

Assessing the Comprehension of Regional Colombian Idioms by Language Models

Santiago Bobadilla Suarez
Universidad de los Andes
s.bobadilla@uniandes.edu.co

Juan Diego Osorio Cardenas
Universidad de los Andes
jd.osorioc1@uniandes.edu.co

Maria Alejandra Pinzon Roncancio
Universidad de los Andes
ma.pinzonr1@uniandes.edu.co

Ignacio Chaparro Rodriguez
Universidad de los Andes
i.chaparro@uniandes.edu.co

Ruben Francisco Manrique
Universidad de los Andes
rf.manrique@uniandes.edu.co

Abstract

This study queries 23 large language models (LLMs) and reports analyses for 22 models on Colombian idioms and regionally marked expressions. We introduce an evaluation set derived from Colombian lexicographical sources and assess models on three tasks: positive-class idiom recognition, definition generation, and context-based paraphrase generation. Automatic metrics show substantial variation across models, with Gemini-3-Flash-Preview achieving the highest average score on the meaning-oriented tasks. Human evaluation of four selected models qualifies these rankings: the best Task 2 definition accuracy reached only 58.57%, while Task 3 accuracy was higher for all four models. Exploratory regional analysis of the tagged subset shows uneven performance across Colombian macro-regions, with lower results for expressions from the Llanos and high Caribbean regions. These findings suggest that current LLMs only partially capture Colombian idiomatic meaning and motivate more regionally balanced resources and evaluation protocols for Spanish.

1 Introduction

Large language models (LLMs) are increasingly used to draft and revise professional text (Liang et al., 2025), but their performance remains uneven for linguistic phenomena that require non-literal interpretation. Idioms exemplify this challenge: their meanings often cannot be derived from the literal meanings of their component words (Chakrabarty et al., 2022). Recent evaluations show that prompted LLMs remain imperfect on idiomaticity tasks, especially when compared with task-specific fine-tuned models (Phelps et al., 2024).

This challenge is especially relevant in Spanish, where idiomatic meaning intersects with regional variation. Mayor-Rocher et al. (2025) show that LLMs do not identify Spanish varieties equally

well, with Peninsular Spanish performing best across the evaluated models. This pattern aligns with broader concerns about Digital Linguistic Bias in Spanish LLMs (Muñoz-Basols et al., 2024).

This study assesses LLM comprehension of Colombian idioms using an evaluation set drawn from DICOL and the *Brief Dictionary of Colombianisms* (Instituto Caro y Cuervo, 2024; Academia Colombiana de la Lengua, s.f.). We query 23 LLMs and report final analyses for 22 models on positive-class idiom recognition, definition generation, and context-based paraphrase generation, complementing automatic metrics with expert evaluation of sampled outputs.

2 Related Work

To our knowledge, no prior benchmark evaluates LLMs on Colombian idioms. Existing idiomaticity evaluations focus mainly on English resources, with limited multilingual coverage, and show that large prompted LLMs remain behind task-specific fine-tuned encoder models on several idiomaticity tasks (Phelps et al., 2024). This leaves open whether LLMs can interpret idiomatic expressions tied to a specific regional variety of Spanish.

This gap matters because Spanish LLM evaluation must account for dialectal variation. Mayor-Rocher et al. (2025) show that LLMs do not identify Spanish varieties equally well, while Muñoz-Basols et al. (2024) describe Digital Linguistic Bias as a problem that can disadvantage less visible regional Spanish varieties relative to dominant ones. Spanish-specific models such as MarIA expand NLP resources for Spanish (Gutiérrez-Fandiño et al., 2021), but they do not remove the need for targeted evaluation of regional expressions in Colombian Spanish.

Idiomaticity also complicates both representation and evaluation. He et al. (2025) show that static and contextual word representations can as-

sign high similarity without accurately capturing idiomatic meaning. This motivates evaluations that compare model outputs with reference definitions and supplement automatic similarity metrics with human evaluation. Related figurative-language work shows that models benefit from contextual and knowledge-based cues, but still fall short of human interpretation (Chakrabarty et al., 2022). Together, these findings motivate the evaluation design used in this work.

3 Dataset

We build the dataset from two Colombian lexicographical resources: DICOL (*Diccionario de colombianismos*; Instituto Caro y Cuervo, 2024) and the fourth edition of the *Brief Dictionary of Colombianisms* (BDC; Academia Colombiana de la Lengua, s.f.). From these sources, we compile idioms, idiomatic expressions, and regionally marked Colombian lexical items, preserving source definitions, examples, and metadata.

A form is an idiom or regionally marked expression as listed in the source. An entry pairs one form with one of its meanings. A form with multiple meanings therefore yields multiple entries and may also appear in both sources. Table 1 illustrates this distinction with two forms represented by five entries.

The final dataset contains 8,129 entries: 5,702 from DICOL and 2,427 from BDC. Each entry has four fields: word, meaning, example, and region. The dataset includes 5,286 unique forms; 61.5% of entries include contextual examples, while 38.5% lack them. Regional metadata are available for 557 entries, corresponding to 504 unique region-tagged forms. Most forms (89.2%) have one or two meanings, and single-word forms account for 93.3% of the dataset.

Three subsets support the evaluation. Tasks 1 and 2 use all 5,286 unique forms. Task 3 uses the 3,916 unique forms with contextual examples. The regional analysis uses the region-tagged entries.

Regional coverage is uneven. Counting each form once per macro-region, the Andean region contains the largest number of tagged forms (320), followed by the Caribbean and Insular (162), Pacific (107), and Orinoco (39) regions. In total, 113 forms appear in more than one macro-region. These imbalances limit regional comparison, but the dataset provides a consolidated resource for evaluating LLM comprehension of Colombian id-

iomatic and regionally marked Spanish.

4 Methodology

We queried 23 LLMs (Appendix A, Table 3) and report final analyses for 22 models after excluding Perplexity Sonar because access to external information could not be ruled out. We evaluate the remaining 22 models through three independent zero-shot tasks designed to probe unassisted comprehension of Colombian idioms and regionally marked expressions. We store all outputs in JSON format for transparency and reproducibility¹. We use one fixed Spanish prompt template per task across models, without model-specific prompt tuning. To test whether results were driven by Spanish prompt comprehension, we also run a prompt-language robustness check using English translations of the same templates, while keeping models, inputs, required outputs, and metrics fixed. Responses that request clarification, violate the required format, or provide irrelevant content are treated as errors; the corresponding metric is set to zero (Appendix B.1–B.3).

Task 1 measures whether each model recognizes a source expression as a Colombian idiom or idiomatic expression. We use the 5,286 unique forms in the curated dataset and prompt each model in Spanish to return a strict JSON response (Appendix B.1). Because all inputs derive from lexicographical sources, the task contains only positive examples and therefore cannot assess discrimination between idioms and non-idioms. We therefore report positive-class coverage, the proportion of forms a model classifies as idiomatic. A system that always answers *Sí* would score 1.0, above every model in our pool. We report format validity separately.

Task 2 asks each model to generate a brief definition in Colombian Spanish for each form, with a maximum length of 60 words (Appendix B.2). We compare model outputs with source definitions using BERTScore (Zhang et al., 2019), computed with BETO (Cañete et al., 2020) and SciBETO (Flaglab, 2025); Sentence-BERT cosine similarity (Reimers and Gurevych, 2019); and chrF (Popović, 2015).

Task 3 evaluates context-based paraphrase generation. For forms with contextual examples, each model receives the expression and its example sen-

¹ Repository https://github.com/SBobi/Modismos_EMNLP_2026

Form	Entry	Meaning	Example	Region	Source
<i>dar casquillo</i>	1	provocar a una persona, incitarla a pelear mediante insultos o burlas	—	Antioquia, Cundinamarca, Santander	BDC
	2	burlarse de una persona	—	Santander	BDC
	3	despertar celos en alguien	—	Cundinamarca, Nariño, Valle del Cauca	BDC
<i>poner conejo</i>	1	engañar, incumplir las promesas hechas	[...] el gobierno le puso conejo al congreso con las reformas a la seguridad social [...]	—	BDC
	2	abandonar un establecimiento público sin pagar el servicio que se ha recibido	[...] tiene que ir a ponerle sus conejos a otra más pendeja que yo	—	BDC

Dashes mark fields the source leaves empty; [...] marks an example we shortened. The sources label regions by department. We map these labels to macro-regions for the regional analysis.

Table 1: Two forms and their five entries. *Dar casquillo* shows that the regional labels attached to a form vary from one entry to the next.

tence, then generates a short paraphrase that preserves the expression’s meaning in context and defines that paraphrase (Appendix B.3). We apply the same automatic metrics by comparing the generated paraphrase definition with the source definition. This comparison captures whether the model recovered the source meaning, not whether the paraphrase is a valid substitute for the expression in context.

To complement automatic evaluation, four Colombian experts assessed Tasks 2 and 3 for four selected models (Appendix F). Each expert evaluated one model. We drew one common sample per task and evaluated all four models on the same sampled items (Table 9). The annotation interface displayed neither the model nor the provider. We conducted no calibration phase and collected no overlapping annotations. For Task 2, experts assessed whether each generated definition preserved the source meaning. For Task 3, they evaluated two criteria: Paraphrase Definition, which assesses whether the definition generated for the paraphrase preserves the source meaning, and Paraphrase Appropriateness, which assesses whether the generated paraphrase is a valid substitute for the target expression in the provided example and preserves its intended meaning in context. Each criterion was assigned 1 when satisfied and 0 otherwise.

5 Results

5.1 Automatic Evaluation

Results for Task 1 are reported in Appendix D, Table 5. Because all inputs are drawn from lexicographical sources, Task 1 is a positive-class recog-

nition task rather than a balanced idiom/non-idiom detection task. Ten of 22 LLMs recognized more than half of the forms as idiomatic. Gemma-2-27B-it achieved the highest coverage (0.862), followed by Llama-3.3-70B-Instruct (0.821), Command-R (0.740), and Llama-4-Maverick (0.735). Every model therefore falls below the always-positive baseline. Format-validity rates are reported in Appendix E, Table 6.

Task 1 captures positive-class recognition, but not whether models can recover idiomatic meaning. We therefore evaluate definition generation (Task 2) and context-based paraphrase generation (Task 3) using similarity-based automatic metrics, summarized in Figure 1. The ranking changes under these meaning-oriented tasks: Gemini-3-Flash-Preview achieved the highest average score, followed by Claude-Sonnet-4, GPT-4.1, GPT-5-mini, and DeepSeek-V3.2. These rankings measure similarity to reference definitions under the automatic metrics described above, but should not be read as direct evidence of idiomatic understanding.

Because these tasks are prompted in Spanish, we also test whether model rankings are sensitive to prompt language. The prompt-language robustness check shows stable rankings across Spanish and English prompts. Across the 22 reported models, the two conditions show high agreement (Pearson $r = 0.965$, Spearman $\rho = 0.936$), with a mean absolute score change of 0.0077 and a mean absolute rank change of 1.64 (Appendix C, Table 4).

5.2 Human Evaluation

Human evaluation was conducted on approximately one-third of the outputs for four selected

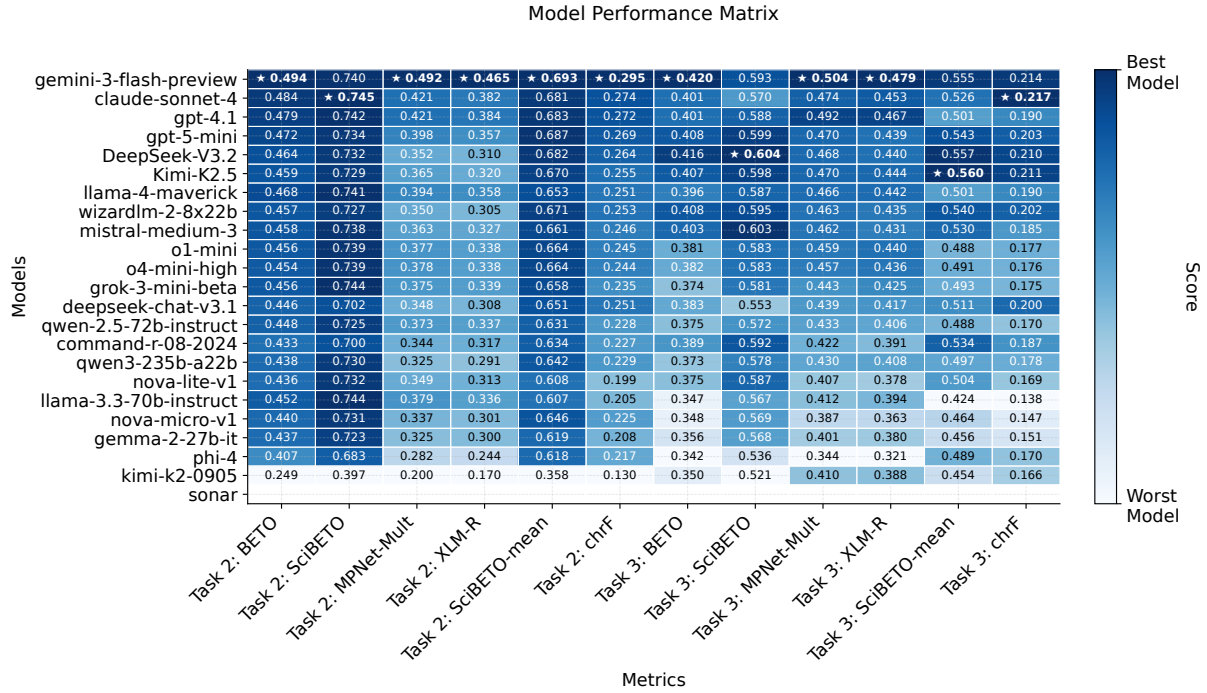


Figure 1: Automatic performance on Tasks 2 and 3. The twelve columns are six similarity metrics per task. We order models by their average score, defined as the unweighted mean of those twelve values, excluding Task 1. Displayed values are raw scores; shading is min-max normalized within each column, so color indicates rank within a metric rather than absolute performance. Stars mark the best-performing model for each metric. Invalid or irrelevant responses receive a score of zero.

Model	Task 2 Definition	Task 3 Paraphrase Definition	Task 3 Paraphrase Appropriateness
Claude	43.56	75.92	80.23
GPT-4.1	35.67	71.57	68.03
Gemini	58.57	75.98	75.10
Llama-4	14.38	59.68	67.63

Table 2: Human-evaluation accuracy for selected models. Values are percentages.

models: Claude-Sonnet-4, GPT-4.1, Gemini-3-Flash-Preview, and Llama-4-Maverick. Table 2 summarizes the results; sample sizes appear in Appendix F, Table 9.

Human evaluation contextualizes the rankings produced by automatic metrics. Gemini-3-Flash-Preview achieved the highest Task 2 definition accuracy. Task 3 Paraphrase Definition accuracy was higher for all four models, but the two tasks differ in their item sets, requested outputs, sample sizes, and human criteria. This difference therefore cannot be attributed to context alone. Across the four models, Task 3 Paraphrase Definition accuracy exceeds Task 2 Definition accuracy by 17 to 45 percentage points.

6 Discussion

Human evaluation shows that Colombian idiom comprehension remains limited, even among the evaluated models. The best Task 2 result, achieved by Gemini-3-Flash-Preview, reached only 58.57% definition accuracy, while the other models performed substantially lower. The human results partially qualify the automatic rankings: Gemini-3-Flash-Preview remains the strongest model on Task 2, but the ordering differs across the human-evaluation criteria. This shows that similarity-based screening is only directionally informative. In particular, high automatic similarity scores do not necessarily correspond to human-judged semantic correctness.

Task 3 accuracies are higher for all four models, but this pattern should not be interpreted as evidence of robust idiom comprehension or as a direct effect of context. Tasks 2 and 3 differ in their item sets, requested outputs, sample sizes, and human criteria. The consistent difference is compatible with sentence-level cues helping models recover plausible meanings, but the current design does not isolate that effect. Nor does stronger Task 3 performance demonstrate stable internal knowledge of

Colombian idioms.

We also examined Task 2 and Task 3 performance across Colombian macro-regions and departments using the subset with geographic metadata. This analysis is exploratory because regional tags are sparse. The results therefore reflect only the tagged subset and should not be interpreted as a complete regional benchmark. We adjusted departmental estimates using Empirical Bayes shrinkage to account for sparsity.

As shown in Appendix G, Figure 2 (Task 2 definition accuracy) and Figure 3 (Task 3 paraphrase-definition agreement), performance is uneven across regions. Both maps show stronger results in the Pacific and parts of the Andean region, while expressions from the Llanos and parts of the Caribbean region remain more difficult. The Amazonian region lacks sufficient tagged data for robust evaluation. Overall, the results point to partial and uneven model competence, reinforcing the need for more regionally balanced Spanish evaluation resources.

7 Conclusion

We report results for 22 LLMs, drawn from an initial pool of 23, on Colombian idioms and regionally marked expressions across three tasks: positive-class idiom recognition, definition generation, and context-based paraphrase generation. Models showed substantial variation, and Colombian idiom comprehension remains limited even among the strongest systems.

Human evaluation qualifies the automatic rankings. Gemini-3-Flash-Preview achieved the highest Task 2 definition accuracy (58.57%), while Task 3 accuracy was higher across all four evaluated models. Because Tasks 2 and 3 differ in their item sets, requested outputs, sample sizes, and human criteria, this difference should not be attributed to context alone or interpreted as evidence of stable internal knowledge of Colombian idioms.

Exploratory regional analysis over the tagged subset shows uneven performance: expressions from the Llanos and parts of the Caribbean region show lower performance than those from the Pacific and parts of the Andean region, though these findings are bounded by sparse regional metadata.

This work contributes a benchmark for Colombian idiomatic and regionally marked Spanish and highlights the need for more regionally balanced resources, broader human validation of automatic

metrics, and evaluation protocols that account for dialectal diversity.

8 Limitations

Human evaluation covers four selected models rather than all 22 reported LLMs. Due to resource constraints, each of the four models was assessed by a single expert on the same sampled items, with no overlapping annotations across evaluators. This design does not allow us to measure inter-annotator agreement and may introduce evaluator-specific bias. We also conducted no calibration phase. Comparisons between automatic metrics and human judgments should therefore be interpreted with caution. The automatic metrics capture surface and embedding-level similarity rather than full semantic correctness. In Task 3, they measure recovery of the source meaning but do not directly assess whether the generated paraphrase is a valid substitute in context. Task 1 is also a positive-class recognition task and does not test false positives against non-idiomatic expressions.

Tasks 2 and 3 differ in their item sets, requested outputs, sample sizes, and human-evaluation criteria. Their performance difference therefore cannot be used to isolate the effect of contextual information.

Model access also remains partly opaque. To mitigate external-information leakage, we configured API calls with external tools and retrieval disabled when the provider supported this. Still, proprietary deployments and provider-specific changes may affect exact reproducibility. The zero-shot fixed-template design supports comparability but does not estimate best achievable performance under few-shot or dialect-aware prompting.

Finally, the evaluation set reflects its lexicographical sources. Some entries lack examples, and regional metadata are sparse and uneven across Colombian macro-regions. Regional findings should therefore be interpreted as evidence from the tagged subset, not as a complete account of Colombian dialectal variation.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Academia Colombiana de la Lengua. s.f. [Base de datos de la academia colombiana de la lengua](https://www.academiacolombianadelalengua.co/bdc/). [En línea]. Available: <https://www.academiacolombianadelalengua.co/bdc/>.
- José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.
- T. Chakrabarty, Y. Choi, and V. Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.
- Cohere Labs. 2024. C4ai command r 08-2024 model card. <https://huggingface.co/CohereLabs/c4ai-command-r-08-2024>.
- DeepSeek-AI. 2025a. Deepseek-v3.1 model card. <https://huggingface.co/deepseek-ai/DeepSeek-V3.1>.
- DeepSeek-AI. 2025b. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.
- Flaglab. 2025. Scibeto-large (version 48d915e). <https://huggingface.co/Flaglab/SciBETO-large>. Modelo de lenguaje para textos científicos en español.
- Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. Maria: Spanish language models. *arXiv preprint arXiv:2107.07253*.
- Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2025. *Investigating idiomaticity in word representations*. *Computational Linguistics*, 51:505–555.
- Instituto Caro y Cuervo. 2024. *Diccionario de colombianismos*. Lexicc / Instituto Caro y Cuervo. [En línea]. Available: <https://lexicc.caroycuervo.gov.co/diccionario/diccionario-de-colombianismos/>.
- Kimi Team. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.
- W. Liang, Y. Zhang, M. Codreanu, J. Wang, H. Cao, and J. Zou. 2025. The widespread adoption of large language model-assisted writing across society. *Patterns*, page 101366.
- M. Mayor-Rocher, C. Pozo, N. Melero, G. Martínez, M. Grandury, and P. Reviriego. 2025. It's the same but not the same: Do llms distinguish spanish varieties? *arXiv*. Online.
- Meta AI. 2024. Llama 3.3 70b instruct model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md.
- Meta AI. 2025. Llama 4 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md.
- Moonshot AI. 2026. Kimi-k2.5 model card. <https://huggingface.co/moonshotai/Kimi-K2.5>.
- J. Muñoz-Basols, M. d. M. Palomares Marín, and F. Moreno Hernández. 2024. El sesgo lingüístico digital (sld) en la inteligencia artificial: implicaciones para los modelos de lenguaje masivos en español. *Lengua y Sociedad. Revista de Lingüística Teórica y Aplicada*, 23(2):623–647.
- Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. *Sign of the times: Evaluating the use of large language models for idiomaticity detection*. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- WizardLM Team. 2024. Wizardlm-2-8x22b model card. <https://huggingface.co/microsoft/WizardLM-2-8x22B>.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Model Pool

Table 3 lists the 23 models we queried. We report parameter counts where documented by the provider, drawing on the corresponding technical reports and model cards (Abdin et al., 2024; Cohere Labs, 2024; DeepSeek-AI, 2025a,b; Gemma Team, 2024; Meta AI, 2024, 2025; WizardLM Team, 2024; Kimi Team, 2025; Moonshot AI, 2026; Qwen Team, 2024, 2025). Mixture-of-experts models show total and active parameters per token; dense models show a single count. We mark undisclosed sizes as n/r.

Provider	Model	Parameters
Amazon	nova-lite-v1	n/r
Amazon	nova-micro-v1	n/r
Anthropic	claude-sonnet-4	n/r
Cohere	command-r-08-2024	32B
DeepSeek	DeepSeek-V3.2	685B
DeepSeek	deepseek-chat-v3.1	671B / 37B
Google	gemini-3-flash-preview	n/r
Google	gemma-2-27b-it	27B
Meta	llama-3.3-70b-instruct	70B
Meta	llama-4-maverick	400B / 17B
Microsoft	phi-4	14B
Microsoft	wizardlm-2-8x22b	141B
Mistral AI	mistral-medium-3	n/r
Moonshot AI	Kimi-K2.5	1T / 32B
Moonshot AI	kimi-k2-0905	1T / 32B
OpenAI	gpt-4.1	n/r
OpenAI	gpt-5-mini	n/r
OpenAI	o1-mini	n/r
OpenAI	o4-mini-high	n/r
Perplexity	sonar [†]	n/r
Qwen	qwen-2.5-72b-instruct	72B
Qwen	qwen3-235b-a22b	235B / 22B
xAI	grok-3-mini-beta	n/r

[†]Accessed via a third-party endpoint; external-information access could not be ruled out. Excluded from final analyses.

Table 3: Model pool and documented parameter counts. Perplexity Sonar was accessed through a third-party endpoint and excluded from final analyses.

B Primary Spanish Prompt Templates

The prompts below are the Spanish templates used in the main evaluation. English translations were used only for the prompt-language robustness check (Appendix C).

B.1 Prompt 1: Positive-Class Idiom Recognition

Eres un clasificador experto en modismos y expresiones idiomáticas de Colombia.

Instrucciones:

1. Siempre responde ÚNICAMENTE con JSON válido y nada más.
2. Clasifica si la expresión dada es un modismo colombiano o no.

3. Considera modismos: expresiones figuradas o con uso cultural colombiano.

4. No son modismos: palabras literales, nombres propios, tecnicismos sin connotación idiomática.

5. Usa solo tu conocimiento interno (sin búsquedas externas).

6. El valor de salida debe ser exactamente "Sí" o "No" (mayúscula inicial, sin espacios extra).

INPUT (JSON):

```
{
  "modismo": "{{modismo}}"
}
```

FORMATO DE SALIDA (JSON estricto):

```
{
  "input": "{{modismo}}",
  "output": {
    "es_modismo": "<Sí o No>"
  }
}
```

B.2 Prompt 2: Definition Generation

Eres un modelo experto en modismos y expresiones idiomáticas colombianas.

Instrucciones:

1. Siempre responde ÚNICAMENTE con JSON válido y nada más.

2. Define el modismo dado con una sola oración, breve, clara y objetiva en español, máx. 60 palabras.

3. No incluyas ejemplos, sinónimos ni explicaciones adicionales.

4. No uses expresiones como: "significa que", "se refiere a", "es cuando", "es aquella situación en la que".

5. Usa solo tu conocimiento interno (sin búsquedas externas).

INPUT (JSON):

```
{
  "modismo": "{{modismo}}"
}
```

FORMATO DE SALIDA (JSON estricto):

```
{
  "input": "{{modismo}}",
  "output": {
    "definicion": "<definición breve en español formal>"
  }
}
```

B.3 Prompt 3: Context-Based Paraphrase Generation

Eres un modelo experto en modismos y expresiones idiomáticas colombianas.

Instrucciones:

1. Siempre responde ÚNICAMENTE con JSON válido y nada más.
2. Identifica el modismo en el ejemplo proporcionado.
3. Genera la PARÁFRASIS MÁS PLAUSIBLE que mantenga el MISMO sentido del modismo en el contexto dado.
4. La PARÁFRASIS MÁS PLAUSIBLE debe:
 - Ser una palabra o frase simple y directa.
 - Tener máximo 10 palabras.
5. Define la paráfrasis identificada.
6. No uses expresiones como: "significa", "se refiere a", "es cuando", "es aquella situación en la que".
7. Usa solo tu conocimiento interno (sin búsquedas externas).
8. Si desconoces el modismo, infiere la PARÁFRASIS MÁS PLAUSIBLE según el contexto del ejemplo.

INPUT (JSON):

```
{
  "modismo": "{{modismo}}",
  "ejemplo": "{{ejemplo}}"
}
```

FORMATO DE SALIDA (JSON estricto):

```
{
  "input": {
    "modismo": "{{modismo}}",
    "ejemplo": "{{ejemplo}}"
  },
  "output": {
    "parafrasis": "<parafrasis más plausible>",
    "definicion": "<definición breve de la paráfrasis>"
  }
}
```

C Prompt-Language Robustness

To assess whether results were sensitive to prompt language, we repeated the automatic evaluation using English translations of the Spanish templates and compared model rankings across both conditions.

Statistic	Value
Reported models compared	22
Pearson r	0.965
Spearman ρ	0.936
Mean abs. score change	0.0077
Mean abs. rank change	1.64

Table 4: Prompt-language robustness: Spanish vs. English conditions.

D Task 1: Positive-Class Coverage

Model	Classified idiomatic	Coverage
gemma-2-27b-it	4,557	0.862
llama-3.3-70b-instruct	4,341	0.821
command-r-08-2024	3,912	0.740
llama-4-maverick	3,884	0.735
gemini-3-flash-preview	3,496	0.661
mistral-medium-3	3,493	0.661
phi-4	3,297	0.624
deepseek-chat-v3.1	2,932	0.555
kimi-k2-0905	2,871	0.543
qwen-2.5-72b-instruct	2,813	0.532
Kimi-K2.5	2,635	0.498
claude-sonnet-4	2,627	0.497
DeepSeek-V3.2	2,525	0.478
gpt-4.1	2,230	0.422
qwen3-235b-a22b	2,171	0.411
nova-lite-v1	2,013	0.381
wizardlm-2-8x22b	1,704	0.322
nova-micro-v1	1,490	0.282
gpt-5-mini	1,101	0.208
grok-3-mini-beta	968	0.183
o4-mini-high	855	0.162
o1-mini	847	0.160

Table 5: Task 1 positive-class coverage by model. Denominator: 5,286 forms.

E Output-Format Validity

Invalid responses fell into three recurring patterns: plain prose instead of JSON, JSON objects with incorrect keys, and requests for clarification. These responses did not satisfy the required output format and were treated as errors, with the corresponding metric set to zero.

Model	Valid	Rate
claude-sonnet-4	5,281	0.999
command-r-08-2024	5,280	0.999
deepseek-chat-v3.1	5,144	0.973
DeepSeek-V3.2	5,283	0.999
gemini-3-flash-preview	5,273	0.998
gemma-2-27b-it	5,200	0.984
gpt-4.1	5,262	0.995
gpt-5-mini	5,285	1.000
grok-3-mini-beta	5,277	0.998
kimi-k2-0905	5,280	0.999
Kimi-K2.5	5,285	1.000
llama-3.3-70b-instruct	5,208	0.985
llama-4-maverick	5,277	0.998
mistral-medium-3	5,279	0.999
nova-lite-v1	5,270	0.997
nova-micro-v1	5,278	0.998
o1-mini	5,283	0.999
o4-mini-high	5,279	0.999
phi-4	5,274	0.998
qwen-2.5-72b-instruct	5,272	0.997
qwen3-235b-a22b	5,285	1.000
wizardlm-2-8x22b	5,251	0.993

Table 6: Output-format validity for Task 1. Denominator: 5,286 forms.

Model	Valid	Rate
claude-sonnet-4	5,285	1.000
command-r-08-2024	5,022	0.950
deepseek-chat-v3.1	5,060	0.957
DeepSeek-V3.2	5,284	1.000
gemini-3-flash-preview	5,216	0.987
gemma-2-27b-it	5,156	0.975
gpt-4.1	5,275	0.998
gpt-5-mini	5,285	1.000
grok-3-mini-beta	5,278	0.998
kimi-k2-0905	2,844	0.538
Kimi-K2.5	5,249	0.993
llama-3.3-70b-instruct	5,264	0.996
llama-4-maverick	5,270	0.997
mistral-medium-3	5,263	0.996
nova-lite-v1	5,211	0.986
nova-micro-v1	5,269	0.997
o1-mini	5,282	0.999
o4-mini-high	5,278	0.998
phi-4	4,952	0.937
qwen-2.5-72b-instruct	5,138	0.972
qwen3-235b-a22b	5,284	1.000
wizardlm-2-8x22b	5,236	0.991

Table 7: Output-format validity for Task 2. Denominator: 5,286 forms.

Model	Valid	Rate
claude-sonnet-4	3,736	0.954
command-r-08-2024	3,871	0.989
deepseek-chat-v3.1	3,615	0.923
DeepSeek-V3.2	3,911	0.999
gemini-3-flash-preview	3,822	0.976
gemma-2-27b-it	3,839	0.980
gpt-4.1	3,912	0.999
gpt-5-mini	3,902	0.996
grok-3-mini-beta	3,896	0.995
kimi-k2-0905	3,453	0.882
Kimi-K2.5	3,898	0.995
llama-3.3-70b-instruct	3,857	0.985
llama-4-maverick	3,881	0.991
mistral-medium-3	3,897	0.995
nova-lite-v1	3,877	0.990
nova-micro-v1	3,868	0.988
o1-mini	3,911	0.999
o4-mini-high	3,907	0.998
phi-4	3,616	0.923
qwen-2.5-72b-instruct	3,793	0.969
qwen3-235b-a22b	3,902	0.996
wizardlm-2-8x22b	3,868	0.988

Table 8: Output-format validity for Task 3. Denominator: 3,916 forms with contextual examples.

F Human Evaluation

Model	Task 2	Task 3
Claude-Sonnet-4	1,738	1,245
GPT-4.1	1,738	1,245
Gemini-3-Flash-Preview	1,738	1,245
Llama-4-Maverick	1,738	1,245

Table 9: Number of outputs reviewed in the human evaluation by model and task.

G Geographical Analysis

Figures 2 and 3 report department- and region-adjusted performance for the region-tagged subset. Because regional metadata are sparse and unevenly distributed, these maps should be interpreted as evidence from the tagged subset rather than as a complete regional benchmark. Estimates are adjusted using Empirical Bayes shrinkage to reduce instability in departments with few observations.

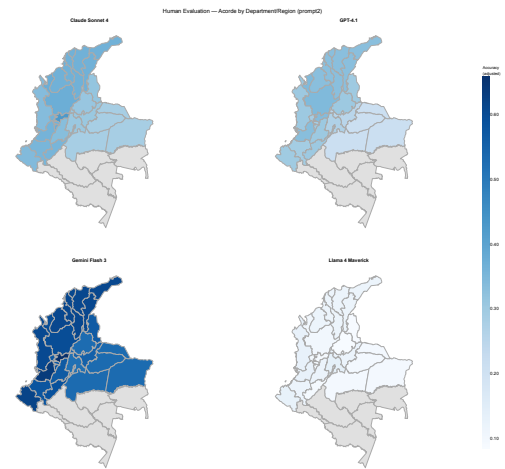


Figure 2: Department- and region-adjusted Task 2 definition accuracy across evaluated models for the region-tagged subset.

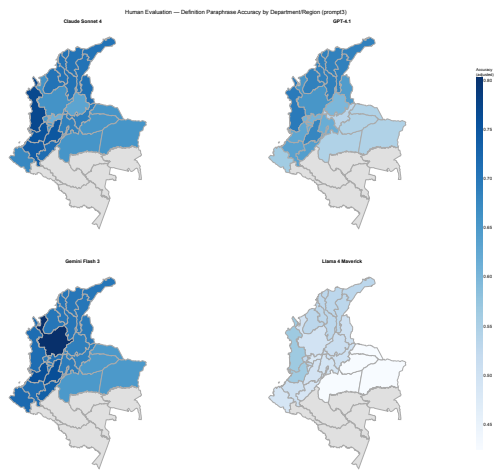


Figure 3: Department- and region-adjusted Task 3 paraphrase-definition agreement across evaluated models for the region-tagged subset.