Digital Linguistic Bias in Spanish: Evidence from Lexical Variation in LLMs

Anonymous ACL submission

Abstract

This study examines the extent to which Large Language Models (LLMs) capture geographic lexical variation in Spanish, a language that exhibits substantial regional variation. Treating LLMs as *virtual informants*, we probe their dialectal knowledge using two survey-style question formats — Yes–No questions and multiple–choice questions. To this end, we exploited a large-scale, expert-curated database of Spanish lexical variation. Our evaluation covers more than 900 lexical items across 21 Spanish-speaking countries and is conducted at both the country and dialectal area levels.

Across both evaluation formats, the results reveal systematic differences in how LLMs represent Spanish language varieties. Lexical variation associated with Spain, Equatorial Guinea, Mexico & Central America, and the La Plata River is recognized more accurately by the models, while the Chilean variety proves particularly difficult for the models to distinguish. Importantly, differences in the volume of country–level digital resources do not account for these performance patterns, suggesting that factors beyond data quantity shape dialectal representation in LLMs.

By providing a fine-grained, large-scale evaluation of geographic lexical variation, this work advances empirical understanding of dialectal knowledge in LLMs and contributes new evidence to discussions of *Digital Linguistic Bias* in Spanish.

1 Introduction

Spanish is spoken as a first language by more than 520 million people (Instituto-Cervantes, 2025) and exhibits substantial geographic lexical variation across the Spanish-speaking world. For many everyday concepts, multiple lexical variants coexist, with their distribution shaped by historical, regional, and sociolinguistic factors. A well-known example is a variation in referring to a vehicle equivalent to English *car*; they mainly use *coche* in Spain, whereas *carro* is preferred in many Latin American countries (Ueda and Moreno-Fernández, 2016). Documenting and modeling this variation has long been a central concern of Spanish dialectology (Zamora-Vicente, 1968; Lipski, 1994; Alvar, 1996a,b; Fernández-Ordóñez, 2011; Gonçalves and Sánchez, 2014; Ueda and Moreno-Fernández, 2016; Moreno-Fernández, 2020a,b; Moreno-Fernández and Caravedo, 2023; Ueda, 2023). According to diverse linguistic features, the 21 principal Spanish-speaking countries can be grouped into eight dialectal areas[1].

Large Language Models (LLMs) are now widely used in downstream applications that interact directly with speakers of different regional varieties. However, these models are trained primarily on large-scale web data, whose composition is estimated to be uneven across regions and to privilege standardized or hegemonic varieties in detriment of dialect fairness (Kantharuban et al., 2023; Muñoz-Basols et al., 2024; Lin et al., 2025). As a result, concerns have been raised that LLMs may inadequately represent dialectal diversity, thereby producing *Digital Linguistic Bias* (DLB) (Muñoz-Basols et al., 2024). DLB refers to linguistic hybridity generated by Artificial Intelligence at an inter-linguistic level (e.g., preponderance of English) and at an intra-linguistic level (e.g., preponderance of a dominant variety of a language). It may potentially undermine the real-world linguistic diversity in the digital sphere.

[1] **Spain**, **Equatorial Guinea**, **Antilles** (Cuba, Dominican Republic, Puerto Rico), **Mexico & Central America** (Mexico, Guatemala, Honduras, El Salvador, Nicaragua), **Continental Caribe** (Costa Rica, Panama, Venezuela), **Andes** (Colombia, Ecuador, Peru, Bolivia), **Chile**, and **La Plata River** (Argentina, Uruguay, Paraguay). Note that Puerto Rico, which is currently a self-governing commonwealth of the United States, was counted as a country for uniformity. The division of Latin American varieties is based on Moreno-Fernández (2020a). Spain and Equatorial Guinea are treated separately in this work.

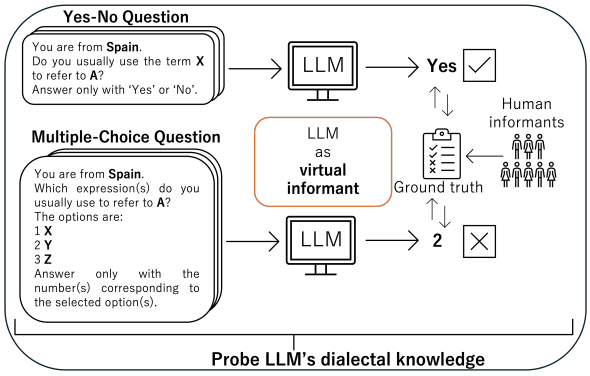

Figure 1: Schematic overview of the paper. The questions are translated into English to facilitate understanding.

LLMs presumably encode dialectal information implicitly and can be prompted to exhibit region-specific lexical or stylistic features. However, systematic evaluation of this knowledge remains limited, particularly for languages such as Spanish (Mayor-Rocher et al., 2025), where geographic variation is rich and well documented (Moreno-Fernández, 2020a,b; Moreno-Fernández and Caravedo, 2023). Therefore, clarifying the extent to which LLMs can handle Spanish language varieties is essential for developing LLMs that represent its linguistic and cultural diversity (Grandury et al., 2025).

In this paper, we provide a large-scale evaluation of geographic lexical variation in Spanish as represented in LLMs (see Figure 1). We treat LLMs as *virtual informants* and probe their dialectal knowledge using two survey-style question formats — Yes–No questions and multiple–choice questions. To this end, we exploited VARILEX, a large-scale, expert-curated database of Spanish lexical variation (Ueda and Moreno-Fernández, 2016).

Our evaluation covers 934 lexical items across 21 Spanish-speaking countries and assesses performance at both the country and dialectal area levels. To support robust comparisons across items with varying degrees of lexical diversity, we adopt evaluation metrics tailored to each question format, including a chance-corrected, adjusted Jaccard coefficient for multi-variant responses. This evaluation framework enables fine-grained analysis of dialectal knowledge while controlling for the task's structural properties.

Our results show that LLMs capture geographic lexical variation unevenly across Spanish language varieties[2]. Varieties associated with Spain, Equatorial Guinea, Mexico & Central America, and the La Plata River are recognized most accurately by the models, while Chilean Spanish proves particularly challenging. Importantly, we find that differences in the volume of country–level digital resources do not explain these patterns, suggesting that factors beyond data quantity shape dialectal representation in LLMs.

This work makes three main contributions: (i) a large-scale, expert-grounded evaluation of Spanish geographic lexical variation in LLMs; (ii) a survey-style evaluation framework that bridges traditional dialectology and LLM assessment; and (iii) new empirical evidence informing discussions of DLB in Spanish.

The rest of the paper is organized as follows. In Section 2, we review related research. Section 3 describes our methodology. In Section 4, we present the experimental setup and results, followed by discussion in Section 5. Section 6 concludes the paper.

2 Related Work

Explicit evaluations of dialectal knowledge in LLMs have so far focused mainly on English and other high-resource languages, using classification tasks or targeted prompts. These studies consistently report uneven performance across varieties (Kantharuban et al., 2023; Lin et al., 2025), thereby raising concerns about dialect fairness (Davidson et al., 2019; Gallegos et al., 2024) and DLB (Muñoz-Basols et al., 2024).

For Spanish, large-scale and fine-grained evaluations of geographic lexical variation remain limited, except for Mayor-Rocher et al. (2025), who addressed 30 items, of which 20 relate to morphosyntactic variation and 10 to lexical variation. The authors prompted various LLMs with multiple–choice questions and assessed performance across the conventional dialectal areas. They found that Peninsular Spanish and La Plata River varieties were better distinguished by the models, whereas varieties associated with the Andes and Mexico & Central America were more difficult for the models to recognize. They attributed the varying performance to differences in the amount of digital resources across areas, given that the two variables present an almost perfect positive correlation.

[2]Experimental results will be publicly available upon acceptance.

This study builds on Mayor-Rocher et al. (2025) and addresses the aforementioned gap by evaluating extensively Spanish lexical variation in LLMs.

3 Methods

We evaluate the extent to which LLMs capture geographic lexical variation in Spanish by treating them as *virtual informants* and eliciting dialectal judgments through structured question–answering tasks. This section describes the reference dataset, the question formats used to probe the models, and the evaluation metrics.

3.1 Reference Dataset: VARILEX

As a gold-standard reference, we work with *VARILEX (Variación léxica del español en el mundo)*, a large-scale, expert-curated database documenting geographic lexical variation across the Spanish-speaking world (Ueda and Moreno-Fernández, 2016). We rely on its revised version[3], in which local linguists annotated the usage of lexical variants at the country level using a binary scheme: "+" denotes predominant usage, while "–" indicates marginal or absent usage. The database comprises 934 lexical items and 9,057 variants, covering 21 Spanish-speaking countries.

Each item includes an index[4], a Spanish description[5], an English gloss, and a list of lexical variants. Table 1 illustrates the structure of a representative item (A141), referring to a vehicle equivalent to English *car*, for which VARILEX lists six variants (*auto*, *automóvil*, *carro*, *coche*, *concho*, *máquina*). The number of variants per item varies substantially, ranging from 1 to 114, with a median of 8. Moreover, the number of predominant variants differs across countries per item; while Spain predominantly uses only one variant (*coche*) for item A141, Argentina exhibits multiple predominant forms (*auto*, *automóvil*, *coche*), as shown in Table 2. These characteristics make VARILEX particularly suitable for evaluating fine-grained geographic lexical knowledge, as they reflect both cross-item and cross-country variation in lexical inventories.

[3] https://h-ueda.sakura.ne.jp/varilex-r/

[4] A first letter in the item index, either of "A", "B", "C", "D", "E", "F", "I", represents when the original survey was conducted and does not hold linguistic relevance.

[5] Example sentences included in some entries were excluded for uniformity.

Item index	A141
Description	*vehículo destinado al transporte de personas* "*vehicle intended for the transport of people*"
English equivalent	CAR
Variants	*auto*, *automóvil*, *carro*, *coche*, *concho*, *máquina*

Table 1: Item A141 referring to a vehicle equivalent to English *car*.

Country	Predominant variants
Spain	*coche*
Equatorial Guinea	*auto*, *coche*
Dominican Republic	*automóvil*, *carro*, *concho*
Mexico	*auto*, *automóvil*, *carro*, *coche*
Venezuela	*automóvil*, *carro*
Peru	*auto*, *automóvil*, *carro*
Chile	*carro*, *coche*
Argentina	*auto*, *automóvil*, *coche*

Table 2: Predominant variants in some countries for the item A141 referring to a vehicle equivalent to English *car*.

3.2 Question Formats

To probe lexical variation, we employ two complementary question–answering formats that simulate traditional dialectological surveys. In both formats, prompts explicitly instruct the model to ignore previous questions, ensuring that each response is treated as independent. See Appendix A for question templates.

Yes–No Question format (YNQF) The *Yes–No Question Format* simulates a survey in which an informant from a given country is asked whether they typically use a specific lexical variant to refer to a certain concept. The model is instructed to respond exclusively with *Sí* "Yes" or *No* "No". The predominant variants for each country are labeled "+" in the dataset. Each prompt specifies the country, the lexical variant, and the Spanish description of the item. This format allows us to assess the model's ability to recognize whether a given variant is predominant in a particular country.

Multiple–Choice Question Format (MCQF) The *Multiple–Choice Question Format* simulates a survey in which an informant is asked to select all lexical variants they would typically use to refer to a given concept. The variants associated with an item are enumerated as options, and the model is instructed to return only the corresponding option numbers, separated by a forward slash if multiple options are selected. The set of gold variants for

each country is those labeled "+" in the dataset. This format captures the fact that multiple variants may be simultaneously predominant within a country and enables set-based comparison between gold annotations and model responses.

3.3 Evaluation Metrics

We evaluate model performance separately for YNQF and MCQF using metrics tailored to the structure of each task.

F1 Score For YNQF, country–level performance is measured by means of the conventional F1 score computed using `sklearn.metrics.f1_score`[6] with `average='binary'`. Higher F1 scores indicate better alignment between model responses and gold annotations.

Adjusted Jaccard Coefficient For MCQF, we measure performance using the adjusted Jaccard coefficient (J_{adj}), a chance-corrected measure of overlap between two sets. By correcting for chance agreement, J_{adj} enables meaningful comparison across items with differing numbers of variants and supports robust aggregation at the country and area levels. Country–level performance is computed as the average across questions.

Let A denote the set of gold-standard predominant variants for a given item and country, and B the set of variants predicted by the model. The conventional Jaccard coefficient (Real and Vargas, 1996) is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{X}{s+t-X}, \qquad (1)$$

where $|A| = s$, $|B| = t$, and $X = |A \cap B|$. A larger value means better performance. However, because overlap can arise spuriously when sets are large, we correct for chance agreement by defining:

$$J_{\text{adj}} = \frac{J - \mathbb{E}[J]}{1 - \mathbb{E}[J]}, \qquad (2)$$

where J is the conventional Jaccard and $\mathbb{E}[J]$ its expected value.

Assume the null hypothesis that A and B are formed by selecting s and t variants uniformly at random from the N variants, independently of each other. Under this model, the size of the intersection X follows a hypergeometric distribution:

$$X \sim \text{Hypergeometric}(N, t, s).$$

Thus,

$$\mathbb{E}[X] = \frac{st}{N}.$$

This expected value quantifies the chance similarity arising from the random overlap between two randomly selected sets of sizes s and t.

Because $|A \cup B| = s + t - X$, we obtain

$$\mathbb{E}[|A \cup B|] = s + t - \mathbb{E}[X] = s + t - \frac{st}{N}.$$

Therefore, a common first-order approximation to the expected Jaccard coefficient is

$$\begin{aligned} \mathbb{E}[J] \approx \frac{\mathbb{E}[X]}{\mathbb{E}[|A \cup B|]} &= \frac{\frac{st}{N}}{s + t - \frac{st}{N}} \\ &= \frac{st}{N(s+t) - st}. \end{aligned} \qquad (3)$$

The adjusted Jaccard coefficient J_{adj} provides a chance-corrected, normalized measure of similarity between two sets.

This transformation has the following properties:

- $J_{\text{adj}} = 1$ if and only if $A = B$.
- $J_{\text{adj}} = 0$ when the observed similarity equals its expected value under the null model.
- $J_{\text{adj}} < 0$ indicates less overlap than would be expected by chance.

We clipped negative values to zero, ensuring that the metric is bounded between 0 and 1.

4 Experiments

4.1 Experimental Setting

Prompts We evaluated LLMs using the two question formats described in Section 3.2: the Yes–No Question Format (YNQF) and the Multiple–Choice Question Format (MCQF). For YNQF, we randomly sampled 10,000 questions from the complete set of 190,158 possible country–variant–item combinations. For MCQF, we used all 17,911 questions corresponding to lexical items with at least two variants. To mitigate ordering effects, the variants associated with each item were randomly shuffled before enumeration in MCQF, following established findings on option-order sensitivity in multiple–choice evaluation (Pezeshkpour and Hruschka, 2024). In both formats, questions were presented to the models in randomized order. Only responses that strictly adhered to the prompt instructions (i.e.,

[6] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

Model	N_Q	N_A	$F1$
baseline	10,000	10,000	0.249
gpt-4o	10,000	9,999	**0.514**
gpt-5.1	10,000	10,000	0.499
gpt-5.2	10,000	10,000	0.480

Table 3: Overall performance on YNQF. N_Q and N_A represent the number of questions and valid responses, respectively.

Model	N_Q	N_A	J_{adj}
baseline	17,911	17,911	0.110
gpt-4o	17,911	17,853	0.314
gpt-5.1	17,911	17,911	**0.338**
gpt-5.2	17,911	17,910	0.336

Table 4: Overall performance on MCQF.

Country	N_A	$F1$	baseline	$\Delta F1$
Spain	469	**0.723**	0.318	**0.405**
Equatorial Guinea	480	0.448	0.190	0.258
Cuba	434	0.619	0.386	0.233
Dominican Republic	507	0.463	0.246	0.217
Puerto Rico	462	0.527	0.263	0.264
Mexico	466	0.692	0.363	0.329
Guatemala	506	0.485	0.154	0.331
Honduras	479	0.557	0.248	0.309
El Salvador	476	0.325	0.159	0.166
Nicaragua	458	0.571	0.332	0.239
Costa Rica	489	0.325	0.137	0.188
Panama	486	0.355	0.206	0.149
Colombia	464	0.282	0.117	0.165
Venezuela	515	0.539	0.223	0.316
Ecuador	447	0.509	0.245	0.264
Peru	466	0.409	0.257	0.152
Bolivia	451	0.415	0.161	0.254
Chile	475	0.372	0.292	0.080
Paraguay	490	0.588	**0.388**	0.200
Uruguay	498	0.500	0.211	0.289
Argentina	481	0.665	0.359	0.306

Table 5: Country–level performance of gpt-4o on YNQF. $\Delta F1$ indicates performance gain against baseline.

valid *Sí/No* answers for YNQF and properly formatted numeric selections for MCQF) were retained for evaluation.

Baselines For YNQF, we set the baseline by responding *Sí* "Yes" to all the questions. Overall, only approximately 25% of variant–country pairs are labeled positive with varying percentages across countries. For MCQF, we set the baseline by selecting the first three options (e.g., 1/2/3) for items with three or more variants. The value of three derives from rounding off the average number of gold-standard variants, 2.66, to the nearest integer. For items with two variants, the model returns both. Random shuffling prior to enumeration is expected to favor no specific variants.

Models We evaluated three models accessed via the official API: gpt-4o, which achieved the best performance in prior work (Mayor-Rocher et al., 2025), and its successor models gpt-5.1 and gpt-5.2. We did not include other model families, as they were previously shown to be substantially outperformed by GPT models on similar tasks (Mayor-Rocher et al., 2025). Our focus is therefore on expanding lexical coverage and deepening dialectal analysis rather than on broad model comparison.

4.2 Results

4.2.1 Overall Performance

Table 3 summarizes overall performance on YNQF, measured using the F1 score. N_Q and N_A represent the number of questions and valid responses, respectively. gpt-4o achieves an F1 score of 0.514, outperforming newer gpt-5.1 (0.499) and gpt-5.2 (0.480). Every model improves by almost twice the baseline (0.249). These results indicate that all evaluated models demonstrate a reasonable ability to recognize geographic lexical variation, with modest differences across model versions.

Table 4 presents overall performance on MCQF, measured using the adjusted Jaccard coefficient (J_{adj}). The results show that gpt-5.1 achieves the highest score (0.338), closely followed by gpt-5.2 (0.336), while gpt-4o performs slightly worse (0.314). The newer models also adhere more closely to prompt constraints, producing fewer invalid responses. Every model improves by almost three times over the baseline (0.110). Taken together, these results suggest incremental improvements in both instruction-following and dialectal recognition across model generations.

4.2.2 Country–Level Performance

In the remainder of this section, we focus on the best-performing gpt-4o and gpt-5.1 for YNQF and MCQF, respectively.

Table 5 reports country–level performance of gpt-4o on YNQF. $\Delta F1$ indicates performance gain against baseline. The highest F1 scores are observed in Spain (0.723), Mexico (0.692), Argentina (0.665), Cuba (0.619), and Paraguay (0.588). In contrast, the lowest scores are obtained for Colombia (0.282), Costa Rica (0.325), El Salvador (0.325), Panama (0.355), and Chile (0.372).

Table 6 presents country–level performance of gpt-5.1 on MCQF. ΔJ_{adj} indicates performance

Country	N_A	J_{adj}	baseline	ΔJ_{adj}
Spain	907	**0.508**	0.110	**0.398**
Equatorial Guinea	826	0.412	0.087	0.325
Cuba	920	0.317	0.120	0.197
Dominican Republic	892	0.339	0.102	0.237
Puerto Rico	917	0.306	0.138	0.168
Mexico	928	0.381	**0.140**	0.241
Guatemala	918	0.329	0.101	0.228
Honduras	921	0.332	0.114	0.218
El Salvador	717	0.310	0.102	0.208
Nicaragua	917	0.306	0.121	0.185
Costa Rica	563	0.308	0.103	0.205
Panama	832	0.279	0.092	0.187
Colombia	510	0.288	0.090	0.198
Venezuela	908	0.352	0.132	0.220
Ecuador	910	0.330	0.106	0.224
Peru	853	0.318	0.100	0.218
Bolivia	908	0.335	0.107	0.228
Chile	818	0.141	0.101	0.040
Paraguay	918	0.366	0.100	0.266
Uruguay	907	0.344	0.103	0.241
Argentina	921	0.448	0.125	0.323

Table 6: Country–level performance of `gpt-5.1` on MCQF. ΔJ_{adj} indicates performance gain against baseline.

gain against baseline. Spain (0.508) achieves the strongest result, followed by Argentina (0.448), Equatorial Guinea (0.412), Mexico (0.381), and Paraguay (0.366). The weakest performance is again observed in Chile (0.141), with Panama (0.279), Colombia (0.288), Nicaragua (0.306), and Puerto Rico (0.306) also yielding comparatively low scores.

The two metrics, F1 score and adjusted Jaccard coefficient, exhibit a moderate positive correlation[7] ($\rho = 0.558$ ($p < 0.01$)), indicating that the YNQF and MCQF evaluations provide seemingly consistent and complementary results. That said, we demonstrate that the F1 score is not an appropriate metric for evaluating performance, as discussed below.

5 Discussion

This section interprets the experimental results to clarify which aspects of geographic lexical variation LLMs successfully capture, which factors influence their performance, and what these findings imply for DLB in Spanish.

5.1 Yes–No Question Format (YNQF)

We assessed whether differences in country–level performance on YNQF reflect genuine dialectal knowledge rather than artifacts of label imbalance. F1 scores do not correlate significantly with the proportion of positive model responses ($\rho = 0.247$ ($p = 0.279$)). However, we observed a strong positive correlation between F1 scores and the proportion of positive variants in the gold annotations ($\rho = 0.891$ ($p < 0.001$)). Even stronger tendency was seen between the baseline and the proportion of positive gold labels ($\rho = 0.903$ ($p < 0.001$)). F1 scores also correlate strongly with the baseline ($\rho = 0.794$ ($p < 0.001$)). Therefore, F1 scores should not be viewed as reflecting models' genuine dialectal knowledge but as spurious artifacts derived from systematic differences in label distribution across countries[8].

$\Delta F1$, which represents performance gain against baseline, can be an alternative metric, as it does not correlate significantly with the random baseline ($\rho = 0.169$ ($p = 0.464$)). Note, however, that it still moderately correlates with the proportion of positive gold labels ($\rho = 0.438$ ($p = 0.047$)). In the following, we rely on $\Delta F1$ for further discussion on YNQF, although the insights should be taken with due reservations. The highest $\Delta F1$ scores are observed in Spain (0.405), Guatemala (0.331), Mexico (0.329), Venezuela (0.316), and Honduras (0.309). In contrast, the lowest scores are obtained for Chile (0.080), Panama (0.149), Peru (0.152), Colombia (0.165), and El Salvador (0.166).

$\Delta F1$ exhibits a moderate positive correlation with J_{adj} ($\rho = 0.679$ ($p < 0.01$)), as illustrated in Figure 2, where countries are colored by dialectal area. The positive correlation indicates that the YNQF and MCQF evaluations provide consistent and complementary results.

5.2 Multiple–Choice Question Format (MCQF)

For MCQF, we first examined whether the models were sensitive to question configuration. A moderate positive correlation was observed between the number of gold-standard variants and the number of model-selected variants ($\rho = 0.437$ ($p < 0.01$)). The positive correlation suggests that the models approximate, to some extent, the number of responses appropriate for a given item, regardless of correctness.

Performance as measured by the adjusted Jac-

[7]This study uses Spearman's rank correlation ρ and sets the significance threshold at $p = 0.05$.

[8]Other metrics —Matthew's correlation coefficient and Cohen's κ — also present similar patterns, suggesting difficulties in controlling for the effect of the varying proportion of positive gold labels.

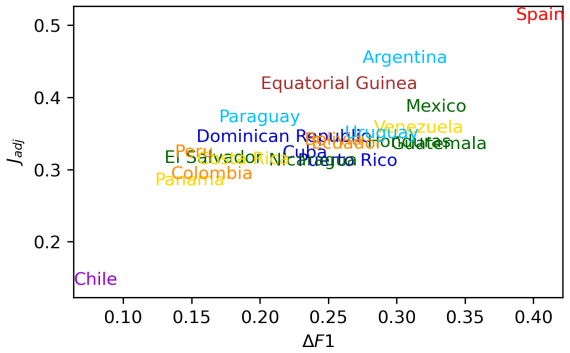

Figure 2: J_{adj} plotted against $\Delta F1$, with countries colored by dialectal area: $\rho = 0.679$ ($p < 0.01$).

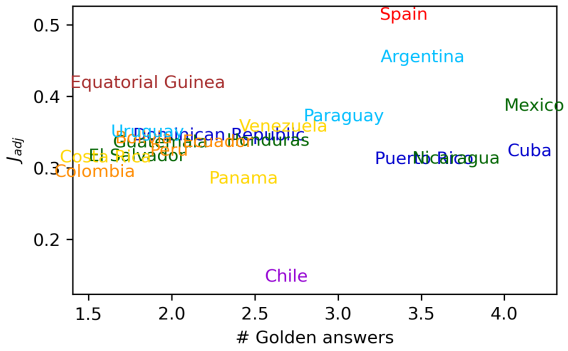

Figure 3: J_{adj} at country–level plotted against the number of the golden answers, with countries colored by dialectal area: $\rho = 0.183$ ($p = 0.427$).

card coefficient (J_{adj}) shows a slight negative correlation with both the number of gold variants ($\rho = -0.114$ ($p < 0.01$)) and the number of model responses ($\rho = -0.224$ ($p < 0.01$)). This pattern arises naturally from the chance-correction built into J_{adj}; as the number of potential overlaps increases, chance agreement becomes more likely, and the metric penalizes spurious coincidence. Importantly, this behavior indicates that J_{adj} appropriately controls for item difficulty and response set size. Consistent with this interpretation, items with an increased number of variants are substantially more difficult for the models, as reflected in the moderate negative correlation between J_{adj} and the number of variants per item ($\rho = -0.591$ ($p < 0.01$)).

At the country level, however, J_{adj} shows no significant correlation with the average number of gold variants ($\rho = 0.183$ ($p = 0.427$))[9], as illustrated in Figure 3. J_{adj} does not correlate significantly with the average number of model responses, either ($\rho = 0.057$ ($p = 0.806$)). This absence of correlation indicates that country–level performance differences are not driven by structural properties of the evaluation task, reinforcing the validity of the observed geographic patterns. Contrary to YNQF, no significant correlation was observed between the metric, J_{adj}, and the baseline ($\rho = 0.246$ ($p = 0.283$)). These properties endorse the appropriateness of the metric and the results.

[9]Use of the conventional Jaccard coefficient results in a moderate positive correlation with the number of gold variants ($\rho = 0.655$ ($p < 0.01$)), which would complicate the assessment of model's genuine ability.

5.3 Digital Resources and Dialectal Knowledge

5.3.1 Country level

A common assumption in the literature is that LLMs more accurately represent varieties associated with larger volumes of web data (Mayor-Rocher et al., 2025)[10]. Note, however, that the actual data volume used to train LLMs is not made public. We tentatively test this hypothesis, following Mayor-Rocher et al. (2025). Specifically, we examined the relationship between country–level performance and estimated amounts of digital resources derived from CEREAL (Corpus del Español REAL) (España-Bonet and Barrón-Cedeño, 2024)[11]. The amount of digital resources was measured in terms of tokens. Contrary to the hypothesis, we found no significant correlation between data volume and performance for YNQF ($\rho = 0.158$ ($p = 0.493$)) nor MCQF ($\rho = -0.305$ ($p = 0.178$)) (see Figure 4 for the latter).

This result suggests that the quantity of digital data alone does not explain how well LLMs capture geographic lexical variation. Thus, simply adding more data from a country may not enable the models to capture its regional nuances. Rather, more direct intervention, such as fine-tuning LLMs on dialect-rich corpora, may be effective.

While countries with huge digital footprints —most notably Spain (by several orders of magnitude), Argentina, and Mexico— do exhibit strong

[10]The volumes of web data correlate strongly with the country's economy ($\rho = 0.794$ ($p < 0.01$)), when it was measured by Gross Domestic Product (GDP). Thus, the volumes of web data can be considered as a proxy for a country's economy. The GDP data was retrieved from: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD.

[11]https://zenodo.org/records/14771240

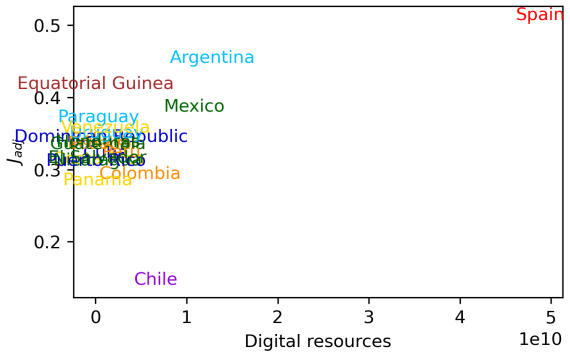

Figure 4: J_{adj} plotted against the amount of digital resources, with countries colored by dialectal area: $\rho = -0.305$ ($p = 0.178$).

Area	$\Delta F1$	J_{adj}
Spain	**0.405**	**0.508**
Equatorial Guinea	0.258	0.412
Antilles	0.238	0.321
Mexico & Central America	0.275	0.332
Continental Caribe	0.218	0.313
Andes	0.209	0.318
Chile	0.080	0.141
La Plata River	0.265	0.386

Table 7: Area–level average performance.

performance, this pattern does not generalize across the dataset. Two cases are particularly informative.

Equatorial Guinea Equatorial Guinea yields a decent score (tenth out of twenty-one) in YNQF and ranks third in MCQF, despite having the smallest estimated data volume. This outcome is plausibly explained by the strong lexical affinity between Equatoguinean and Peninsular Spanish (Ueda, 1995, 2007), allowing knowledge acquired from Spanish data to generalize effectively.

Chile In contrast, Chile performs worst across both evaluation formats, despite having one of the largest digital footprints. For instance, in MCQF, the model's prediction (*máquina*, *auto*, *automóvil*) includes none of the gold-standard variants (*carro*, *coche*) for item A141. The model also fails to recognize exclusively Chilean variants (*furgón*, *furgonita*) for item A157 referring to a vehicle equivalent to English *van*. This finding is noteworthy given the well-documented lexical distinctiveness of Chilean Spanish (Núñez, 2023), and suggests a disconnect between traditional dialectological descriptions and the lexical patterns represented in web data. This unexpected outcome warrants further investigation in future work.

5.3.2 Dialectal Area level

Following Mayor-Rocher et al. (2025), we examined the average performance across the dialectal areas. The results are summarized in Table 7. At the dialectal area level, performance patterns largely mirror those observed at the country level. Spain remains the strongest-performing area. Within the Americas, La Plata River and Mexico & Central America consistently yield slightly higher scores. Then, Antilles, Continental Caribe, and Andes present a practically identical level of performance, followed by a substantial margin by Chile.

Importantly, area–level performance shows no meaningful correlation with digital resource volume; $\rho = 0.262$ ($p = 0.531$) in YNQF and $\rho = 0.119$ ($p = 0.779$) in MCQF. The result disagrees with the finding of Mayor-Rocher et al. (2025), who observed a nearly perfect positive correlation. This finding reinforces the conclusion that factors beyond data quantity —such as lexical similarity across regions, standardization pressures, or corpus composition— play a decisive role.

6 Conclusions

This study evaluated the extent to which LLMs capture geographic lexical variation in Spanish. Treating LLMs as virtual informants, we probed their dialectal knowledge using survey-style Yes–No and multiple–choice questions. To this end, we used an expert-curated database covering more than 900 lexical items across 21 Spanish-speaking countries.

The results reveal systematic differences in how Spanish language varieties are represented. Lexical variation associated with Spain, Equatorial Guinea, Mexico & Central America, and the La Plata River is recognized more accurately by the models, whereas the Chilean variety remains particularly difficult for the models to distinguish. Crucially, differences in country–level digital resource volume do not explain these patterns, indicating that factors beyond data quantity influence dialectal representation in LLMs.

By providing a large-scale, fine-grained evaluation of Spanish geographic lexical variation, this work contributes new empirical evidence to discussions of dialectal knowledge and *Digital Linguistic Bias* in LLMs. Our findings highlight the need for more nuanced approaches to modeling and evaluating linguistic diversity in LLMs, beyond assumptions based solely on data availability.

Limitations

This study has several limitations that point to directions for future research. First, treating LLMs as virtual informants implicitly assumes that their responses approximate those of an average speaker of each dialect. In doing so, we abstract away from sociolinguistic factors such as age, gender, educational background, and social context, all of which are known to influence linguistic choice (Holmes, 2008). Explicitly encoding such variables in prompts may affect variant selection and warrant systematic investigation.

Second, for analytical simplicity, we treat countries as the minimal units of geographic variation and assume internal homogeneity within national boundaries. This abstraction overlooks well-documented intra-country variation, including regional, urban–rural, and social distinctions (Gonçalves and Sánchez, 2014). Future work could refine this approach by incorporating subnational dialectal divisions or by modeling graded variation rather than categorical country–level usage.

Third, our experiments are limited to GPT-family models. While prior work (Mayor-Rocher et al., 2025) suggests that these models substantially outperform other families on similar tasks, extending the analysis to open-weight models would allow for finer-grained probabilistic evaluation and greater transparency regarding model behavior.

Furthermore, the evaluation methodology can be improved. Regarding the Yes–No Question Format, a metric that does not correlate significantly with the number of gold-standard variants in the dataset should be developed to remove confounding factors. On the other hand, Multiple–Choice Question Format may not fully capture the capabilities of LLMs, although they offer an efficient means of automated evaluation (Li et al., 2024). It is also noteworthy that VARILEX's binary labeling ("predominant" vs. "absent/marginal") masks gradation, as a variant could be used often but not "predominantly". Thus, future work might work with frequency data or graded surveys.

In addition to the variant recognition task, evaluating dialect-sensitive text generation (e.g., in Mexican Spanish) would provide a more comprehensive view of how linguistic diversity is represented in model outputs (Dhamala et al., 2021). Such an approach may also support the development of synthetic data for underrepresented varieties (Muñoz-Basols et al., 2024) and facilitate inter-dialect adaptation in practical applications.

Finally, other dimensions of linguistic variation, such as morphosyntactic (Takagaki et al., 2018) and discourse-level phenomena, are equally central to dialectal differentiation and should be examined using comparable evaluation frameworks.

Ethical Considerations

This study involves no human participants, personal data, or human annotation. All analyses are based on publicly available linguistic resources (i.e., VARILEX) and responses generated by LLMs via official APIs.

The work evaluates how geographic lexical variation is represented in language models and does not involve profiling, classification, or inference about individual speakers. Countries and dialectal areas are used as analytical abstractions for evaluation purposes and should not be interpreted as homogeneous linguistic entities. The reported results do not constitute normative judgments about varieties or their speakers and should not be construed as assessments of linguistic quality or legitimacy.

Our analysis is diagnostic in nature and aims to identify potential sources of DLB arising from data distributions and modeling choices. Because the evaluated models are proprietary and their training data are not transparent, our conclusions are limited to observed model behavior and do not make claims about internal representations.

We used AI-assisted tools solely to improve grammatical correctness, clarity, and style.

References

Manuel Alvar, editor. 1996a. *Manual de dialectología hispánica. El Español de América*. Ariel.

Manuel Alvar, editor. 1996b. *Manual de dialectología hispánica: El español de España*. Ariel.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872. ACM.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2024. Elote, choclo and mazorca: On the varieties of Spanish. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3689–3711. Association for Computational Linguistics.

Inés Fernández-Ordóñez. 2011. La lengua de Castilla y la formación del español.

Isabel Orlanes Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50:1097–1179.

Bruno Gonçalves and David Sánchez. 2014. Crowdsourcing dialect characterization through Twitter. *PLoS ONE*, 9:e112074.

María Grandury, Javier Aula-Blasco, Júlia Falcão, Clémentine Fourrier, Miguel González Saiz, Gonzalo Martínez, Gonzalo Santamaria Gomez, Rodrigo Agerri, Nuria Aldama García, Luis Chiruzzo, Javier Conde, Helena Gomez Adorno, Marta Guerrero Nieto, Guido Ivetta, Natàlia López Fuertes, Flor Miriam Plaza del Arco, María-Teresa Martín-Valdivia, Helena Montoro Zamorano, Carmen Muñoz Sanz, and 6 others. 2025. La leaderboard: A large language model leaderboard for Spanish varieties and languages of Spain and Latin America. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32482–32524. Association for Computational Linguistics.

Janet Holmes. 2008. *An Introduction to Sociolinguistics*, 3 edition. Longman.

Instituto-Cervantes. 2025. *El español en el mundo 2025: Anuario del Instituto Cervantes*. Instituto Cervantes.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245. Association for Computational Linguistics.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2819–2834. ELRA Language Resource Association.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J. Wooldridge, Janet B. Pierrehumbert, and Furu Wei. 2025. Assessing dialect fairness and robustness of large language models in reasoning tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6317–6342. Association for Computational Linguistics.

John Lipski. 1994. *El español de América*. Cátedra.

Marina Mayor-Rocher, Cris del Pozo Huerta, Nina Melero, Gonzalo Martínez, María Grandury, and Pedro Reviriego. 2025. Es igual pero no es lo mismo: ¿Distinguen los LLMs las variedades del español? *Procesamiento del Lenguaje Natural*, 75:137–146.

Francisco Moreno-Fernández. 2020a. *La lengua española en su geografía: Manual de dialectología hispánica*, 5 edition. Arco/Libros.

Francisco Moreno-Fernández. 2020b. *Variedades de la lengua española*. Routledge.

Francisco Moreno-Fernández and Rocío Caravedo, editors. 2023. *Dialectología hispánica*. Routledge.

Javier Muñoz-Basols, María del Mar Palomares Marín, and Francisco Moreno-Fernández. 2024. El Sesgo Lingüístico Digital (SLD) en la inteligencia artificial: implicaciones para los modelos de lenguaje masivos en español. *Lengua y Sociedad*, 23:623–647.

Abelardo San Martín Núñez. 2023. *El español en Chile*, pages 216–226. Routledge.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017. Association for Computational Linguistics.

Raimundo Real and Juan M. Vargas. 1996. The probabilistic basis of Jaccard's index of similarity. *Systematic Biology*, 45:380–385.

Toshihiro Takagaki, Noritaka Fukushima, Masami Miyamoto, Antonio Ruiz Tinoco, Hiroto Ueda, Kimiyo Nishimura, and Kahori Umezaki. 2018. Variación gramatical del español en el mundo (VARIGRAMA).

Hiroto Ueda. 1995. Zonificación del español del mundo: Palabras y cosas de la vida urbana. In *Lingüística (ALFAL)*, volume 7, pages 43–86.

Hiroto Ueda. 2007. Zonificación múltiple de las ciudades hispanohablantes según el léxico urbano moderno: Análisis clúster y análisis de componentes principales. In *Jornadas sobre métodos informáticos en el tratamiento de las lenguas ibéricas*, pages 121–140. Centro de Estudios Hispánicos de la Universidad Sofía.

Hiroto Ueda. 2023. *Dialectología del español y dialectometría*, pages 87–104. Routledge.

Hiroto Ueda and Francisco Moreno-Fernández. 2016. VARILEX-R: Variación léxica en español del mundo / Datos revisados.

Alonso Zamora-Vicente. 1968. *Dialectología española (segunda edición muy aumentada)*. Gredos.

A Question Templates

Each question specifies three components in brackets: *country*, *variant*, and *description*.

A.1 Yes–No Question Format (YNQF)

Yes–No Question Format is as follows:
Responda a la siguiente pregunta. No tenga en cuenta las preguntas anteriores.
*Usted es de [**country**].*
*¿Suele utilizar el término «**[variant]**» para referirse a «**[description]**»? Responda únicamente con «Sí» o «No».*

"Answer the following question. Do not take the previous questions into account.
*You are from **[country]**.*
*Do you usually use the term '**[variant]**' to refer to '**[description]**'? Answer only with 'Yes' or 'No'."*

A.2 Multiple–Choice Question Format (MCQF)

Multiple–Choice Question Format is as follows:
Responda a la siguiente pregunta. No tenga en cuenta las preguntas anteriores.
*Usted es de **[country]**.*
*¿Qué expresión(es) suele usar para referirse a «**[description]**»? Las opciones son:*
*1 **[variant 1]***
*2 **[variant 2]***
*3 **[variant 3]***
...
Conteste solo con el número correspondiente a la opción. Puede elegir más de una opción; en ese caso, los números deberán ir separados por el signo «/» en orden ascendente.

"Answer the following question. Do not take the previous questions into account.
*You are from **[country]**.*
*Which expression(s) do you usually use to refer to '**[description]**'? The options are:*
*1 **[variant 1]***
*2 **[variant 2]***
*3 **[variant 3]***
...
Answer only with the number(s) corresponding to the selected option(s). You may choose more than one option; in that case, the numbers should be separated by the '/' sign in ascending order."