Evaluating LLMs or Evaluating "to" LLMs? LLMs and Differential Object Marking in Spanish

Marina Mayor-Rocher[1], Gonzalo Martínez[2], Elena Álvarez-Mellado[1], Nina Melero[2,3] and Pedro Reviriego[2]

[1] *Universidad Autónoma de Madrid, Madrid, Spain*
[2] *Universidad Politécnica de Madrid, Madrid, Spain*
[3] *New York University, Madrid Campus, Spain*

Abstract

This paper analyzes the linguistic competence of Large Language Models (LLMs) regarding Differential Object Marking (DOM) in Spanish. Through the creation of two specific multiple-choice datasets, the performance of 17 state-of-the-art models is evaluated in both obligatory and optional DOM contexts. Results from the first experiment indicate that most models successfully handle obligatory structures, achieving near-perfect precision. However, in optional contexts, significant variability is observed: models show uneven sensitivity to semantic factors such as animacy, definiteness, and agentivity. Furthermore, a tendency is identified in smaller-scale models toward a simplified or excessive use of the prepositional marker. This work underscores the need for fine-grained linguistic analysis to understand how artificial intelligence reproduces complex and gradient phenomena of the Spanish language.

Keywords

LLM, Differential Object Marking, Evaluation.

1. Introduction

The rapid development of Large Language Models (LLMs) has intensified the need to evaluate their linguistic capabilities systematically, in order to understand how they function and, in particular, to guide their improvement. However, these models tend to reproduce various biases derived both from the data on which they have been trained and from the corpus construction processes used. For example, the models in the MarIA project (RoBERTa-base, RoBERTa-large, GPT-2, GPT-large-bne) were trained on data from the Spanish National Library, whose main variety is Peninsular Spanish (Muñoz-Basols et al. 2024: 631). The overrepresentation of certain varieties or registers in training data can be reflected in disproportionately frequent linguistic patterns in model-generated texts. An illustrative example is the excessive use of the word 'delve' in texts generated by ChatGPT. Although uncommon in everyday English, it has become recurrent in recent scientific articles, possibly due to the influence of English used by subcontracted annotators in countries such as Nigeria, where this term is more habitual (Shapira, 2024).

Despite these possible biases, the language generated by current models has managed to achieve a high degree of grammatical correctness and fluency, which means that many of their deviations easily go unnoticed by users. In this vein, the TELEIA benchmark (Mayor-Rocher et al. 2025b) introduces a set of tests designed specifically to evaluate linguistic knowledge of

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

marina.mayor@uam.es (M. Mayor-Rocher).
0000-0002-4177-7559 (M. Mayor-Rocher); 0000-0002-9125-6225 (G. Martínez); 0000-0001-5952-6902 (E. Álvarez-Mellado), 0009-0004-3160-7734 (N. Melero), 0000-0003-2540-5234 (P. Reviriego).

Spanish in language models, indicating that, although models show good general performance, they still exhibit limitations in more specific grammatical and lexical aspects of the linguistic system. Consequently, evaluating these systems requires going beyond general comprehension or text generation tests and moving toward finer-grained analyses of specific linguistic phenomena. Previous works have studied LLM behavior on different grammatical phenomena, such as dative alternation (Yao et al. 2025), disjunctions (Boguraev et al. 2026), or sociolinguistic characteristics (Grieve et al. 2025). Recently, datasets in various languages have been published compiling minimal pairs or annotated with acceptability judgements (Warstadt et al. 2020; Bel et al. 2024a; Bel et al. 2024b; Jumelet et al. 2026). Accordingly, other works have inquired into LLM notions of grammaticality and whether their grammaticality preferences align with human judgements (Behzad et al. 2024; Qiu et al. 2024; Ide et al. 2025). In this work we analyze LLM preferences for Differential Object Marking (DOM).

DOM is a grammatical phenomenon whereby some languages explicitly mark certain direct objects depending on some of their semantic and referential properties, such as animacy, specificity or definiteness (Aissen, 2003), and on various semantic factors of the verb (telicity, affectedness and agentivity).

In Spanish this phenomenon manifests itself through the preposition *a* before the direct object, especially when it refers to human or highly animate entities. While in a sentence such as *Vi el coche* no marker appears, in *Vi a María* the presence of *a* indicates that the direct object is human. However, this behavior is not completely categorical: DOM is a gradient, context-sensitive phenomenon (Torrego, 1999; Leonetti, 2004; Rodríguez-Mondoñedo, 2007). In Spanish, moreover, its distribution shows dialectal variation and ongoing change processes (Gómez Seibane and Álvarez Morera, 2022), which makes it particularly interesting for LLM evaluation. To the best of the authors' knowledge, no previous work has specifically addressed LLM preferences for DOM in Spanish.

In this context, this work analyzes how language models handle DOM in Spanish. It seeks to answer two main questions: (1) whether LLMs use DOM systematically in contexts where Spanish grammar requires it, and (2) whether these models are sensitive to the linguistic factors conditioning its appearance in optional contexts (*Busco (a) un ayudante que sepa inglés*).

2. Objectives and Hypotheses

To answer the questions raised, this study has as its main objective to analyze the use of DOM in the Spanish generated by LLMs. It seeks to determine:

- Whether models use DOM systematically in obligatory contexts, evaluating their ability to correctly reproduce the grammatical patterns required by Spanish.
- Whether language models show sensitivity to the linguistic factors (animacy, definiteness, telicity, affectedness, agentivity) conditioning DOM in optional contexts, comparing their behavior with the use observed in the literature for human speakers.

Two hypotheses are proposed. The first is that LLMs will reproduce DOM consistently in obligatory contexts, showing a high degree of grammatical correctness as a result of the training process. The second is that in optional contexts, models will show less sensitivity to the linguistic factors conditioning DOM, tending to generate more uniform or simplified patterns than those observed in human speakers.

3. Methodology

This section briefly presents the methodology used in the study, describing the datasets created, the models evaluated, and the procedure used.

3.1. Dataset

Two *datasets*[2] in multiple-choice test format were created to study the use of DOM in LLMs, a common and reproducible method for evaluating models (Guo et al. 2023) and for ranking them in leaderboards according to their performance on different tasks (Grandury et al. 2025).

Since five semantic categories were to be evaluated (Table 1), both datasets are composed of 32 sentences balanced for each category, designed as a binary category matrix in which each feature is coded as either present (1) or absent (0). The first dataset was designed to observe LLM performance in relation to general

[2] https://zenodo.org/records/20628671

DOM use. Each question presents two possible options (A and B), one of which corresponds to the grammatically appropriate use of the direct object marker. The second dataset was constructed by selecting only those sentences in which DOM is optional. For these sentences, both answer options (A and B) can be considered grammatically correct. This set allows analysis of model preferences in variation contexts, where grammar admits both the presence and absence of the marker, and observes whether LLMs show systematic tendencies toward one of the two alternatives. The semantic categories analyzed are summarized in Table 1.

Table 1
Semantic categories analyzed. Examples from Torrego (1999).

Category	Explanation	Example
Animacy/Agent	Animate (human/animal) or reactive referent.	*Encontraron a un montañero. / El ácido afecta (a) los metales.*
Definiteness	Specific and known (definite) or non-specific referent	*Trajeron al policía con ellos.*
Telicity	Event with a defined endpoint	*María insultó a un compañero.*
Affectedness	Object undergoes a change as result of the action	*Golpearon a un extranjero.*
Agentivity	Subject acts with intention, control and volition	*Este abogado escondió a muchos prisioneros / Esta montaña escondió (*a) muchos prisioneros.*

The *prompt* used for both tests was the same in order to maintain consistent experimental conditions between the two datasets. In each case, models were asked to select the option they considered most natural among the alternatives provided. The prompt used was:

"Complete the following sentence in the way that feels most natural to you. Stick to the given options only."

In this way, models were asked to choose between the two proposed options (A and B) based on their linguistic preference, without generating additional text or reformulating the sentence. This format allows for a direct comparison of model choices regarding the presence or absence of DOM.

Furthermore, all sentences used in the tests were manually selected, produced and reviewed by three expert linguists, guaranteeing the quality and relevance of the items included in the datasets. The creation and review process involved multiple rounds of discussion, during which any sentences that raised doubts or disagreements were revisited and debated on successive occasions until a consensus was reached and the dataset was refined accordingly.

3.2. Models

To obtain responses, 17 state-of-the-art LLMs representative of the state of the art at the beginning of 2026 were selected. The selection was made following criteria of architectural diversity and market representativeness, in order to obtain a global overview of DOM handling in current artificial intelligence. It should also be noted that no models specifically trained in Spanish were selected, as the aim was to obtain a general overview of DOM handling across the models currently available on the market.

First, models from the most used families were included, such as GPT-5 (OpenAI), Gemini 3.1 (Google), and Claude 4.5/4.6 (Anthropic). These "frontier" models allow evaluation of the upper bound of grammatical competence in Spanish. Second, a group of Chinese-developed models was incorporated—DeepSeek-v3.2, Qwen 3.5 (Alibaba), GLM-5, and Kimi-k2.5—to observe whether there are biases or differences in the treatment of Romance languages in models whose main training may differ from the Western standard. Finally, two models from the Mistral family were included to represent European models.

Within each developer and model family, versions of different sizes and capacities were included (Pro, Mini, Lite or Nano versions), which will allow analysis of whether parameter

scale and model optimization influence sensitivity to the linguistic factors governing DOM. The details of the model versions evaluated are summarized in Table 2.

Table 2
Models evaluated[3].

Developer	Model
OpenAI	gpt-5, gpt-5-mini, gpt-5-nano, gpt-5.4
Google	gemini-3.1-pro-preview, gemini-3.1-flash-lite-preview
Anthropic	claude-opus-4.6, claude-sonnet-4.6, claude-haiku-4.5
DeepSeek	deepseek-v3.2
Alibaba	qwen3.5-397b-a17b, qwen3.5-35b-a3b, qwen3.5-plus-02-15
Z-AI	glm-5
Moonshot ai	kimi-k2.5
Mistral	mistral-large-2512, mistral-small-creative

3.3. Implementation

During implementation, the standard procedure for language model evaluation was followed. The prompts and questions previously defined in an Excel file were extracted and sent to each model via the OpenRouter API, which acts as an intermediary between providers, avoiding the particularities derived from each platform. Specifically, models were queried directly with the question without providing task examples, known in the literature as *zero-shot* (Brown et al. 2020). This decision was taken because examples can introduce biases in responses. Moreover, in our case they are not necessary, since the task is sufficiently clear for current models. As is customary in model evaluation, temperature was set to 0 to facilitate reproducibility of results (Holtzman et al. 2019).

For the analysis of responses, an automatic process was carried out to extract the chosen option and enable statistical analysis. Each question was sent independently, so that no question influenced the rest in any way. There were cases for which models gave more extensive explanations in addition to the option. In these cases this text was also reviewed in order to verify and understand the reasons for their choice. To avoid any extraction issues, all results were also manually verified. Ambiguous or contradictory responses were resolved through this manual review process.

The questions, prompts, results, and the code used are available online for public access and replication.

4. Discussion of Results

The results obtained for each of the datasets are presented and discussed below.

4.1. Dataset 1: Obligatory DOM

In this first experiment, the performance of the different models in the use of DOM was evaluated. Figure 1 presents the accuracy results obtained on this first dataset. It can be observed that most models achieved perfect accuracy, while three of them recorded a single error. These results suggest that, in general terms, this is a relatively simple task for this type of system.

The only exception corresponds to the creative writing model from Mistral AI, which presented five errors, achieving an accuracy around 80%. Unfortunately, no detailed public information is available for this model; therefore, it is difficult to draw firm conclusions. However, it is possible that this creative approach influences the results, favoring freer constructions that could lead to this type of error.

Carrying out a more detailed analysis of the errors, no common patterns are generally observed among the failures. However, interestingly in the case of the sentence: *Juan acaricia lentamente ____. A) el gato B) al gato*, two models from the same family, GPT-5.4 and GPT-5 Nano, did make the same error.

4.2. Dataset 2: Optional DOM

Once the good performance of these models on obligatory DOM had been confirmed, the decision was made to analyze behavior when DOM is optional. In addition, the different semantic categories explained in Table 1 were considered.

[3] The exact OpenRouter identifiers follow the {developer}/{model} format (accessed March 17, 2026)

Figure 2 shows the models' preference for the presence or absence of the marker in their choices. Two groups are apparent: on the one hand, models that clearly prefer the presence of DOM; and on the other, models that fall at an intermediate point. This latter group is made up of the GPT models and the largest Gemini model. Interestingly, Claude Opus from Anthropic, the model most similar to these, does not obtain a similar result and DOM use clearly predominates. Furthermore, it is interesting that in these models, with the exception of GPT-nano, as model size decreases, preference for DOM use increases, as can be seen in models such as GPT-5 Mini, Gemini 3.2 Flash, or Claude Haiku. The latter tends to always use DOM. On the other hand, the Mistral Creative model, which in the previous test had obtained the worst accuracy, does not show results very different from those of Mistral Large. Finally, models from Chinese companies do not show a different functioning from Western models: their results are similar to those observed in GPT-5 and Gemini, except for DeepSeek, which more closely approximates the behavior of Anthropic models.

On the other hand, Figures 3–7 show these same results considering the different semantic categories. Among them, animacy is the variable that presents the clearest relationship with DOM use. As observed in Figure 3, when the sentence does not present animacy, models tend to prefer the absence of DOM, except for Haiku, due to its strong bias toward its use, while when animacy is present, most opt to employ it, as do real Spanish speakers (Torrego, 1999).

Other variables show a smaller influence. Agentivity and definiteness also affect the choice, although in a more moderate way: when these features are absent, the absence of DOM predominates, while their presence favors its use in most models. Finally, in the case of affectedness and telicity, the results are not so clear. Although many models follow the same general tendency, some exceptions appear, such as in the GPT-5, GPT-5.4, and Gemini models for affectedness, or in models such as GPT-5 Nano, GPT-5.4, and the Qwen models for telicity.

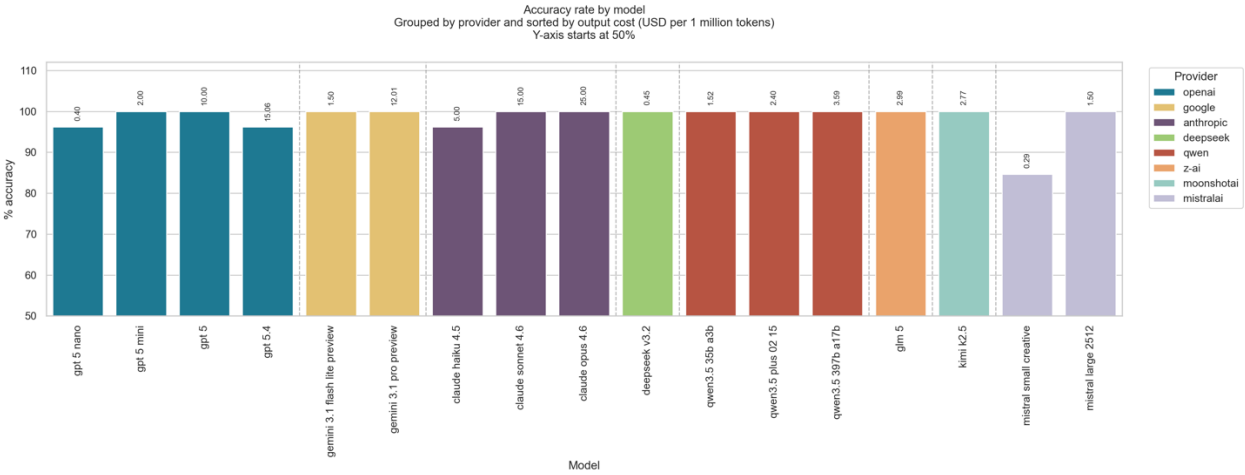

Figure 1. Accuracy of models in the use of DOM in obligatory contexts; above the bars, cost of each model in USD per million tokens. (First Experiment)

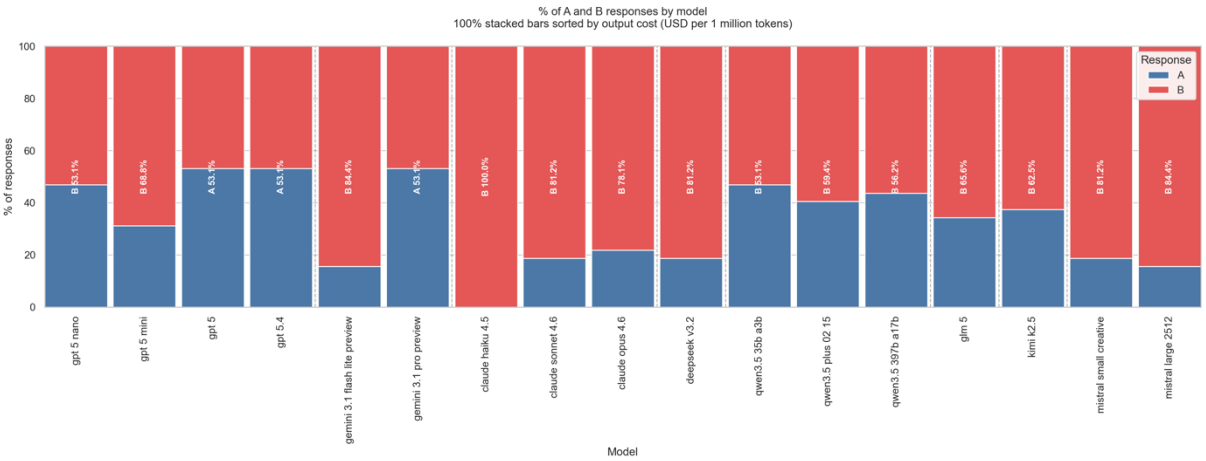

Figure 2. Percentage of choices for options A (no DOM) and B (DOM) by model in the optional DOM benchmark.

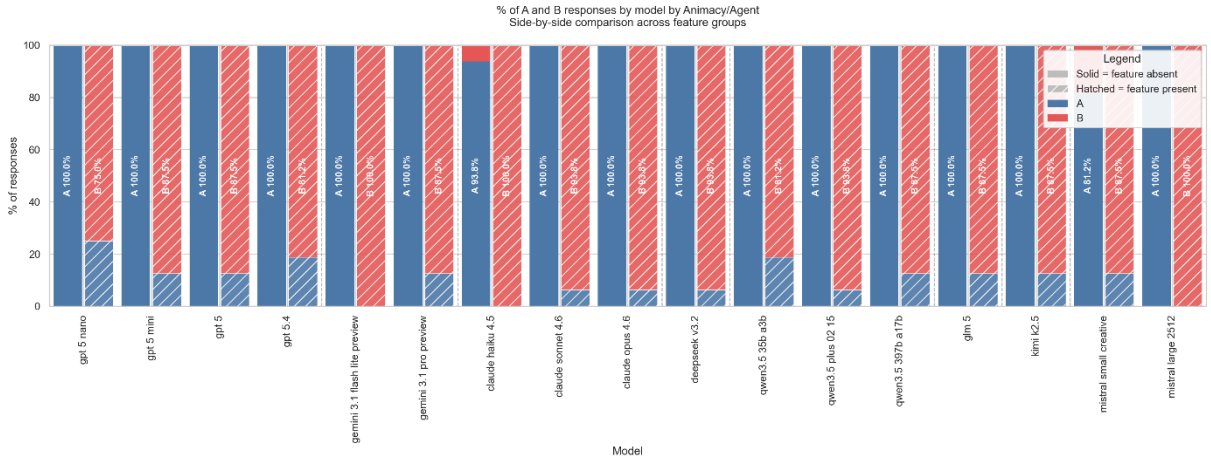

Figure 3. Percentage distribution of responses (A/B, no DOM/DOM) by model according to the animacy feature.

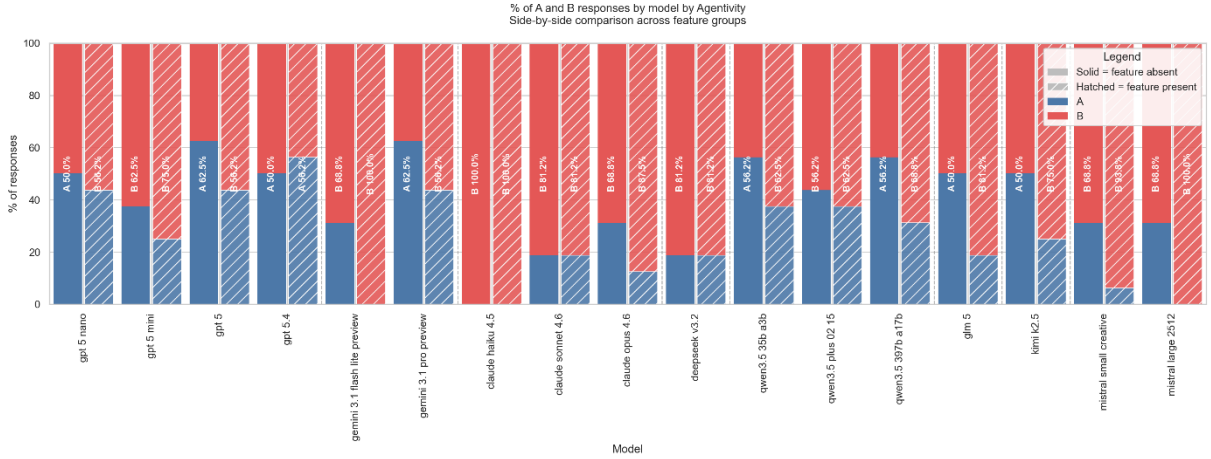

Figure 4. Percentage distribution of responses (A/B, no DOM/DOM) by model according to the agentivity feature.

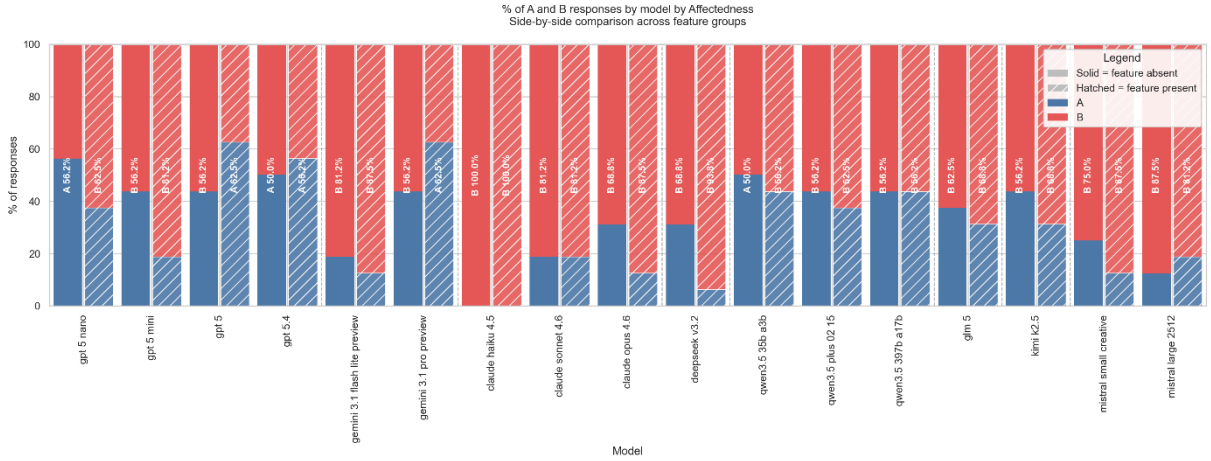

Figure 5. Percentage distribution of responses (A/B, no DOM/DOM) by model according to the affectedness feature.

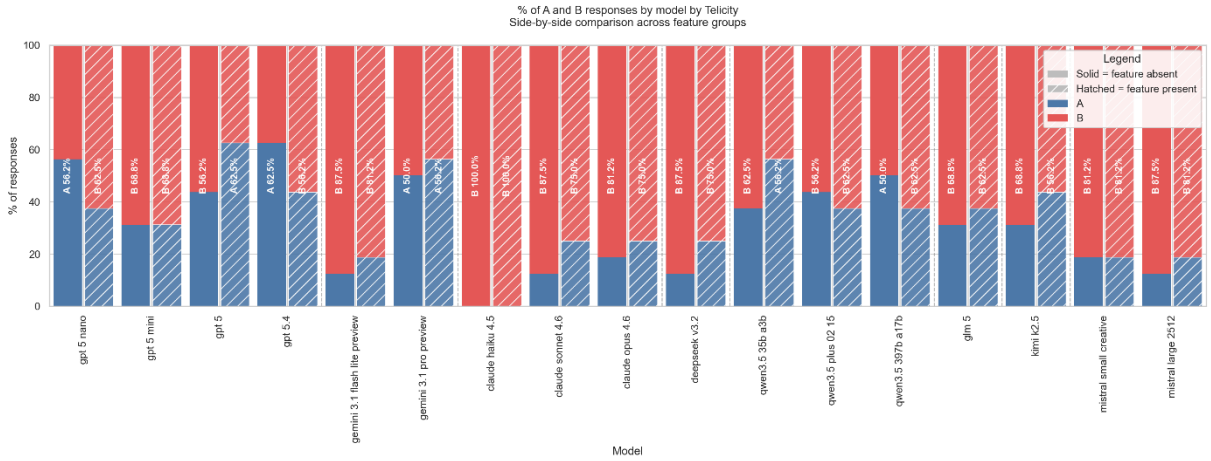

Figure 6. Percentage distribution of responses (A/B, no DOM/DOM) by model according to the telicity feature.

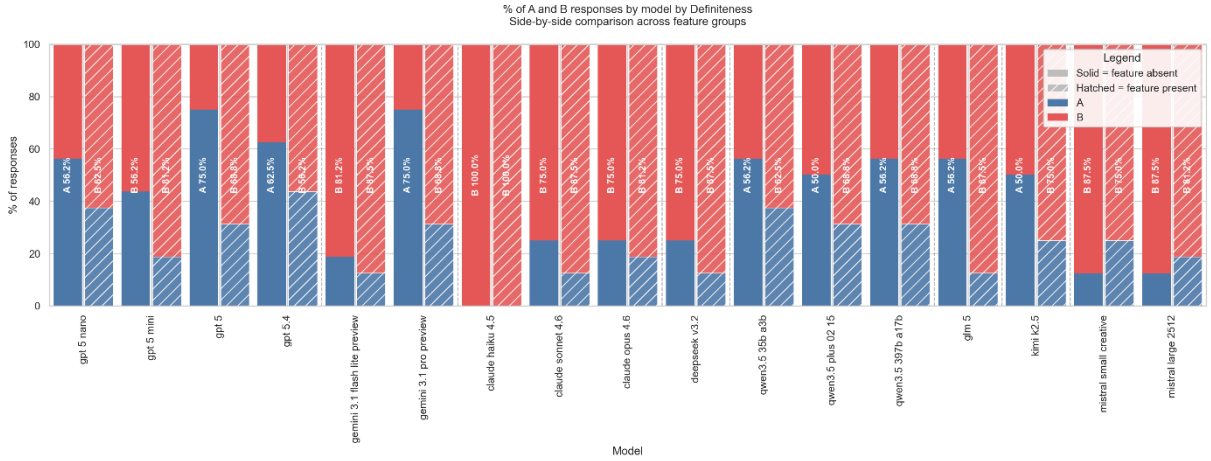

Figure 7. Percentage distribution of responses (A/B, no DOM/DOM) by model according to the definiteness feature.

5. Conclusions

The results of this work allow several fundamental conclusions to be drawn about the behavior of LLMs with respect to DOM in Spanish.

First, the first research hypothesis is confirmed: current LLMs present a high degree of grammatical correctness in contexts where DOM use is obligatory. Most evaluated models achieved perfect accuracy on the first dataset, suggesting that basic normative structures have been successfully internalized during training. The few exceptions, such as the Mistral Creative model, could be due to a prioritization of narrative fluency over grammatical rigor.

Second, regarding the second hypothesis, model behavior becomes significantly more inconsistent in optionality contexts. LLMs do not exactly replicate the contextual sensitivity of human speakers but show divergent preferences. While some models maintain a balance between the presence and absence of the marker, others present a marked bias toward constant use of the preposition "*a*".

Third, the analysis by semantic categories reveals that animacy is the factor that most clearly conditions the choices of the models. When the referent is animate, most opt to use DOM; when it is not, they tend to omit it. However, other factors such as definiteness, agentivity, telicity, and affectedness show a much more moderate and variable influence depending on the model analyzed.

Finally, a pattern related to model scale is identified: smaller versions (such as the "Mini" or "Haiku" variants) tend to simplify the phenomenon by increasing the preference for DOM use compared to their larger versions.

These findings underscore that, although LLMs are powerful tools for text generation, they still have limitations in capturing the gradient and complex nature of specific linguistic phenomena of the Spanish language.

6. Limitations

This study presents several limitations that should be considered in order to contextualize the results obtained:

- Scope of dialectal varieties: although it is acknowledged that DOM presents dialectal variation and ongoing change processes, the datasets focused on general semantic categories without delving into specific differences between varieties of Spanish.
- Restriction of the evaluation format: the research was based exclusively on multiple-choice tests with two alternatives (A and B). This format, while useful for statistical analysis, does not necessarily reflect model behavior in free text generation tasks or real communicative contexts.
- Opacity in model architecture: 17 state-of-the-art models were evaluated, but the lack of detailed public information about some of them, such as Mistral AI's creative writing model, makes it difficult to draw definitive conclusions about the exact causes of their errors or biases.
- Prompting methodology: the study used a zero-shot configuration to avoid example-induced biases. However, results could vary if few-shot techniques were employed or if the prompt used to request the natural response were modified.
- Interaction of semantic factors: the analysis was conducted evaluating sensitivity to isolated features such as animacy, definiteness, or agentivity predominantly. In real language use, these factors typically interact in complex ways, constituting a gradient phenomenon difficult to fully capture in a controlled experimental setting.

Acknowledgements

This work was supported by the FUN4DATE (PID2022-136684OB-C22) and SMARTY (PCI2024-153434) projects funded by the Spanish Agencia Estatal de Investigación (AEI) 10.13039/501100011033, by TUCAN6-CM (TEC-2024/COM-460), funded by CM (ORDEN 5696/2024) and by the Chips Act Joint Undertaking project SMARTY (Grant no. 101140087).

References

[1] Aissen, J. (2003). Differential object marking: Iconicity vs. economy. Natural Language & Linguistic Theory, 21(3), 435–483. https://doi.org/10.1023/A:1024109008573.

[2] Behzad, S., Zeldes, A., & Schneider, N. (2024). To ask LLMs about English grammaticality, prompt them in a different language. In Findings of the ACL: EMNLP 2024 (pp. 15622–15634).

[3] Bel, N., Punsola, M., & Ruiz-Fernández, V. (2024a). Catcola, catalan corpus of linguistic acceptability. Procesamiento del Lenguaje Natural, 73, 177–190.

[4] Bel, N., Punsola, M., & Ruiz-Fernández, V. (2024b). EsCoLA: Spanish corpus of linguistic acceptability. In Proceedings of LREC-COLING 2024 (pp. 6268–6277).

[5] Boguraev, S., Yao, Q., and Mahowald, K. (2026). France or Spain or Germany or France: A Neural Account of Non-Redundant Redundant Disjunctions. arXiv:2602.23547.

[6] Brown, T., et al. (2020). Language models are few-shot learners. Advances in neural information processing systems 33: 1877–1901.

[7] Gómez Seibane, S. y Álvarez-Morera, G. (2022). El marcado diferencial de objeto en el español del siglo XIX en Cataluña. Revista de filología española, 102(1), 87–109.

[8] Grandury, M., Aula-Blasco, J., Falcao, J., et al. (2025). La Leaderboard: A Large Language Model Leaderboard for Spanish Varieties and Languages of Spain and Latin America. arXiv:2507.00999.

[9] Grieve, J., et al. (2025). The sociolinguistic foundations of language modeling. Frontiers in Artificial Intelligence, 7, 1472411.

[10] Guo, Z., et al. (2023). Evaluating Large Language Models: A Comprehensive Survey. arXiv:2310.19736.

[11] Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. arXiv:1904.09751.

[12] Ide, Y., et al. (2025). How to make the most of LLMs' grammatical knowledge for acceptability judgments. In Proceedings of NAACL 2025: HLT (Volume 1: Long Papers) (pp. 7416–7432).

[13] Jumelet, J., Weissweiler, L., Nivre, J., & Bisazza, A. (2026). Multiblimp 1.0: A

massively multilingual benchmark of linguistic minimal pairs. Transactions of the ACL, 14, 193–216.
[14] Leonetti, M. (2004). Specificity and differential object marking in Spanish. Catalan Journal of Linguistics, 3, 75–114.
[15] Mayor-Rocher, M., et al. (2025). TELEIA: A Spanish Language Dataset for Evaluating Artificial Intelligence Models. Data in Brief.
[16] Muñoz-Basols, J., Palomares Marín, M. y Moreno Fernández, F. (2024). El Sesgo Lingüístico Digital (SLD) en la inteligencia artificial. Lengua y Sociedad, 23(2), pp. 623–647.
[17] Qiu, Z., Duan, X., & Cai, Z. (2024). Evaluating grammatical well-formedness in LLMs. In Proceedings of the workshop on cognitive modeling and computational linguistics (pp. 189–198).
[18] Rodríguez-Mondoñedo, M. (2007). The syntax of objects: Agree and differential object marking. Doctoral dissertation, University of Connecticut.
[19] Shapira, P. (2024). "Delving into delve." https://pshapira.net/2024/03/31/delving-into-delve/
[20] Torrego, E. (1999). El complemento directo preposicional. En I. Bosque & V. Demonte (eds.), Gramática descriptiva de la lengua española (vol. 2, pp. 1779–1805). Madrid: Espasa Calpe.
[21] Warstadt, A., et al. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. Transactions of the ACL, 8, 377–392.
[22] Yao, Q., Misra, K., Weissweiler, L., & Mahowald, K. (2025). Both direct and indirect evidence contribute to dative alternation preferences in language models. arXiv:2503.20850.