 (1.11) [60.7]	1.24 (1.98) [152.4]
GPT-4o-mini	Ratio	×2.26	×2.89	×3.11	×2.81

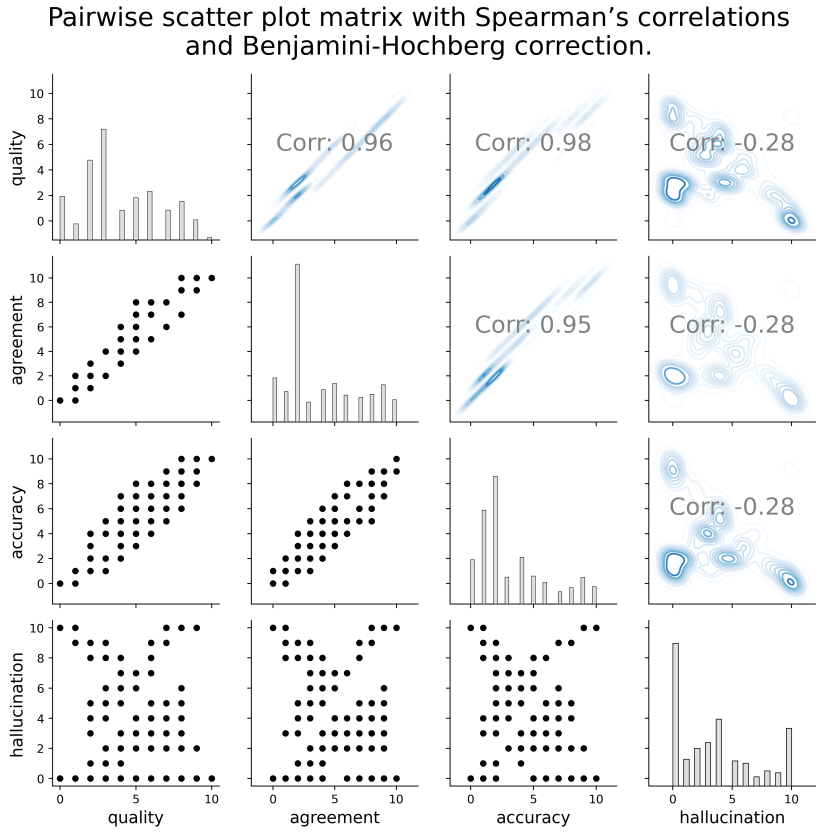

Fig. 2. Pairwise Spearman correlations: Shows scatter and correlation analyses between the parameters.

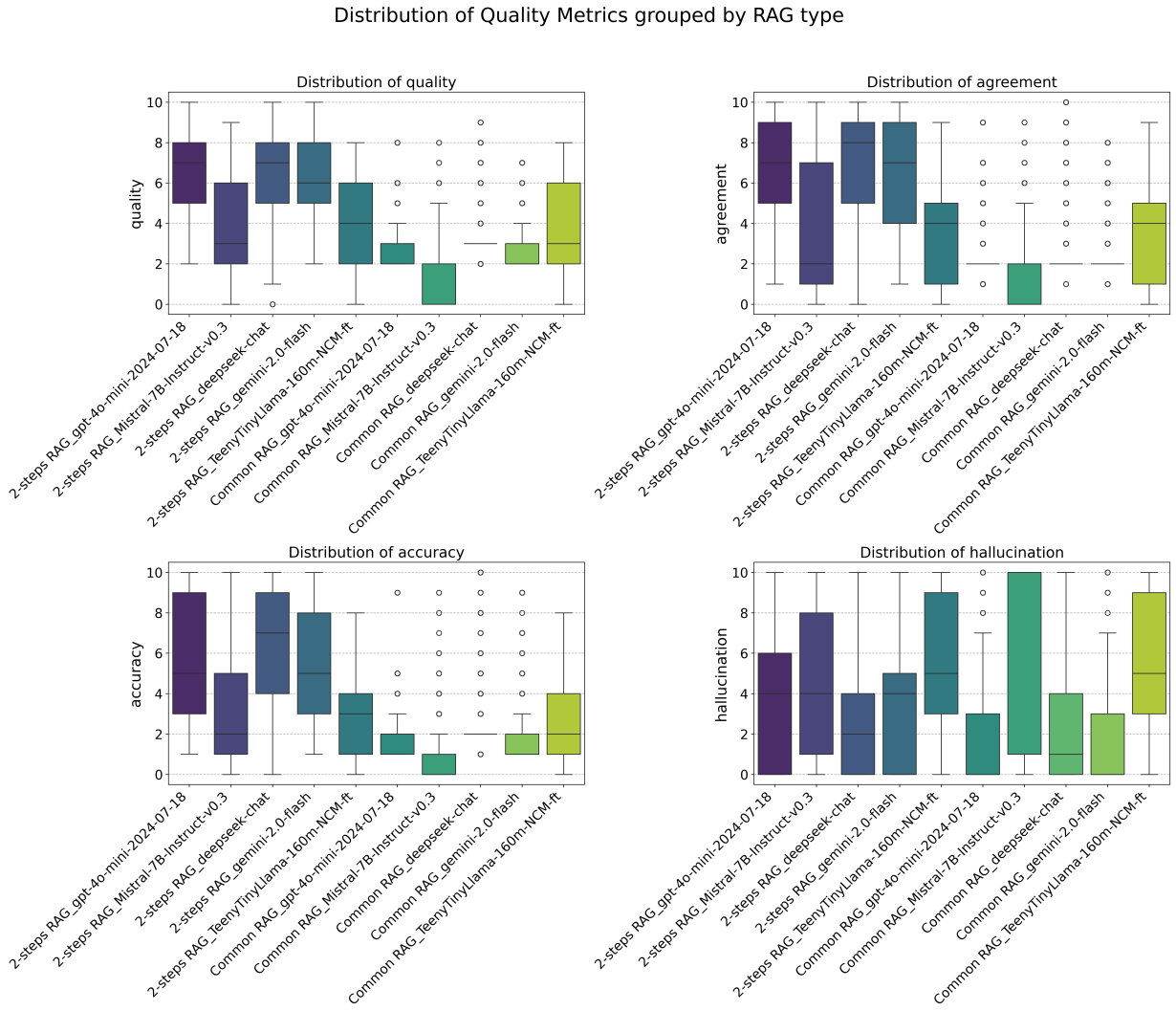

Fig. 3. Results: The boxplots metrics, quality, agreement, accuracy and hallucination distributions by model and RAG type.

TABLE V
BOOTSTRAP COEFFICIENTS WITH STANDARD ERRORS (SE) AND 95% CONFIDENCE INTERVALS (CI)

Coefficient	Coef. Mean (SE)	95% CI (Lower, Upper)
Intercept	3.9103 (0.0993)	(3.7219, 4.1138)
model[T.TeenyTinyLlama]	2.0183 (0.1179)	(1.7984, 2.2549)
model[T.deepseek-chat]	1.3930 (0.0985)	(1.1971, 1.5855)
model[T.gemini-2.0-flash]	0.8835 (0.0924)	(0.6960, 1.0619)
model[T.gpt-4o-mini]	0.9936 (0.0944)	(0.8092, 1.1784)
temperature[T.1.0]	0.1253 (0.0351)	(0.0579, 0.1973)
temperature[T.1.9]	0.3559 (0.0410)	(0.2789, 0.4427)
top_p[T.0.5]	0.1665 (0.0370)	(0.0916, 0.2427)
top_p[T.0.9]	0.3131 (0.0466)	(0.2238, 0.4020)
rag_type[T.Common RAG]	-2.3191 (0.0836)	(-2.4858, -2.1500)
model[T.TeenyTinyLlama]:temp[T.1.0]	-0.5922 (0.0558)	(-0.7020, -0.4909)
model[T.deepseek-chat]:temp[T.1.0]	-0.0451 (0.0452)	(-0.1363, 0.0396)
model[T.gemini-2.0]:temp[T.1.0]	-0.0565 (0.0434)	(-0.1522, 0.0258)
model[T.gpt-4o-mini]:temp[T.1.0]	-0.0638 (0.0415)	(-0.1514, 0.0132)

schema, penalty rules, and uniform prompts. Results are also domain-specific, as performance may vary in less structured or noisier contexts. Our metrics capture retrieval relevance and factual accuracy but not all aspects of user-perceived utility. Finally, while the factorial design and bootstrap-based linear mixed models reduce statistical error, residual confounders such as prompt variation cannot be fully excluded. These considerations guide the interpretation of findings and inform future replication.

F. Limitations

- The approach depends on metadata quality—poorly structured data reduces effectiveness.
- Generalisability beyond the NCM domain remains untested. Our evaluation compares Two-Step RAG only with a conventional RAG baseline and not with other metadata-enhanced methods such as Multi-Meta-RAG or BlendFilter. Applying these approaches in the NCM context would require substantial schema engineering, which we identify as future work.
- Evaluation relies on semi-subjective metrics; future work should pursue more standardised frameworks.

VI. Conclusions

This study proposed a statistically grounded framework for evaluating Retrieval-Augmented Generation (RAG) strategies, highlighting the effectiveness of the Two-Step RAG method in structured classification tasks, particularly under the Mercosur Common Nomenclature (NCM). Using a bootstrap-based multivariate linear mixed model, we decomposed the total variability into fixed effects ($\sigma_f^2 = 2.14$), prompt-induced variability ($\sigma_P^2 = 0.66$) and residual error ($\sigma_e^2 = 6.42$), allowing for precise attribution of performance variation across experimental conditions.

Two-Step RAG significantly improved results over conventional RAG: quality rose from 2.73 to 5.30, agreement from 2.41 to 5.56, and accuracy from 2.12 to 4.70, with hallucination dropping from 4.67 to 3.93. Models like *GPT-4o-mini* and *deepseek-chat* were especially effective, showing over $3\times$ accuracy gains and low variability. Even smaller models such as *Mistral-7B* showed relative accuracy improvements of $4.5\times$.

While the method has been designed with potential applications in high-stakes domains such as legal and fiscal classification, our empirical validation is currently limited to the NCM domain; in e-commerce, it enhances product categorisation and search relevance. Its separation of semantic retrieval and metadata filtering ensures flexibility, scalability and reliability across public-sector use cases. Tested on real NCM data from Brazil's Secretariat of Economy, it supports practical adoption under institutional constraints.

Nonetheless, caution is warranted regarding the use of LLMs for metadata generation in step M. Since this process relies on the model's semantic interpretation of retrieved content, the extracted metadata may, in some cases, reflect inductive biases or introduce hallucinated elements not grounded in the original documents. For deployment in real-world settings, these risks necessitate safeguards such as automated metadata validation layers, cross-checking extracted attributes against controlled vocabularies, and ensemble prompting strategies that require consensus among multiple extraction attempts. These measures can reduce the likelihood of spurious metadata and improve semantic fidelity in operational environments.

Future Work: The success of Two-Step RAG in the NCM domain opens avenues for further exploration. Developing complementary safeguards for metadata extraction, such as automated consistency checks, controlled vocabularies, or ensemble prompting, should be prioritised to counteract hallucination risks. Once such mechanisms are in place, applying the approach to less structured datasets, including legal taxonomies, healthcare coding systems, or open-domain collections, would provide a rigorous test of robustness beyond highly standardised contexts. In addition, direct comparisons with alternative metadata-enhanced strategies would further substantiate the empirical claims and situate Two-Step RAG within the broader methodological landscape.

Acknowledgments

This study is partially funded by the Brazilian National Council for Scientific and Technological Development.

References

[1] M. Poliakov and N. Shvai, "Multi-meta-rag: Improving RAG for multi-hop queries using database filtering with LLM-extracted metadata," in *Proc. Int. Conf. on Information and Communication Technologies in Education, Research, and Industrial Applications*, Springer, 2024, pp. 334–342. DOI 10.1007/978303181372625.

[2] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park, "Database-Augmented Query Representation for Information Retrieval," *ArXiv*, 2024. DOI: 10.48550/arXiv.2406.16013.

[3] H. Wang, T. Zhao, and J. Gao, "BlendFilter: Advancing Retrieval-Augmented Large Language Models via Query Generation Blending and Knowledge Filtering," *ArXiv*, 2024. DOI: 10.48550/arXiv.2402.11129.

[4] L. Mombaerts, T. Ding, A. Banerjee, F. Felice, J. Taws, and T. Borogovac, "Meta Knowledge for Retrieval Augmented Large Language Models," *ArXiv*, 2024. DOI: 10.48550/arXiv.2408.09017.

[5] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Routledge, 1988. DOI: 10.4324/9780203771587.

[6] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020. DOI 10.48550/arXiv.2005.11401.

[7] H. Wang, F. Liu, Y. Dong, and D. Yu, "Entropy of eye movement during rapid automatized naming," *Frontiers in Human Neuroscience*, vol. 16, 2022. DOI: 10.3389/fnhum.2022.945406.

[8] C.-H. Liu, C.-W. Chang, J. Hung, J. J. H. Lin, P. Sung, L.-A. Lee, C.-T. Hsiao, Y.-P. Chao, E. S. Huang, and S.-L. Wang, "Brain computed tomography reading of stroke patients by resident doctors from different medical specialities: An eye-tracking study," *Journal of Clinical Neuroscience*, vol. 117, pp. 173–180, 2023. DOI: 10.1016/j.jocn.2023.10.004.

[9] J. Rystrøm, H. R. Kirk, and S. Hale, "Multilingual $\neq$ multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs," *arXiv preprint arXiv:2502.16534*, 2025. DOI: 10.48550/arXiv.2502.16534.

[10] J. P. Verma and P. Verma, "Determining Sample Size in General Linear Models," in *Statistics and Research Methods in Psychology with Excel*, Springer, 2020, pp. 89–119. DOI 10.1007/978-981-13-3429-0.

[11] V. Di Oliveira, Y. Bezerra, L. Weigang, P. Brom, and V. Celestino, "SLIM-RAFT: A Novel Fine-Tuning Approach to Improve Cross-Linguistic Performance for Mercosur Common Nomenclature," in *Proc. 20th Int. Conf. on Web Information Systems and Technologies (WEBIST)*, SciTePress, 2024, pp. 234–241. DOI: 10.5220/0012943400003825.

[12] V. Di Oliveira, L. Weigang, and G. P. R. Filho, "ELEVEN DataSet: A Labeled Set of Descriptions of Goods Captured from Brazilian Electronic Invoices," in *Proc. 18th Int. Conf. on Web Information Systems and Technologies (WEBIST)*, SciTePress, 2022, pp. 257–264. DOI: 10.5220/0011524800003318.

[13] Edge, Darren, et al., ”From local to global: A graph rag approach to query-focused summarization.” in *arXiv preprint arXiv:2404.16130*, 2024, DOI: 10.48550/arXiv.2404.16130..

[14] O. Khattab and M. Zaharia, “ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT,” in *Proc. 43rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR ’20)*, ACM, 2020, pp. 39–48. DOI: 10.1145/3397271.3401075.

[15] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, “BEIR: A Heterogeneous Benchmark for Zero-Shot Evaluation of Information Retrieval Models,” in *Proc. 35th Conf. on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track*, 2021. arXiv:2104.08663. DOI: 10.48550/arXiv.2104.08663

[16] K. Lin, K. Lo, J. E. Gonzalez, and D. Klein, “Decomposing Complex Queries for Tip-of-the-Tongue Retrieval,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Association for Computational Linguistics, 2023, pp. 5521–5533. DOI: 10.48550/arXiv.2305.15053

Vinícius Di Oliveira received a B.Sc. in Civil Engineering from the Federal University of Goiás (1997), an M.Sc. in Applied Computing from the University of Brasília (UnB), and is currently a Ph.D. candidate in Computer Science at UnB. He is a Tax Auditor at the Secretariat of Economy of the Federal District, focusing on Data Science in tax compliance and the development of Large Language Models for Portuguese.

Pedro Carvalho Brom holds degrees in Mathematics and Statistics and an M.Sc. in Statistics from the University of Brasília and is currently a Ph.D. candidate in Computer Science at UnB. He is a professor at the Federal Institute of Brasília, working in applied mathematics, statistical modelling, and R programming. His current projects include AI-based trajectory management, financial modelling, and intelligent data processing systems.

Li Weigang holds a D.Sc. from the Aeronautics Institute of Technology (ITA), Brazil (1994), and was a postdoctoral fellow at the University of Calgary, Canada (2001–2002). He is Full Professor and vice head of the Department of Computer Science at UnB, a CNPq Researcher, Boeing Inventor, and IEEE Senior Member. His work spans AI, Machine Learning, NLP, and computational modelling for air traffic management, where he introduced the “Once-Learning” approach.