https://doi.org/10.64142/jeai.2.2.54

Journal of Engineering and Artificial Intelligence

ISSN: 3117-3357

Research Article

JEAI-26-54

Does Linguistic Register Affect Faithfulness in Spanish Cache-Augmented Generation? An Exploratory Study with Open-Source LLMs

Angel Manuel Garcia-Carmona*

Independent Researcher, Spain

***Corresponding author:** Angel Manuel Garcia-Carmona, Independent Researcher, Spain, E-mail: am-gc@live.com

Received date: Sep 09, 2026; **Accepted date:** Sep 25, 2026; **Published date:** Oct 09, 2026

Citation: Garcia-Carmona AMG (2026) Does Linguistic Register Affect Faithfulness in Spanish Cache-Augmented Generation? An Exploratory Study with Open-Source LLMs. J Eng Artif Intell Vol.2 No.2: 54.

Abstract

Cache-Augmented Generation (CAG) provides an alternative to retrieval-based knowledge augmentation by preloading external knowledge into a reusable Key-Value (KV) cache. However, the extent to which CAG remains stable when semantically equivalent queries are expressed through different linguistic registers remains insufficiently understood, particularly in Spanish. This exploratory study investigates whether formal-colloquial register variation affects the faithfulness of CAG responses. A controlled paired design was constructed from 10 Spanish concepts and two complementary tasks definition generation and concept identification yielding 20 semantic cases and 40 formal-colloquial queries per model. Three open-source instruction-tuned LLMs (Qwen2.5-1.5B-Instruct, SmolLM2-1.7B-Instruct and Ministral-3B-Instruct) were evaluated under deterministic generation conditions using a fixed preloaded knowledge context. Faithfulness was assessed with RAGAS using granite 3.2 as the LLM judge. Of 120 planned generations, 113 produced valid faithfulness evaluations, yielding 53 complete formal-colloquial pairs. Neither paired parametric nor non-parametric analyses identified statistically significant register effects for any individual model and the between-model analysis did not detect systematic differences in register sensitivity. Descriptively, however, SmolLM2 tended to favor formal formulations, Ministral showed the opposite tendency and Qwen2.5 exhibited the most stable profile, combining comparatively high faithfulness, limited formal-colloquial separation and relatively low within-condition variability. These findings provide initial evidence that formal-colloquial variation does not produce a detectable systematic effect on CAG faithfulness under the conditions examined, while suggesting that descriptive sensitivity to linguistic register may depend on the underlying model and task configuration.

Keyword: Cache-augmented generation; CAG; Large language models; Linguistic register; Faithfulness; Spanish; Linguistic robustness; RAGAS

Introduction

Large Language Models (LLMs) increasingly operate in settings where users can express semantically equivalent requests through substantially different linguistic formulations. Such variation may involve lexical choice, syntactic structure, degree of formality or conversational style while preserving the underlying communicative intent. Robustness to this variation is therefore relevant to reliable LLM-based systems: When the information available to a model and the intended meaning of a query remain unchanged, differences in linguistic realization should not, in principle, lead to substantial changes in the extent to which generated responses remain grounded in the provided information.

This issue becomes particularly relevant when linguistic variation occurs within a single language. Spanish offers a suitable setting for examining this phenomenon because semantically equivalent requests can be expressed across different degrees of formality and conversationality without necessarily altering their underlying intent. A formal query may rely on more standardized lexical and syntactic choices, whereas a colloquial formulation may incorporate more conversational constructions and everyday expressions.

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

https://doi.org/10.64142/jeai.2.2.54
Journal of Engineering and Artificial Intelligence
ISSN: 3117-3357
Research Article

Evaluating whether such register variation affects model behaviour allows linguistic form to be examined separately from the semantic content of the request.

Cache-Augmented Generation (CAG) provides a particularly suitable setting for investigating this question. While Retrieval-Augmented Generation (RAG) has become a widely explored paradigm for grounding LLM responses in external knowledge, CAG offers an alternative approach in which relevant knowledge can be preloaded into the model's context and the resulting Key-Value (KV) cache reused across subsequent queries. Beyond its interest as an alternative to retrieval-based augmentation, this architecture provides a useful controlled setting for studying linguistic variation: The available knowledge and generation configuration can remain fixed while only the formulation of the query changes. Consequently, differences in the faithfulness of generated responses can be examined without introducing variability from a retrieval stage.

Within this setting, faithfulness constitutes a particularly relevant dimension of evaluation. A faithful response should remain supported by the knowledge made available to the model rather than introduce claims that cannot be attributed to that context. Although context-grounded generation has received considerable attention in LLM evaluation, the extent to which intralinguistic register variation affects faithfulness in a CAG setting remains insufficiently understood, particularly for Spanish. Examining formal and colloquial formulations of semantically equivalent queries therefore provides a controlled means of assessing whether changes in linguistic realization are associated with changes in the model's ability to remain grounded in the same preloaded knowledge.

This study investigates whether linguistic register affects the faithfulness of CAG responses when semantically equivalent queries are expressed in formal and colloquial Spanish. To isolate this factor, the experimental design keeps the underlying knowledge, semantic intent and generation configuration fixed while varying only the register of each query. Three open-source instruction-tuned LLMs are evaluated across two complementary task types: Definition generation and concept identification. Accordingly, the study addresses the following research question: Does linguistic register affect the faithfulness of Cache-Augmented Generation (CAG) when semantically equivalent queries are expressed in formal and colloquial Spanish? By examining this question in a controlled CAG setting, this work provides an exploratory evaluation of faithfulness robustness to intralinguistic variation in Spanish and investigates whether the observed patterns remain consistent across different language models and task formulations.

Related Work

Retrieval-Augmented Generation (RAG) has become a widely adopted approach for grounding Large Language Model (LLM) outputs in external knowledge, improving factuality and contextual relevance while enabling access to information beyond the model's parametric knowledge. By decoupling external knowledge from model parameters, RAG also allows knowledge sources to be updated without requiring model retraining. However, its effectiveness remains dependent on the retrieval stage: Variations in query formulation, vocabulary, ambiguity or domain-specific terminology may affect which evidence is retrieved and, consequently, the context available to the generator. This dependency raises an important robustness question when semantically equivalent information needs are expressed through different linguistic formulations, particularly in languages and varieties that have received less attention in LLM evaluation [1].

Beyond architectural considerations, the formulation of natural-language prompts represents an additional source of variability in LLM behaviour. Previous research has noted that relatively small changes in wording, sentence construction and contextual cues may alter model outputs. This issue is particularly relevant in non-English settings, where prompting behaviour and its relationship with model performance remain comparatively less explored. Carrasco-Sáez et al. examined prompting strategies used by higher-education students interacting with ChatGPT in Spanish and found that contextualization-related techniques were associated with response quality, providing empirical evidence that prompt construction is relevant to LLM performance in Spanish. Their study also highlights the comparatively limited body of research addressing prompting in Spanish and calls for further investigation in non-English linguistic contexts. However, their analysis focuses on prompt-engineering strategies and educational interactions rather than controlled intralinguistic variation between semantically equivalent formulations [2].

Prompt formulation itself has also been shown to constitute a relevant source of variation in Spanish-language LLM interactions. Viveros-Muñoz et al. examined structural properties of Spanish prompts, including utterance length, verbal mood and sentence complexity and found significant associations between these linguistic characteristics and the quality of generated responses. In particular, prompts combining different sentence structures received more favorable response evaluations, while prompt length was closely associated with grammatical complexity. The authors further suggest that longer prompts may benefit generation by

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

providing more specific instructions. These findings indicate that apparently surface-level characteristics of prompt formulation can be associated with measurable differences in LLM behaviour, motivating more controlled analyses in which semantic content is preserved while linguistic realization is systematically varied [3].

These limitations have motivated alternative approaches to knowledge-grounded generation that reduce or eliminate the dependency on real-time retrieval. Cache-Augmented Generation (CAG) has recently been proposed as one such paradigm, leveraging the extended context capabilities of modern LLMs to preload relevant knowledge and precompute its Key-Value (KV) cache before inference. Unlike conventional RAG pipelines, in which relevant information is dynamically retrieved for each query, CAG allows subsequent queries to be processed against the same preloaded knowledge without an additional retrieval stage. This architecture can reduce retrieval latency and errors while simplifying the overall generation pipeline, particularly when the underlying knowledge base is sufficiently constrained to fit within the model's context window [4].

Precomputed Key-Value (KV) caches provide an important mechanism for reducing repeated computation over previously processed context. Approaches such as TurboRAG have exploited precomputed KV caches to accelerate retrieval-augmented generation by avoiding redundant computation over retrieved content [5]. CAG builds on the broader principle of cache reuse in a retrieval-free setting: Relevant external knowledge is preloaded and encoded into a KV cache before inference, allowing the same cached context to be reused across subsequent queries. Consequently, the computational cost of processing the knowledge is incurred only once, while each query is processed conditionally on the precomputed cache without requiring an additional retrieval stage [4].

The broader literature on KV-cache reuse also indicates that caching is not merely a computational optimization, as the cached representations are conditioned by the context in which they were originally computed. Cache-Craft, for example, investigates the reuse of precomputed KV caches for knowledge chunks in RAG systems and shows that naive reuse across different contextual configurations can degrade generation quality. To address this issue, the approach selectively recomputes context-sensitive representations while preserving reusable portions of the cache. These findings highlight the interaction between cache reuse, contextualization and generation quality, extending the evaluation of caching mechanisms beyond computational efficiency alone.

Research on cache reuse has therefore considered not only computational efficiency but also the preservation of generation quality under different caching configurations. Cache-Craft demonstrates that precomputed KV representations cannot necessarily be reused independently of their surrounding context, since changes in chunk composition and ordering may alter their contextualization and affect the resulting output. The proposed system consequently combines cache reuse with selective re-computation to balance computational savings against generation quality [6].

Recent extensions of CAG have further shown that the information preserved in KV representations can influence downstream generation quality. CACHENOTES introduces a task-aware compression strategy in which a corpus is distilled into a reusable, query-agnostic KV cache, allowing subsequent queries to be answered without retrieval or reprocessing of the original corpus. Its experiments show that downstream performance depends on which contextual information is preserved during compression, with task-misaligned compression producing substantial performance degradation on some tasks. However, these evaluations primarily characterize downstream answer quality through task-specific metrics such as F1, ROUGE-L, exact or string matching, rather than explicitly measuring whether the claims generated by the model remain supported by the cached knowledge [7].

Beyond task-specific measures of answer quality, faithfulness provides a distinct perspective on knowledge-grounded generation by assessing whether the information contained in a generated response is supported by the evidence provided to the model. Rashkin et al. distinguish this property from other dimensions of response quality, such as relevance and fluency and argue that reference-based metrics alone do not directly capture whether a generation remains grounded in its supporting evidence. Their evaluation therefore considers faithfulness explicitly as the extent to which a response is fully supported by the available evidence, showing that improvements in groundedness may occur independently of other quality dimensions. This distinction is particularly relevant to knowledge-augmented generation, where a response may be appropriate or factually correct while still introducing information that is not supported by the supplied context [8].

More recent evaluation frameworks have operationalized faithfulness as an automated measure of the relationship between generated claims and their supporting context. RAGAS was introduced as a reference-free framework for evaluating knowledge-augmented generation along distinct quality dimensions, including faithfulness, answer relevance

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

https://doi.org/10.64142/jeai.2.2.54
Journal of Engineering and Artificial Intelligence
ISSN: 3117-3357
Research Article

and context relevance. In particular, faithfulness is defined in terms of whether the claims contained in a generated answer can be inferred from the provided context. The metric first decomposes the answer into individual statements and subsequently verifies each statement against the context, computing the final score as the proportion of supported statements. This formulation explicitly evaluates contextual grounding rather than similarity to a reference answer. In the original WikiEval experiments, the RAGAS faithfulness metric showed high agreement with human judgments, supporting its use as an automated proxy for assessing whether generated information remains grounded in the supplied knowledge [9].

Importantly, contextual support should not be interpreted as evidence of causal reliance on the supplied knowledge. Wallat et al. distinguish between whether a generated statement is supported by an attributed document and the stronger notion of citation faithfulness, which additionally requires that the document causally influence the generation of the corresponding claim. Their analysis shows that an attribution may be correct while still being post-rationalized, for instance when a model generates an answer from parametric memory and subsequently associates it with supporting contextual evidence. Accordingly, the notion of faithfulness adopted in the present study concerns the consistency of generated claims with the knowledge supplied in context, as operationalized by RAGAS, rather than the causal provenance of those claims within the model's generation process. This distinction delimits the scope of the present investigation: The objective is to determine whether formal-colloquial register variation affects the degree to which generated responses remain supported by a fixed preloaded knowledge context, not to establish whether that context causally determines the model's internal generation process [10].

Methodology

Experimental design

This study adopts a controlled paired experimental design to investigate whether linguistic register affects the faithfulness of responses generated through Cache-Augmented Generation (CAG).

The independent variable is linguistic register, operationalized through two conditions: Formal and colloquial Spanish. Faithfulness constitutes the dependent variable and measures the extent to which the claims contained in each generated response are supported by the knowledge supplied to the model.

For each semantic case, two query formulations were constructed to express the same underlying information needed while differing in linguistic register. Each formal query was therefore paired with a semantically equivalent colloquial formulation. This paired design enables within-case comparison while controlling for differences in semantic content. The underlying knowledge base, system instructions, generation configuration and evaluation procedure were held constant across both conditions, such that the linguistic register constituted the experimental factor intentionally varied between paired queries.

The experiment comprised 20 semantic cases derived from 10 concepts and two complementary task types: Definition generation and concept identification. Each semantic case was represented by one formal and one colloquial query, resulting in 40 queries per model. The same experimental procedure was independently applied to three open-source instruction-tuned LLMs, yielding 120 planned generations. This multi-model design was intended to examine both the overall effect of registers and whether descriptive patterns were consistent across different model architectures.

Figure 1 summarizes the complete experimental workflow. A fixed knowledge context was preloaded and represented through a reusable KV cache before query-specific generation. Formal and colloquial versions of each semantic case were subsequently processed under the same generation configuration. The resulting paired responses were evaluated using RAGAS faithfulness, producing a score between 0 and 1 according to the proportion of generated claims supported by the supplied context. These scores constitute the observations used in the subsequent paired statistical analysis.

Knowledge base and query construction

The experimental knowledge base was constructed from ten Spanish concepts related to emotional states and feelings. The corresponding definitions were manually identified in the Diccionario De La Lengua Española (DLE) of the Real Academia Española (RAE) and transcribed into the experimental knowledge base. The knowledge content was therefore obtained directly from the lexical source rather than synthetically generated or paraphrased by an LLM. These entries constituted the fixed external knowledge supplied to the models throughout the experiment and remained unchanged across models, tasks and register conditions.

Two complementary task types were defined for each concept. In the definition task, the target concept was provided and the model was required to generate its meaning from the supplied knowledge. In the concept-identification task, the direction was reversed: The query expressed the

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

https://doi.org/10.64142/jeai.2.2.54
Journal of Engineering and Artificial Intelligence
ISSN: 3117-3357
Research Article

relevant meaning or characteristics and required the model to identify the corresponding concept. The combination of ten concepts and two task types resulted in 20 semantic cases.

For each semantic case, formal and colloquial Spanish query formulations were generated using GPT-5 through the ChatGPT interface. The model was instructed to preserve the underlying semantic intent and requested information while modifying the linguistic register of the query. Formal formulations favored comparatively standardized and explicit language, whereas colloquial formulations employed more conversational and everyday expressions. All generated query pairs were subsequently reviewed manually to verify that the intended semantic content and requested information were preserved across both formulations. The register manipulation therefore concerned the linguistic realization of the query rather than its underlying task or target information.

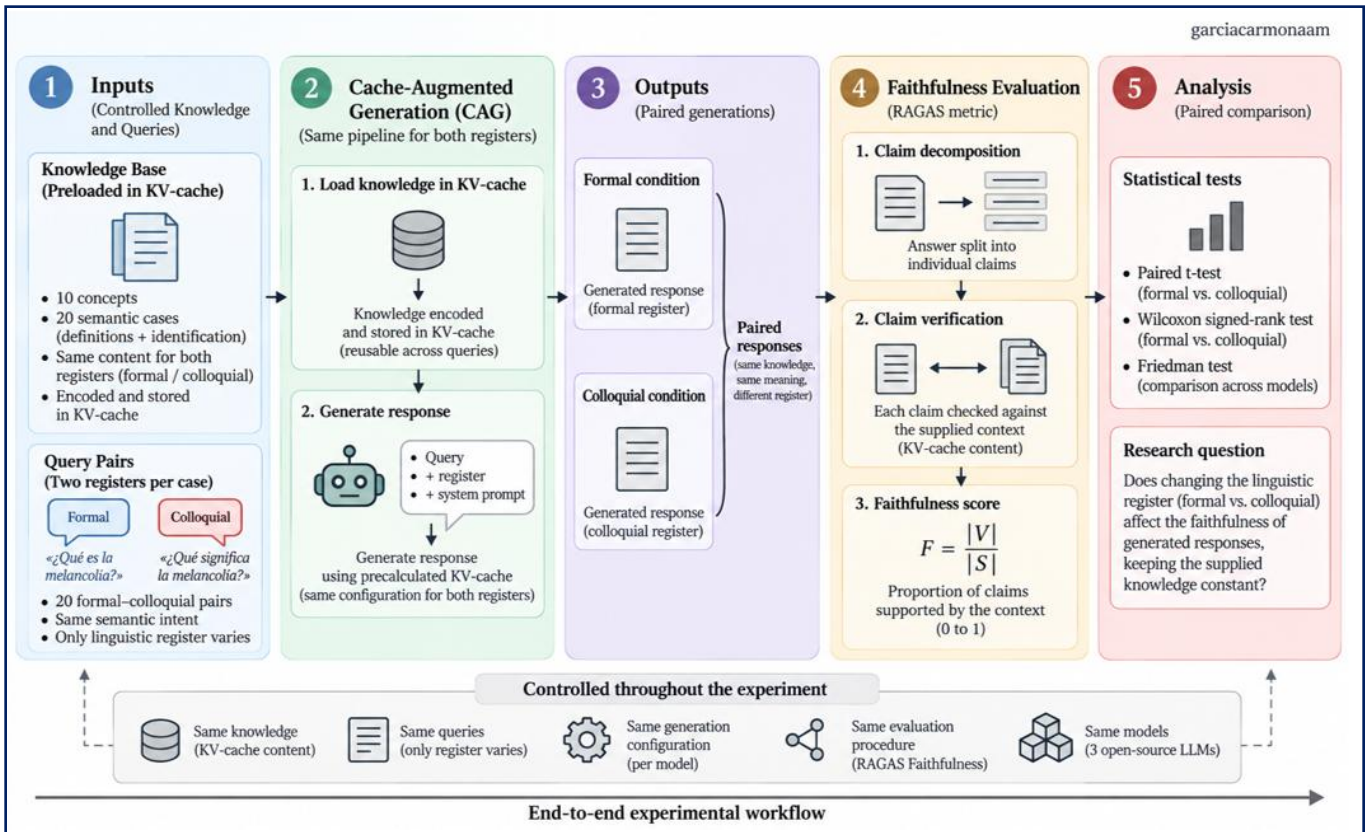

Figure 1: Overview of the experimental workflow used to evaluate the effect of linguistic register on faithfulness in Cache Augmented Generation (CAG). Formal and colloquial queries expressing the same semantic intent are processed against a fixed preloaded knowledge context under identical generation condition. Generated responses are evaluated using RAGAS faithfulness and subsequently compared through paired statistical analysis.

The resulting experimental dataset comprised 40 queries organized into 20 formal-colloquial pairs: Ten pairs for definition generation and ten for concept identification. Within each pair, the target concept, semantic intent, expected information and available knowledge were held constant, while linguistic register constituted the manipulated factor.

Cache-augmented generation pipeline

The generation pipeline followed a Cache-Augmented Generation (CAG) architecture in which the complete experimental knowledge base was incorporated into a fixed context before query processing. The entries stored in the structured knowledge base were first converted into a textual context shared across all experimental conditions. This context was preceded by a fixed instruction requiring the model to answer briefly and exclusively on the basis of the supplied information.

Before processing individual queries, the fixed prefix and knowledge context were tokenized and passed through the model to compute their Key-Value (KV) representations. The resulting KV cache constituted a reusable inference state containing the preloaded knowledge. Rather than recomputing this fixed context for every query, the base cache was preserved and independently reused for subsequent

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

https://doi.org/10.64142/jeai.2.2.54
Journal of Engineering and Artificial Intelligence
ISSN: 3117-3357
Research Article

generations. For each experimental query, a copy of the same base KV cache was created, after which the query-specific tokens were appended and generation proceeded conditionally on the cached context.

In operational terms, the reusable KV cache encoded the fixed instructional prefix together with the complete knowledge context. For each semantic case, this cached state remained unchanged, while only the query formulation was appended during query-specific processing. Formal and colloquial variants of the same semantic case therefore differed in their linguistic formulation but were conditioned on the same precomputed knowledge representation.

This procedure ensured that formal and colloquial formulations of the same semantic case were processed against an identical precomputed knowledge state. No retrieval operation, document ranking or query-dependent context selection was performed during inference. Consequently, a change in query formulation could not modify which knowledge entries were made available to the generator, thereby removing retrieval variability as a potential confounding factor when comparing linguistic registers.

The fixed textual prefix used to construct the cached context was identical for all generations and instructed the models in Spanish to respond concisely using only the supplied context. Apart from the query-specific formulation, the contextual input and generation procedure remained unchanged across paired observations.

Language models and generation configuration

The experiment was conducted using three open-source instruction-tuned language models: Qwen2.5-1.5B-Instruct, SmolLM2-1.7B-Instruct and Ministral-3B-Instruct. The models were selected to provide a multi-model evaluation within a relatively compact parameter range while allowing the same CAG experimental procedure to be applied independently to different language models. Each model processed the complete set of 40 formal and colloquial queries, resulting in 40 planned generations per model and 120 planned generations overall.

Generation was performed deterministically by disabling sampling (do_sample=False). Stochastic token sampling was intentionally avoided because the objective of the experiment was not to characterize the diversity of responses that could be obtained from repeated random generations, but to isolate the potential effect of linguistic register under controlled generation conditions. If tokens had been sampled probabilistically from the model output distribution, differences between paired formal and colloquial responses could have reflected both register variation and random generation variability. Disabling sampling reduced this additional source of variation and made the paired comparison more directly attributable to the experimental manipulation. The maximum number of newly generated tokens was fixed at 200 (max_new_tokens=200) for all three models. Although responses were instructed to be concise, this value was deliberately set as a sufficiently permissive upper bound to accommodate definitions requiring more extensive explanation and to reduce the risk of truncating otherwise complete responses. The same generation configuration was maintained across all models and register conditions.

The same fixed Spanish instruction and preloaded knowledge context described in section 3.3 were used for all three models. No model-specific prompt engineering or query reformulation was performed. Each model independently constructed and reused its own KV cache from the same textual knowledge and instructional prefix, ensuring that the experimental manipulation remained restricted to the formal or colloquial formulation of each query.

Faithfulness evaluation

Generated responses were evaluated using the faithfulness metric implemented in RAGAS (version 0.3.9). Faithfulness was selected as the sole evaluation metric because the objective of the study was specifically to determine whether linguistic register affects the extent to which generated information remains supported by the fixed knowledge context. The evaluation therefore focused on contextual grounding rather than on general response quality, stylistic appropriateness or similarity to a reference answer.

RAGAS faithfulness evaluates a generated response by decomposing its content into individual claims and determining whether each claim can be supported by the supplied context. The resulting score represents the proportion of generated claims that are supported by the contextual information and ranges from 0 to 1, with higher values indicating greater contextual support. Consequently, the metric does not require a reference answer and directly evaluates the relationship between the generated response and the knowledge made available to the model.

Claim evaluation was performed using granite 3.2 (granite3.2:latest) as the LLM judge through Ollama. The model was selected in part to leverage its reasoning-oriented capabilities, including Chain-of-Thought (CoT)-style reasoning, with the aim of facilitating a more deliberate assessment of whether individual claims were supported by the supplied context rather than relying primarily on superficial

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

https://doi.org/10.64142/jeai.2.2.54
Journal of Engineering and Artificial Intelligence
ISSN: 3117-3357
Research Article

lexical correspondence. The judge was configured with a temperature of 0 to minimize stochastic variation during evaluation. The same judge model, evaluation configuration and knowledge context were used for every generated response, ensuring a consistent faithfulness assessment procedure across models, tasks and register conditions.

Faithfulness scores were successfully obtained for 113 of the 120 planned generations. Seven responses could not be assigned a valid faithfulness score because of technical failures during the automated evaluation process. These observations were treated as missing values and excluded from analyses requiring the corresponding score; they were neither assigned a score of zero nor otherwise imputed. Consequently, descriptive analyses were conducted using all available valid evaluations, whereas paired analyses included only semantic cases for which valid faithfulness scores were available for both the formal and colloquial conditions.

Statistical analysis

The primary statistical analysis followed the paired structure of the experimental design. For each model and semantic case, the effect of linguistic register was operationalized as the difference between the faithfulness score obtained for the formal formulation and that obtained for the corresponding colloquial formulation. Positive differences therefore indicated higher faithfulness under the formal condition, whereas negative differences indicated higher faithfulness under the colloquial condition.

Only complete formal-colloquial pairs were included in the inferential analyses; missing faithfulness scores were not imputed. This resulted in 53 complete pairs from the 60 originally planned across the three models.

The distribution of paired differences was assessed separately for each model using the Shapiro Wilk test. Two-sided paired t-tests were subsequently used to test whether the mean formal colloquial difference differed from zero, with 95% confidence intervals calculated for the corresponding mean differences.

Given the bounded and discrete nature of faithfulness scores and the relatively small number of complete pairs, two-sided Wilcoxon signed-rank tests were additionally performed as a non-parametric sensitivity analysis **(Table 1)**.

Table 1: Paired analysis of formal-colloquial faithfulness differences by model.

Model	Complete pairs	Mean D	Median D	95% CI for Mean D	t (df)	p (t-test)	p (Wilcoxon)
SmolLM2-1.7B-Instruct	18	0.19	0.133	(−0.044, 0.424)	1.716 (17)	0.104	0.133
Qwen2.5-1.5B-Instruct	19	0.052	0	(−0.106, 0.210)	0.691 (18)	0.498	0.459
Ministral-3B-Instruct	16	−0.107	0	(−0.379, 0.165)	−0.837 (15)	0.416	0.329

To examine whether register sensitivity differed systematically across models, paired differences were additionally compared at the model level. Homogeneity of variance was explored using the median-centered Levene test. Because the same semantic cases were evaluated across the three models, an ordinary independent-samples analysis of variance was not considered appropriate. Instead, the Friedman test was applied to semantic cases with complete paired differences for all three models, preserving the repeated-measures structure of the data.

Statistical significance was primarily assessed at $\alpha = 0.05$. Given the exploratory nature and limited sample size of the study, $\alpha = 0.10$ was additionally considered as a less conservative sensitivity threshold for detecting potentially weak register effects.

Following the inferential analysis, descriptive statistics were used to examine the structure of the observed patterns in greater detail. Faithfulness distributions were characterized using the mean, median, standard deviation and interquartile range and were examined by model, linguistic register and task type. Directional counts of paired differences were also inspected to distinguish cases in which formal queries produced higher Faithfulness, colloquial queries produced higher Faithfulness or both formulations obtained the same score. These descriptive analyses were used to characterize model- and task-specific patterns rather than as additional tests of statistical significance.

Results

Inferential analysis of linguistic register

The primary inferential analysis examined whether faithfulness differed systematically between formal and colloquial formulations of the same semantic cases. Of the 60 formal-colloquial pairs originally planned across the three models, 53 contained valid faithfulness scores for both register conditions and were therefore included in the paired analyses. The Shapiro-Wilk tests applied to the paired differences did not indicate statistically significant departures from normality

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

https://doi.org/10.64142/jeai.2.2.54
Journal of Engineering and Artificial Intelligence
ISSN: 3117-3357
Research Article

for SmolLM2 (W = 0.938, p = 0.265), Qwen2.5 (W = 0.953, p = 0.436) or Ministral (W = 0.950, p = 0.494). These results provide no statistical evidence against the normality assumption for the paired differences, although they should be interpreted cautiously given the relatively small sample sizes and the discrete nature of the faithfulness scores. While the Central Limit Theorem provides a theoretical basis for the increasing normality of sampling distributions as sample size grows [11], the number of complete pairs available for each model is limited. For this reason, the parametric analysis was complemented with a non-parametric signed-rank test rather than relying exclusively on asymptotic arguments.

The paired comparisons did not identify a statistically significant formal-colloquial difference in faithfulness for any of the three models (**Table 1**). SmolLM2 exhibited the largest positive mean difference (D = 0.190), indicating descriptively higher faithfulness for formal queries, but the paired t-test did not reach statistical significance ($t(17) = 1.716$, $p = 0.104$; 95% CI [−0.044, 0.424]). Qwen2.5 showed a smaller positive difference (D = 0.052), likewise without statistical significance ($t(18) = 0.691$, $p = 0.498$; 95% CI [−0.106, 0.210]). Ministral showed a difference in the opposite direction (D = −0.107), corresponding to descriptively higher faithfulness for colloquial queries, although this difference was also not statistically significant ($t(15) = -0.837$, $p = 0.416$; 95% CI [−0.379, 0.165]). All confidence intervals included zero and none of the paired comparisons reached either the conventional $\alpha = 0.05$ threshold or the exploratory $\alpha = 0.10$ threshold.

The Wilcoxon signed-rank tests yielded conclusions consistent with the parametric analysis. No statistically significant formal-colloquial differences were detected for SmolLM2 (W = 24, p = 0.133), Qwen2.5 (W = 47, p = 0.459) or Ministral (W = 18, p = 0.329). Thus, the absence of statistically significant register effects at the individual-model level was consistent across both the parametric and non-parametric paired analyses.

A Friedman test was subsequently conducted to examine whether sensitivity to linguistic register differed systematically across the three models. This analysis was restricted to the 14 semantic cases for which complete paired differences were available for SmolLM2, Qwen2.5 and Ministral, thereby preserving the repeated-measures structure across models. The test did not reveal a statistically significant difference in register sensitivity among the three models ($\chi^2(2) = 2.44$, $p = 0.295$). Consequently, the null hypothesis of equal model-level distributions of paired register differences could not be rejected and no post-hoc pairwise comparisons were conducted.

Thus, although the direction and magnitude of the observed formal-colloquial differences varied descriptively across models, the inferential analysis did not provide sufficient evidence that these differences represented systematic between-model variation in register sensitivity.

Descriptive analysis of model-specific patterns

Despite the absence of statistically significant register effects, inspection of the descriptive distributions revealed patterns that warranted further examination. Across the 113 valid faithfulness evaluations, the overall mean score was 0.612 and the median was 0.667. As shown in **Figure 2**, the distribution exhibited a pronounced concentration at the boundaries of the scale, particularly at its upper limit. Specifically, 24 observations (21.2%) obtained a faithfulness score of 0, while 45 (39.8%) reached the maximum score of 1. The remaining 44 observations (38.9%) were distributed across intermediate values. Thus, 61.0% of all valid evaluations were located at either boundary of the scale. This distributional structure indicates that measures of central tendency alone provide an incomplete characterization of the observed faithfulness scores and motivates a more detailed examination of register-, model- and task-specific patterns.

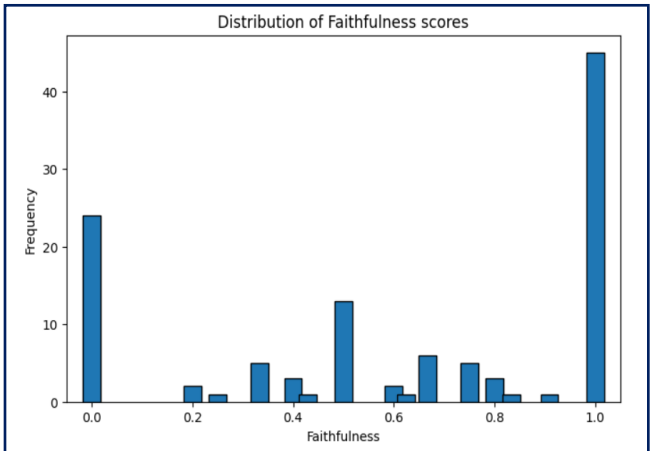

Figure 2: Frequency distribution of the 113 valid faithfulness scores. Bars represent the frequency of each observed faithfulness value. A pronounced concentration is observed at the boundaries of the scale, particularly at the maximum score of 1.

When observations were aggregated by linguistic register, formal queries showed a slightly higher mean faithfulness (0.628) than colloquial queries (0.596), with corresponding median scores of 0.750 and 0.667. This aggregate difference was small and should be interpreted descriptively, particularly because pooling observations across models conceals heterogeneous directional patterns. As examined in greater

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

https://doi.org/10.64142/jeai.2.2.54
Journal of Engineering and Artificial Intelligence
ISSN: 3117-3357
Research Article

detail below, SmolLM2 and Qwen2.5 displayed descriptively higher faithfulness under the formal condition, whereas Ministral showed the opposite tendency, with higher faithfulness under the colloquial condition. Consequently, the overall register-level difference partially reflects the combination of opposing model-specific patterns rather than a uniform tendency across the three models **(Figure 3)**.

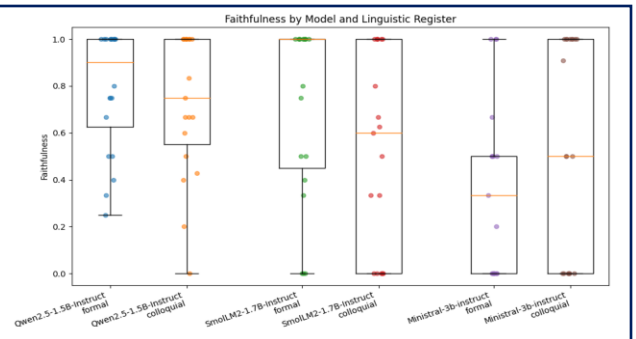

Figure 3: Faithfulness scores by model and linguistic register. Boxplots show the distribution of faithfulness scores for formal and colloquial queries across SmolLM2, Qwen2.5, and Ministral. Individual observations are overlaid to illustrate the dispersion and discrete structure of the scores.

SmolLM2 exhibited the largest descriptive tendency in favor of formal formulations. Mean faithfulness increased from 0.519 under the colloquial condition to 0.699 under the formal condition, while the corresponding medians were 0.600 and 1.000.

Among the 18 complete formal-colloquial pairs, the formal formulation produced a higher faithfulness score in 10 cases, both formulations obtained the same score in 5 cases and the colloquial formulation produced a higher score in 3 cases. The resulting mean paired difference was 0.190. Thus, although the inferential tests did not establish a statistically significant register effect for SmolLM2, its descriptive distribution showed a comparatively pronounced tendency toward higher faithfulness for formal queries.

Qwen2.5 produced the highest absolute faithfulness scores under both register conditions, with mean scores of 0.722 for colloquial queries and 0.785 for formal queries and corresponding medians of 0.750 and 0.900.

Its mean paired difference was comparatively small (0.052) and the directional counts were more balanced: Formal queries produced higher faithfulness in 8 of the 19 complete pairs, colloquial queries in 7 and the remaining 4 pairs resulted in equal scores. Among the three evaluated models, Qwen2.5 therefore exhibited the greatest descriptive stability across the two register conditions, while maintaining comparatively high faithfulness scores overall.

Ministral displayed a different descriptive pattern. Mean faithfulness was 0.548 for colloquial queries and 0.365 for formal queries, with corresponding medians of 0.500 and 0.333. Among the 16 complete pairs, colloquial formulations produced higher faithfulness in 6 cases, formal formulations in 4 and 6 pairs obtained equal scores. The mean paired difference was negative (−0.107), indicating a descriptive tendency in the opposite direction to that observed for SmolLM2 and Qwen2.5. Ministral also showed substantial dispersion, particularly under the colloquial condition, suggesting that its responses were comparatively variable across semantic cases.

Further disaggregation by task type showed that the descriptive register patterns were not equally pronounced across tasks. For SmolLM2, mean faithfulness in the concept-identification task was 0.469 for colloquial queries and 0.537 for formal queries, corresponding to a formal-colloquial difference of +0.068.

The difference was considerably larger for definition generation, where mean faithfulness increased from 0.563 under the colloquial condition to 0.845 under the formal condition (+0.282). Thus, the overall descriptive tendency toward formal formulations observed for SmolLM2 was particularly concentrated in the definition task.

Qwen2.5 displayed comparatively small differences in both task types. In concept identification, mean faithfulness was 0.665 for colloquial queries and 0.750 for formal queries (+0.086), while definition generation yielded means of 0.785 and 0.820, respectively (+0.035). This pattern is consistent with the comparatively small overall paired difference observed for Qwen2.5 and suggests that its descriptive performance remained relatively similar across registers in both task formulations.

Ministral showed the opposite directional pattern in both tasks. For concept identification, mean faithfulness decreased from 0.500 under the colloquial condition to 0.191 under the formal condition, producing the largest task-specific difference observed in absolute terms (−0.310).

Definition generation showed the same direction but a smaller difference, with mean scores of 0.591 for colloquial queries and 0.487 for formal queries (−0.104). The descriptive tendency toward colloquial formulations observed for Ministral was therefore present across both tasks but was most pronounced for concept identification **(Table 2)**.

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

https://doi.org/10.64142/jeai.2.2.54
Journal of Engineering and Artificial Intelligence
ISSN: 3117-3357
Research Article

Further disaggregation by task type revealed differences in both the central tendency and relative variability of faithfulness scores (**Table 2**).

Table 2: Descriptive faithfulness statistics by model, task type and linguistic register.

Model	Task type	Register	N	Mean	Median	CV
SmolLM2	Concept identification	Colloquial	9	0.4694	0.6	1.0012
		Formal	9	0.537	0.5	0.8747
	Definition	Colloquial	10	0.5633	0.5833	0.6959
		Formal	10	0.845	1	0.2711
Qwen 2.5	Concept identification	Colloquial	10	0.6645	0.7083	0.5437
		Formal	10	0.75	0.875	0.3989
	Definition	Colloquial	9	0.7852	0.8333	0.3003
		Formal	10	0.82	0.9	0.273
Ministral	Concept identification	Colloquial	9	0.5	0.5	0.866
		Formal	7	0.1905	0	1.2809
	Definition	Colloquial	10	0.5909	0.9545	0.8619
		Formal	10	0.4867	0.5	0.8707

For SmolLM2, concept identification showed relatively similar mean faithfulness across colloquial (0.469) and formal (0.537) formulations, accompanied by substantial relative variability in both conditions (CV = 1.001 and 0.875, respectively). A more pronounced descriptive pattern emerged for definition generation: Mean faithfulness increased from 0.563 for colloquial queries to 0.845 for formal queries, while the median increased from 0.583 to 1.000 and the CV decreased from 0.696 to 0.271. Thus, the descriptive tendency toward higher faithfulness for formal formulations in SmolLM2 was particularly apparent in the definition task and was accompanied by lower relative variability.

Qwen2.5 showed comparatively high faithfulness and low relative variability across both task types. In concept identification, mean faithfulness was 0.665 for colloquial queries and 0.750 for formal queries, with corresponding medians of 0.708 and 0.875 and CV values of 0.544 and 0.399. Definition generation exhibited an even smaller separation between register-specific means (0.785 and 0.820), accompanied by similar medians (0.833 and 0.900) and comparatively low CV values under both colloquial (0.300) and formal (0.273) conditions. Notably, Qwen2.5 combined the smallest descriptive separation between formal and colloquial faithfulness at the model level with comparatively low relative variability within the register conditions. Thus, its descriptive stability was reflected not only in limited differences between registers but also in comparatively consistent faithfulness scores within them.

Ministral exhibited a contrasting descriptive pattern, with higher faithfulness for colloquial formulations in both task types and comparatively high relative variability. The clearest separation occurred in concept identification, where mean faithfulness was 0.500 for colloquial queries but only 0.191 for formal queries, with corresponding medians of 0.500 and 0.000. Relative variability was also substantial (CV = 0.866 and 1.281, respectively), although the particularly high CV under the formal condition should be interpreted cautiously because of the low corresponding mean and the smaller number of valid observations (N = 7). In definition generation, the separation between register-specific means was smaller (0.591 for colloquial and 0.487 for formal), while the medians were 0.955 and 0.500, respectively. Unlike Qwen2.5, relative variability remained high and nearly identical across both register conditions (CV = 0.862 and 0.871). Overall, Ministral combined a descriptive tendency toward higher faithfulness for colloquial formulations with comparatively variable performance, particularly contrasting with the more stable pattern observed for Qwen2.5.

The coefficients of variation provided additional descriptive evidence of model- and task-specific differences in score dispersion. SmolLM2 showed a marked reduction in relative variability for formal definition queries (CV = 0.271) compared with their colloquial counterparts (CV = 0.696), alongside higher mean and median faithfulness. Qwen2.5 exhibited comparatively low relative variability, particularly in definition generation, where CV values were similar across colloquial (0.300) and formal (0.273) conditions. Ministral remained comparatively variable in definition generation under both registers (CV = 0.862 and 0.871). The particularly high CV observed for formal concept-identification queries in Ministral (1.281) should be interpreted cautiously because the corresponding mean faithfulness was close to zero (0.191). Overall, the CV patterns complement the central-tendency results by suggesting that register-related descriptive

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

https://doi.org/10.64142/jeai.2.2.54
Journal of Engineering and Artificial Intelligence
ISSN: 3117-3357
Research Article

differences concerned not only the level of faithfulness but, in some conditions, also its relative variability.

Taken together, the descriptive analyses revealed model- and task-specific patterns that were not apparent from the aggregate register-level results. SmolLM2 showed a descriptive tendency toward higher faithfulness for formal formulations, particularly in definition generation, whereas Ministral exhibited the opposite tendency, most clearly in concept identification. Qwen2.5 showed the smallest descriptive separation between registers and comparatively low relative variability, particularly in definition generation, resulting in the most stable descriptive profile across the evaluated conditions. These patterns complement, but do not alter, the inferential findings: No statistically significant formal-colloquial differences were detected within any individual model and no significant systematic differences in register sensitivity were identified across models.

Discussion

The results of this exploratory study do not provide sufficient evidence that linguistic register systematically affects the faithfulness of CAG responses when semantically equivalent queries are expressed in formal and colloquial Spanish. None of the three evaluated models exhibited a statistically significant formal-colloquial difference and the complementary parametric and non-parametric analyses led to consistent conclusions. Moreover, the between-model analysis did not identify statistically significant systematic differences in register sensitivity. These findings suggest that, within the experimental conditions examined, variation between formal and colloquial Spanish did not produce a detectable systematic change in the extent to which generated claims remained supported by the same preloaded knowledge context. However, the absence of statistical significance should not be interpreted as evidence of equivalence between registers, particularly given the exploratory design and the limited number of paired observations.

The absence of a statistically significant systematic effect nevertheless coexisted with distinct descriptive patterns across models. SmolLM2 tended to produce higher faithfulness for formal formulations, whereas Ministral showed the opposite tendency, with higher scores for colloquial formulations. Qwen2.5 occupied a different position: it exhibited the smallest mean paired difference between registers while also maintaining comparatively high faithfulness and relatively low score variability across the evaluated conditions. This combination is noteworthy because descriptive stability in Qwen2.5 was observable at two levels: formal and colloquial queries produced comparatively similar faithfulness scores, while the scores within several model-task-register conditions were also comparatively less variable. Thus, although the inferential analysis does not establish significant differences in register sensitivity among models, the descriptive results suggest that robustness to intralinguistic variation may depend partly on the underlying language model rather than solely on the CAG architecture itself.

These findings also have implications for the practical viability of CAG as a knowledge-grounded generation architecture. Previous work has presented CAG as an alternative to retrieval-based augmentation when the relevant knowledge can be accommodated within the model context, allowing a precomputed KV cache to be reused across queries and avoiding query-time retrieval. In such settings, the usefulness of CAG depends not only on computational efficiency or downstream task performance, but also on whether responses remain consistently grounded when users formulate equivalent information needs in different ways. The behavior observed for Qwen2.5 is particularly relevant in this respect: the model combined comparatively high faithfulness with limited descriptive separation between formal and colloquial queries and relatively low within-condition variability. This suggests that, for an appropriate underlying model, CAG can maintain a comparatively stable grounding profile despite intralinguistic variation in user queries.

This observation may complement previous reports of competitive performance and efficiency advantages for cache-based generation relative to retrieval-dependent pipelines under suitable conditions. If CAG can reduce retrieval-related computational overhead while preserving task performance, the descriptive stability observed for Qwen2.5 adds a potentially relevant linguistic dimension to its practical applicability. In a deployed system, users are unlikely to formulate semantically equivalent requests using a uniform register and an architecture whose grounding behavior remains comparatively stable across such variation would therefore be desirable. Nevertheless, the present experiment does not establish that CAG is more robust to linguistic registers than RAG, since no retrieval-based baseline was evaluated. A direct comparison using identical knowledge, semantic cases, models and register manipulations would be required to determine whether removing the retrieval stage actually provides an advantage in register robustness.

From a linguistic perspective, the results indicate that formal-colloquial variation should not necessarily be expected to affect different language models in the same direction. The descriptive tendencies observed for SmolLM2 and Ministral were particularly informative in this regard: Whereas SmolLM2 generally favored formal formulations, Ministral

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

tended to produce higher faithfulness for colloquial formulations. Qwen2.5, by contrast, showed comparatively limited separation between the two registers. This heterogeneity suggests that register variation cannot be characterized simply as a uniform advantage of either formal or colloquial Spanish. Instead, the interaction between linguistic realization and model-specific behavior appears to warrant greater attention, even when the semantic intent and available knowledge remain fixed.

This interpretation is broadly consistent with previous exploratory research showing that the linguistic formulation of Spanish prompts can be associated with differences in LLM outputs. Carrasco-Sáez et al., identified prompt characteristics such as contextualization as relevant to response quality in Spanish, while Viveros-Muñoz et al., reported differences associated with grammatical features including verbal mood, sentence complexity and writing standards. The present study extends this line of investigation in a different direction by isolating formal-colloquial register variation while holding semantic intent and external knowledge constant. Importantly, the findings do not indicate a general advantage for one register. Rather, the opposing descriptive tendencies across models suggest that linguistic robustness in Spanish may need to be evaluated at the level of specific model-task configurations rather than assumed from aggregate performance.

Task type further qualified these descriptive patterns. SmolLM2's tendency toward formal formulations was most pronounced in definition generation, whereas Ministral's tendency toward colloquial formulations was strongest in concept identification. Qwen2.5 maintained comparatively similar faithfulness across registers in both tasks. This suggests that register-related behavior may depend not only on the underlying model but also on the type of transformation required between the query and the supplied knowledge. Definition generation and concept identification reverse the direction of the information mapping and the different patterns observed across these tasks indicate that register sensitivity may emerge differently depending on how the model must use the same contextual knowledge. Given the limited number of cases, however, this interpretation should be regarded as a hypothesis generated by the descriptive results rather than as an independently established task-register interaction.

The interpretation of these findings also depends on the specific notion of faithfulness adopted in the study. RAGAS evaluates whether the claims contained in a generated response are supported by the context supplied to the model, thereby providing an observable measure of contextual grounding rather than a measure of general answer quality. In this sense, the comparatively stable behavior observed for Qwen2.5 indicates that its generated claims remained similarly supported by the fixed knowledge context across formal and colloquial formulations. Conversely, the larger descriptive variations observed for SmolLM2 and Ministral indicate changes in contextual support across register conditions, although these differences were not statistically significant.

However, contextual support should not be conflated with causal reliance on the cached knowledge. As emphasized by Wallat et al., in their distinction between supported attribution and citation faithfulness, a generated claim may be compatible with contextual evidence without that evidence necessarily being the causal source of the generation. Accordingly, the present results do not demonstrate that the models derived each supported claim from the preloaded KV cache rather than from their parametric knowledge. What they establish is the observable degree of support between generated claims and a fixed contextual knowledge source under controlled changes in linguistic register. This distinction is particularly important when interpreting the apparent stability of Qwen2.5: The results characterize stability in contextual faithfulness, not the internal causal mechanisms through which that stability was achieved.

Overall, the findings suggest that formal-colloquial variation in Spanish did not produce a statistically detectable systematic effect on CAG faithfulness under the conditions examined, while simultaneously revealing descriptive differences in how individual models responded to that variation. The contrasting tendencies of SmolLM2 and Ministral caution against assuming that either formal or colloquial formulations are inherently more favorable for contextual grounding. Qwen2.5, meanwhile, combined comparatively high faithfulness, limited separation between registers and relatively low score variability, providing the most stable descriptive profile among the evaluated models. Taken together, these results suggest that the practical robustness of CAG to intralinguistic variation may depend substantially on the underlying language model and task configuration. At the same time, the exploratory nature of the study precludes stronger claims of register invariance or model superiority and motivates further evaluation across larger datasets, additional linguistic varieties and direct comparisons with retrieval-based architectures.

Limitations and Future Work

Several limitations should be considered when interpreting the findings of this exploratory study. First, the experiment was intentionally restricted to a relatively small and controlled

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

dataset comprising 10 concepts, 20 semantic cases and 120 planned generations across three language models. Although the paired design enabled linguistic register to be isolated while holding semantic intent and contextual knowledge constant, the resulting number of complete formal-colloquial pairs was limited. Of the 60 planned pairs, 53 were available for paired analysis and only 14 semantic cases contained complete paired differences across all three models. Consequently, the statistical analyses may have had limited power to detect small register effects or systematic between-model differences. The absence of statistically significant results should therefore not be interpreted as evidence that formal and colloquial formulations are equivalent in their effect on faithfulness.

A second limitation concerns the seven generations for which a valid faithfulness score could not be obtained because of technical failures during automated evaluation. These observations were retained as missing rather than assigned artificial scores or imputed, preserving the integrity of the available measurements but producing unequal sample sizes across some model-task-register conditions.

This issue was particularly relevant to the model-level comparison, which required complete observations across all three models and consequently reduced the Friedman analysis to 14 semantic cases. Future replications should incorporate more robust evaluation and parsing procedures to minimize technical missingness and increase the number of complete paired observations.

A further limitation concerns the linguistic and semantic scope of the experiment. The study examined a specific form of intralinguistic variation in Spanish, contrasting formal and colloquial formulations of semantically equivalent queries. These two conditions do not capture the full range of variation present in naturally occurring Spanish, which may additionally involve regional varieties, dialectal features, sociolectal variation, lexical preferences, pragmatic conventions or different degrees of formality and conversationality. The findings should therefore be restricted to the formal-colloquial contrast operationalized in the present dataset rather than generalized to linguistic variation in Spanish as a whole. Future work could extend the experimental design to additional varieties and registers while preserving the paired semantic structure used in this study.

The semantic domain was also deliberately constrained. The knowledge base comprised ten concepts related to emotional states and feelings, with definition generation and concept identification used as the two experimental task types. This controlled domain facilitated the construction of semantically equivalent query pairs and a fixed knowledge context, but it limits the extent to which the observed patterns can be generalized to other knowledge domains or more complex information needs. Replication with larger and more heterogeneous knowledge bases including technical, scientific, administrative or general-domain content would help determine whether the observed register patterns persist as the semantic and reasoning demands of the task increase.

The scope of the findings is also constrained by the specific models evaluated. The experiment included three relatively compact open-source instruction-tuned LLMs and the contrasting descriptive patterns observed across them indicate that conclusions obtained from one model should not automatically be generalized to others.

In particular, the comparatively stable profile observed for Qwen2.5 should be interpreted as a result of the present experimental configuration rather than as evidence of a general property of Qwen models or CAG systems. Future studies should examine a broader range of model families and parameter scales to determine whether similar patterns emerge and whether model capacity or architecture is associated with sensitivity to intralinguistic register variation.

Faithfulness measurement introduces an additional limitation. The study relied on RAGAS faithfulness with granite 3.2 as a single automated LLM judge. Although this provided a consistent evaluation procedure across all experimental conditions, the resulting scores remain dependent on the claim decomposition and contextual-support judgments produced by the evaluation pipeline.

The study did not include independent human annotation or multiple LLM judges to assess inter-rater agreement. Future work could therefore combine RAGAS with human evaluation and alternative judge models to examine the stability of the observed faithfulness measurements across evaluation methods.

More broadly, the present study should be understood as a controlled exploratory experiment intended to open, rather than exhaust, the investigation of linguistic robustness in CAG. Its deliberately constrained design provides a reproducible setting in which semantic intent, contextual knowledge and generation configuration remain fixed while linguistic register is manipulated.

The resulting findings are therefore better interpreted as initial empirical evidence and as a basis for generating further hypotheses than as definitive estimates of register effects in deployed CAG systems. In particular, the opposing descriptive tendencies observed across models and the comparatively stable profile of Qwen2.5 raise questions that warrant

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

investigation with larger datasets, additional models, broader linguistic variation and more diverse knowledge domains.

A particularly relevant extension would be a controlled comparison between CAG and RAG under the same register manipulation. The present design deliberately removes retrieval variability by exposing formal and colloquial queries to the same preloaded knowledge context. Repeating the experiment with a retrieval-based pipeline while preserving the same models, knowledge base, semantic cases and query pairs would make it possible to examine whether retrieval introduces an additional source of sensitivity to linguistic formulation. Such a comparison could determine whether the descriptive stability observed for some CAG configurations, particularly Qwen2.5 in the present experiment, persists under retrieval or represents a potential advantage of eliminating query-dependent context selection. This question remains open, as the present study did not directly compare CAG and RAG.

Conclusions

This study investigated whether linguistic register affects the faithfulness of cache-augmented generation when semantically equivalent queries are expressed in formal and colloquial Spanish. Under the controlled experimental conditions examined, the inferential analyses did not provide sufficient evidence of a systematic register effect for any of the three evaluated models. Neither the paired parametric and non-parametric analyses nor the between-model comparison identified statistically significant differences attributable to linguistic registers.

These findings should not be interpreted as evidence of equivalence between formal and colloquial formulations, but rather as an absence of statistically detectable systematic effects within the scope and sample size of this exploratory experiment.

The descriptive results nevertheless revealed distinct model- and task-specific patterns. SmolLM2 tended toward higher faithfulness for formal formulations, particularly in definition generation, whereas Ministral showed the opposite tendency, most prominently in concept identification. Qwen2.5 exhibited the most stable descriptive profile, combining comparatively high faithfulness, the smallest formal-colloquial separation and relatively low within-condition variability.

The coexistence of these contrasting patterns suggests that stability across linguistic registers should not be regarded as an inherent property of CAG alone, but may depend substantially on the underlying language model and task configuration. Overall, this work provides a controlled exploratory framework for studying intralinguistic robustness in CAG by holding semantic intent, contextual knowledge and generation conditions constant while manipulating linguistic register. Rather than establishing a definitive advantage for either formal or colloquial Spanish, the findings open a broader empirical question concerning the conditions under which cache-based knowledge grounding remains stable across heterogeneous forms of user language.

Replication with larger and more diverse datasets, additional model families and Spanish varieties and direct CAG-RAG comparisons will be necessary to determine whether the descriptive stability observed in particular configurations generalizes beyond the present experimental setting.

Data and Code Availability

The experimental code, query dataset, generated outputs and analysis resources used in this study are publicly available in the accompanying GitHub repository: Garciacarmonaam/spanish-cag-register-faithfulness.

The repository contains the resources required to reproduce the experimental workflow, including the CAG implementation, formal and colloquial Spanish query pairs, model outputs, faithfulness evaluation results and statistical analysis notebooks. The original source code is released under the MIT License. Third-party lexical content remains subject to the terms and rights of its respective source.

The repository also documents the model configurations and software dependencies used in the experiments. Missing faithfulness evaluations are preserved as missing observations rather than imputed, allowing the reported descriptive and paired statistical analyses to be independently reproduced from the released experimental results.

Declaration of Generative AI Use

Generative AI tools were used during the development of this study for technical consultation, statistical analysis support, code assistance and debugging and language editing.

GPT-5 through the ChatGPT interface was additionally used to generate the initial formal and colloquial query formulations employed in the experimental dataset; all query pairs were subsequently reviewed by the author to verify preservation of semantic intent and requested information.

The author independently defined the research question and experimental design, constructed the knowledge base, executed and validated the experiments and statistical

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/

analyses, evaluated the methodological decisions, interpreted the results and remains solely responsible for the content and conclusions of the study.

References

1. Oche AJ, Folashade AG, Ghosal T, Biswas A (2025) A systematic review of key Retrieval-Augmented Generation (RAG) systems: Progress, gaps and future directions. arXiv. [Crossref], [Google Scholar]
2. Carrasco-Sáez JL, Contreras-Saavedra C, San-Martín-Quiroga S, Contreras-Saavedraand CE, Viveros-Muñoz R (2025) Analyzing higher education students' prompting techniques and their impact on ChatGPT's performance: An exploratory study in Spanish. Appl Sci 15: 7651. [Crossref], [Google Scholar]
3. Viveros-Muñoz R, Carrasco-Sáez J, Contreras-Saavedra C, San-Martín-Quiroga S, Contreras-Saavedra CE (2025) Does the grammatical structure of prompts influence the responses of generative Artificial Intelligence? An exploratory analysis in Spanish. Appl Sci 15: 3882. [Crossref], [Google Scholar]
4. BJ Chan, CT Chen, JH Cheng, HH Huang (2025) Don't do RAG: When cache-augmented generation is all you need for knowledge tasks. ACM Digital Library 893-897. [Crossref], [Google Scholar]
5. Lu S, Wang H, Rong Y, Chen Z, Tang Y (2025) TurboRAG: Accelerating retrieval-augmented generation with precomputed kv caches for chunked text. Assoc Comput Linguistics 6588-6601. [Crossref], [Google Scholar]
6. Agarwal S, Sundaresan S, Mitra S, Mahapatra D, Gupta A, et al. (2025) Cache-craft: Managing chunk-caches for efficient retrieval-augmented generation. ACM Digital Library 3: 1-28. [Crossref], [Google Scholar]
7. Corallo G, Weller O, Petroni F, Papotti P (2026) CacheNotes: Task-aware key-value cache compression for reasoning-intensive knowledge tasks Assoc Comput Linguistics 1: 6571-6590. [Crossref], [Google Scholar]
8. Rashkin H, Reitter D, Tomar GS, Das D (2021) Increasing faithfulness in knowledge-grounded dialogue with controllable features. Assoc Comput Linguistics 1: 704-718. [Crossref], [Google Scholar]
9. Es S, James J, Anke LE, Schockaert S (2024) Ragas: Automated evaluation of retrieval augmented generation. Assoc Comput Linguistics 150-158. [Crossref], [Google Scholar]
10. Wallat J, Heuss M, Rijke M, Anand A (2025) Correctness is not faithfulness in retrieval augmented generation attributions. ACM Digital Library 22-32. [Crossref], [Google Scholar]
11. SG Kwak, JH Kim (2017) Central limit theorem: The cornerstone of modern statistics. Korean J Anesthesiol 70: 144-156. [Crossref], [Google Scholar], [Indexed]

Copyright © 2026 | https://engineering-ai.academicsquare-pub.com/