EvaLex: Automatic Evaluation of Lexical Competence in Language Models

Anonymous ACL submission

Abstract

Existing methods for evaluating lexical competence in language models often rely on discriminative tasks or expensive, human-curated datasets. We introduce EVALEX, an automated framework to evaluate lexical competence in language models without gold labels. EVALEX employs a bidirectional approach: it first prompts a model to define a target word and subsequently tasks the model with retrieving that same word from its own generated definition in a separate inference session. We evaluate EVALEX on two Spanish datasets: CREA-10k-POS (7,784 words) and CREA-10k-Lemmas (4,594 lemmas), and we further validate against an expert-annotated DEA sample (100 words). We compare similarly sized LLMs (Llama, Mistral, Yi, Gemma) alongside Spanish-centric models (Gromenauer, Salamandra). EVALEX scores correlate strongly with expert manual annotations (Pearson $r = 0.8539$, Spearman $\rho = 0.8$) across the four models for which manual annotations were available, supporting its use as a proxy for human judgments. Broader validation across more models and languages remains an open priority. We also observe a clear frequency effect, with Spanish-specific models remaining more robust on low-frequency terms. Overall, EVALEX provides a scalable and low-cost method for model selection and for diagnosing lexical gaps in low-resource and cross-lingual settings.

1 Introduction

Lexical competence, the ability to understand and actively use vocabulary within a language, is a foundational component of linguistic proficiency, necessary for effective speech activity and the expression of ideas (Chunakhova et al., 2023). As Marconi (1997) stated, it comprises two complementary abilities: *inferential competence* (selecting the appropriate word for a given context) and *referential competence* (identifying appropriate contexts for a given word). These skills rely on a network of cognitive processes in the human brain, involving both formal linguistic competence (understanding language structure) and functional linguistic competence (applying language in logical and social reasoning) (Mahowald et al., 2024).

With the rapid advancement of large language models (LLMs), evaluating their lexical competence has become increasingly important for understanding their capabilities and limitations across languages and specialized domains. Traditional evaluation methods (word association tests, multiple-choice assessments, and lexical availability tests) face serious limitations when applied to LLMs: annotated benchmarks are costly to create (Maigeldiyeva et al., 2023; Boch and Sorba, 2020; Palapanidi, 2019), general-purpose benchmarks such as MMLU (Hendrycks et al., 2021) or AGIEval (Zhong et al., 2023) conflate task-specific reasoning with lexical skills, and multiple-choice formats are vulnerable to answer-design biases (Zheng et al., 2024) and transfer poorly across languages.

The purpose of this work is to tackle the fundamental challenge of evaluating lexical competence in language models without requiring costly human annotations for each test instance. Drawing on the theoretical foundation that lexical competence involves mapping words to concepts in the real world (Marconi, 1997), we introduce EVALEX, a novel framework that automatically assesses both directions of lexical knowledge (term to definition and definition to term) within a self-contained pipeline requiring no human annotation beyond an initial word list.

We validate EVALEX on two Spanish frequency-based datasets (CREA-10k-POS and CREA-10k-Lemmas) and against an expert-annotated sample of 100 words from Conde et al. (2024), comparing seven similarly sized models, including Spanish-centric variants. EVALEX scores correlate strongly

with expert rankings (Pearson $r = 0.8539$, Spearman $\rho = 0.8$) across the four models with available manual annotations, supporting its use as a scalable proxy for human judgment while motivating further validation on larger annotated sets.

Our main contributions are:

1. **A novel evaluation framework.** We propose EVALEX, an annotation-free, bidirectional definition-retrieval protocol for assessing lexical competence in language models.

2. **Empirical validation on Spanish.** We evaluate seven similarly sized models, including two Spanish-centric variants, on two large frequency-ranked vocabularies derived from the CREA corpus (7,784 word forms and 4,594 lemmas), demonstrating a consistent inverse relationship between term frequency and lexical competence.

3. **Correlation with expert judgments.** We validate EVALEX against a manually annotated 100-word sample, showing strong rank-order agreement with human evaluators and establishing it as a low-cost proxy for comparative model selection in data-scarce settings.

2 Related work

2.1 Lexical competence evaluation

LLM evaluation has typically followed two main paradigms. The dominant strategy uses comprehensive benchmarks that repurpose human-designed tests such as MMLU (Hendrycks et al., 2021) or AGIEval (Zhong et al., 2023) to proxy broad linguistic ability, but these conflate task-specific reasoning with fundamental lexical skills, making it difficult to isolate vocabulary knowledge from general problem-solving (Chang et al., 2024).

Work specifically targeting lexical evaluation has historically relied on discriminative tasks. The Word-in-Context (WiC) dataset (Pilehvar and Camacho-Collados, 2019) evaluates context-sensitive sense disambiguation using expert-curated pairs, but does not capture the generative depth of active vocabulary use. Along similar lines, Teglia et al. (2025) probed encoder-only models on word sense disambiguation, finding that middle layers outperform output layers, which evidences that sense-level knowledge is encoded internally yet does not always surface in generative behavior.

More recently, the field has shifted toward generative evaluation. Moskvoretskii et al. (2024) tested LLaMA-2 on definition generation, while Ciaccio et al. (2025) proposed a framework for lexical proficiency in Italian using curated definitions. Conde et al. (2024) assessed acquired vocabulary in Spanish through generated definitions. However, all these generative approaches rely on expert-curated datasets or manual annotation to validate outputs, limiting scalability to new languages or domains. Martínez et al. (2024) revived vocabulary tests for LLM evaluation across English and Spanish, demonstrating scalability through automatic test generation, but their approach remains discriminative (multiple-choice and yes/no formats) and does not probe the productive dimension of lexical knowledge that EVALEX targets.

Finally, Pham et al. (2025) studied the alignment between LLM-generated definitions and classical dictionary entries, finding that while model-produced definitions are generally faithful, they exhibit systematic differences in specificity and coverage. This is directly relevant to EVALEX, whose first step assumes models can generate semantically adequate definitions.

EVALEX's bidirectional design also connects to a broader tradition of cycle-consistency evaluation in Natural Language Processing (NLP), where a system's output is used as input to a reverse process to verify round-trip fidelity. This principle has been applied, for example, in machine translation (Sennrich et al., 2016).

2.2 Multilingual benchmarking and low-resource evaluation

Performance degradation for low-resource languages is a well-documented limitation of current LLMs (Hasan et al., 2024; Jayakody and Dias, 2024; Robinson et al., 2023), driven more by data scarcity than linguistic complexity (Pantelidou et al., 2025). A large-scale survey of over 2,000 multilingual benchmarks confirms that English remains heavily overrepresented in evaluation resources, and that performance correlations with human judgments vary markedly across task types (Wu et al., 2025). Native-language evaluation efforts such as INCLUDE (Romanou et al., 2025), which provides 197,243 QA pairs across 44 languages, have shown that culturally grounded benchmarks reveal deficiencies that translation-based approaches miss, a philosophy that informs our choice of the CREA corpus as EVALEX's lexical foundation rather than a translated English word list.

For Spanish specifically, IberBench (González et al., 2026) provides 101 datasets across 22 task categories for Iberian and Latin American languages, while La Leaderboard (Grandury et al., 2025) evaluates generative LLMs on natively authored Spanish and Latin American data, covering several of the same model families as our study. EVALEX complements these efforts with a targeted, lexically focused diagnostic that requires only a word list, making it especially practical in data-scarce settings where full benchmark suites are not feasible.

2.3 The Reversal Curse and long-tail knowledge

EVALEX's bidirectional design is motivated by the *Reversal Curse* (Berglund et al., 2023): autoregressive models trained on "A is B" often fail to infer "B is A", suggesting that fluent definition generation (term → definition) does not imply genuine bidirectional understanding. Wu et al. (2024) showed that this asymmetry is architecture-dependent (decoder models such as GPT and Llama are susceptible, while bidirectional encoders like BERT are not), validating our choice to probe the reverse path (definition → term) as a meaningful test of semantic comprehension for the autoregressive models we evaluate.

Beyond the reversal curse, there is strong evidence for a systematic *frequency bias*: performance degrades sharply for knowledge encountered infrequently during pre-training (Kandpal et al., 2023). The Head-to-Tail benchmark (Sun et al., 2024) found that the best models barely reach 31% accuracy on tail entities, and the LongTail-Swap framework (Algayres et al., 2025) confirmed that both syntactic and semantic abilities degrade on infrequent vocabulary, with architectural differences being most pronounced in the distribution's tail. This frequency effect extends to linguistic competence independently of factual knowledge, as Collado-Montañez et al. (2025) showed that factual knowledge scales faster with model size than linguistic competence. These findings collectively motivate EVALEX's frequency-stratified evaluation design and our focus on rare, domain-specific vocabulary as the most diagnostic region of the word distribution.

3 Methodology

EVALEX assesses lexical competence by testing a model's ability to identify a term from its own generated definition. To prevent in-context learning, the two steps are executed as fully isolated inference sessions: each prompt contains no shared context from the other, exploiting the stateless nature of individual LLM inference. The system's architecture and workflow are illustrated in Figure 1.

Given an initial set of target words $\mathcal{W}$ from a specific language or domain, the evaluation process for each word $w_i \in \mathcal{W}$ proceeds as follows:

Step 1 — Definition generation. The model under evaluation is prompted to generate a set $\mathcal{D}$ of m distinct definitions for w_i, capturing polysemy by associating one word with multiple senses. Informed by Ozturk (2017), who reports that high-frequency vocabulary averages approximately four meanings per word, we set $m = 5$. Generation uses sampling (Table 1) to encourage the model to explore multiple senses of the target word.

To prevent data leakage, we post-process all generated definitions by replacing every surface form of the target word including morphological variants, gendered forms, and orthographic errors, with a placeholder token (_). This step is non-trivial: in our experiments, Llama-3.1-8B-Instruct included the target term in 0.83% of its definitions, while salamandra-7b-instruct did so in 14.93% of cases, making masking especially consequential for the latter. If JSON parsing of the model output fails, the system treats the response as raw text, strips JSON-specific artifacts (braces, quotes), and extracts content following the final colon separator.

Step 2 — Reverse lookup. In a new, separate interaction for each definition $d_j \in \mathcal{D}$, the model is prompted to generate a set of candidate words $\mathcal{W}'_j$ corresponding to that definition, thereby accounting for synonymy. In deliberate contrast to Step 1, generation here uses greedy decoding (`do_sample = False`): Step 1 benefits from diversity to capture polysemy, whereas Step 2 is a retrieval task with a determinate correct answer, and greedy decoding reduces noise while making the matching criterion cleaner.

Candidate generation is capped at 64 output tokens, which, given the fertility of Spanish text across common tokenizers (∼1.32–1.61 tokens per word) (Nayeem et al., 2025), yields an effective up-

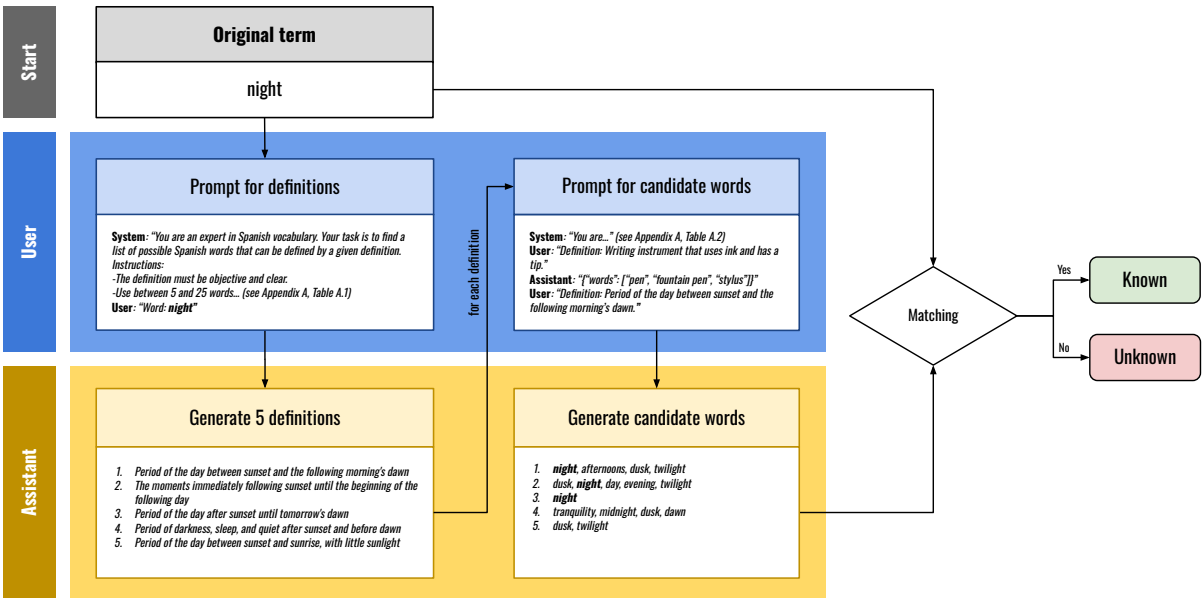

Figure 1: System workflow to evaluate whether a given model knows the word "night". The example is translated into English for illustration.

per bound of approximately 40–48 candidate words per definition. This budget is sufficient to recover the target term from any semantically adequate definition while ensuring no model is advantaged by producing an arbitrarily long candidate list. We additionally employ a one-shot configuration (Table 2) to assist older models that struggle with complex instructions; this configuration is not necessary for modern instruction-tuned LLMs but is retained for comparability with Conde et al. (2024).

As in Step 1, a heuristic extraction strategy is applied if the model fails to produce a valid JSON list: the system parses raw text to identify candidate terms, ensuring that formatting failures are not incorrectly counted as unknown words.

Step 3 — Candidate aggregation and verification. All candidate sets from the m definitions are concatenated into a single set:

$$\mathcal{W}' = \bigcup_{j=1}^{m} \mathcal{W}'_j$$

The system then checks whether the original word w_i appears in $\mathcal{W}'$. If $w_i \in \mathcal{W}'$, the word is classified as *known*; otherwise it is classified as *unknown*. The model's final lexical competence score is the percentage of words in $\mathcal{W}$ successfully identified.

Design rationale. A critical design choice is that the *same* model serves as both definition generator and term retriever. Using an external model or dictionary in Step 1 would decouple what Step 2 measures: a model could succeed without genuinely knowing the word (by matching against a well-formed external definition) or fail simply because the external phrasing differs from its internal representation. By contrast, when the same model both defines and retrieves, success requires that it can articulate the concept behind a word *and* recognise that articulation as pointing back to the same lexical item, a direct operationalisation of the bidirectional lexical competence described by Marconi (1997).

4 Experimental setup

This section details the setup used to evaluate our lexical competence method on a small annotated dataset (DEA) and a comparison between models on a larger unannotated one (CREA), both in Spanish. All the implemented code and data described below are available in the public repository for this work[1]. Every experiment has been run on a single Tesla V100 GPU.

4.1 Data

Unlike traditional evaluation approaches that require annotated corpora or labeled datasets, our method only requires a curated list of target terms. This minimal requirement makes the framework highly adaptable: the term list can be sourced from

[1] repository anonymized for double blind review

language proficiency standards (such as the European Framework of Reference for Languages[2]), domain-specific lexicons (legal, engineering, life sciences), or frequency-ranked vocabularies from reference corpora. This flexibility enables evaluation across different languages, proficiency levels, and specialized domains.

In this paper, we focus on evaluating lexical competence in Spanish. We utilize two complementary datasets that serve different purposes in our evaluation framework:

- **CREA corpus:** A large-scale frequency-based vocabulary for comprehensive quantitative evaluation across varying levels of word difficulty.
- **DEA dictionary sample:** A manually annotated subset for validation against expert human judgments.

4.1.1 CREA frequency-based vocabulary

Our primary dataset is derived from the CREA corpus[3] (*Corpus de Referencia del Español Actual*), which provides a frequency-ranked list of the 10,000 most common words in contemporary Spanish. This corpus-based approach allows us to evaluate lexical competence across a broad spectrum of vocabulary, from high-frequency everyday terms to less common specialized words.

CREA-10k-POS: The raw CREA word list contains non-alphabetic entries (punctuation marks, numerals, symbols) that are not suitable for lexical competence evaluation. We applied the following preprocessing steps to obtain a clean vocabulary:

1. Remove all non-alphabetic entries (punctuation, numerals, symbols).
2. Lowercase all words to ensure consistent matching.
3. Eliminate function words to keep only content words (nouns, adjectives, verbs, and adverbs) that carry substantial semantic information.

The resulting vocabulary, which we refer to as CREA-10k-POS, consists of 7,784 word forms.

CREA-10k-Lemmas: To account for morphological variations in Spanish (number, gender, verb conjugations), we created a lemmatized version of the vocabulary. Lemmatization groups inflected forms under their base form (e.g., *hablo, hablas, habla* → *hablar*), which facilitates more reliable matching between model-generated text and target terms. This allows us to evaluate whether models understand the core lexical concept regardless of its surface form. The lemmatized vocabulary, CREA-10k-Lemmas, comprises 4,594 unique lemmas.

These two vocabulary variants enable us to analyze lexical competence at different granularities: CREA-10k-POS assesses knowledge of specific word forms, while CREA-10k-Lemmas evaluates understanding of underlying lexical concepts.

4.1.2 DEA expert-annotated sample

To validate our automated evaluation framework against human expert judgments, we utilize a dataset from Conde et al. (Conde et al., 2024). This dataset consists of 100 Spanish words randomly sampled from the *Diccionario del Español Actual* (DEA)[4], a comprehensive dictionary containing over 84,000 entries representing contemporary Spanish vocabulary.

Each word in this sample was manually annotated by linguistic experts who compared language model responses against authoritative dictionary definitions. Words were classified as either *known* (the model demonstrated correct understanding) or *unknown* (the model's response was inconsistent with the dictionary definition). While this sample includes some relatively uncommon terms and may not comprehensively represent everyday Spanish vocabulary, it serves as a valuable gold standard for assessing the alignment between our automated scores and expert human judgments. The complete word list is provided in Table 8.

4.2 Models

To evaluate our approach and be able to compare our results to those from Conde et al. (Conde et al., 2024), we selected several of the same models they used, including: Llama-2-7b-chat-hf (Touvron et al., 2023), Mistral-7B-Instruct-v0.1 (Jiang et al., 2023), gemma-7b-it (Gemma Team, 2024), and Yi-6B-Chat (01.AI, 2024). These are widely recognized language models of comparable size, which facilitates a fair comparison of their performance.

In addition to those, we also included a more recent model of similar size (Llama-3.1-8B-Instruct (Llama Team, 2024)) to compare currently prominent LLMs in Spanish vocabulary based on

[2] https://www.coe.int/en/web/common-european-framework-reference-languages

[3] https://www.rae.es/banco-de-datos/crea

[4] https://www.fbbva.es/diccionario/

the CREA datasets. We also included two Spanish-specific models (Gromenauer-7B-Instruct[5] and salamandra-7b-instruct (Gonzalez-Agirre et al., 2025)) to assess how domain-specific training influences performance in our task.

4.3 Implementation details

Table 1 and Table 2 report the generation parameters used in Steps 1 and 2 respectively. The asymmetric sampling strategy (diverse sampling in Step 1, greedy decoding in Step 2) is motivated in Section 3.

Parameter	Value
do_sample	True
temperature	1
top_k	50
max_new_tokens	128
num_return_sequences	5

Table 1: Generation parameters for Step 1 (definition generation).

Parameter	Value
do_sample	False
max_new_tokens	64
num_shots	1

Table 2: Generation parameters for Step 2 (reverse lookup). Greedy decoding is used to maximise retrieval precision.

The prompts used in both steps are provided in full in Appendix A, in both Spanish (original) and English (translated for reader convenience).

5 Results

The performance of the selected LLMs on the CREA-10k-POS and CREA-10k-Lemmas datasets is summarized in Table 3. As anticipated, all models achieved higher scores on the lemmatized dataset. This is attributed to the fact that lemmatization reduces morphological variance by normalizing inflectional differences, such as gender and number, thereby simplifying the retrieval task within the EvaLex framework.

Applying EvaLex across frequency-stratified subsets of CREA reveals a consistent inverse relationship between term frequency and lexical competence (Figure 2), in line with prior work on the long-tail knowledge gap in LLMs (Kandpal et al., 2023). Notably, the two Spanish-centric models (Gromenauer-7B-Instruct and salamandra-7b-instruct) appear to sustain higher competence on rare terms relative to their general-purpose counterparts, even without outperforming them on high-frequency vocabulary. This pattern, visible across both dataset variants, suggests a potential benefit of language-specific pretraining for low-frequency robustness, though we have not formally tested these differences statistically.

5.1 Validation against expert annotations

Table 4 reports EvaLex scores and manual annotations on the 100-word DEA dataset. Automatic scores are consistently lower, as expected from EvaLex's conservative exact-match criterion, yet the relative ranking of models is nonetheless preserved: Pearson $r = 0.8539$ and Spearman $\rho = 0.8$ indicate strong ordering agreement. We stress, however, that this correlation is computed over only the four models for which paired manual annotations are available from Conde et al. (2024), which limits its statistical power and means it should be interpreted with caution as a proof-of-concept alignment rather than a definitive validation. The consistency of this ranking across all four models, together with the precision and recall metrics in Table 5, supports the interpretation of EvaLex as a plausible annotation-free proxy for comparative model assessment, while also motivating further validation on larger annotated sets.

6 Discussion

The comparison between EvaLex scores and the manual annotations from Conde et al. (2024) reveals strong structural alignment despite a substantial gap in absolute values. As shown in Table 4, EvaLex consistently produces more conservative estimates: for instance, Llama-2-7b-chat-hf was credited with 12 known words automatically versus 42 by human evaluators.

We attribute this gap to the conservative design of EvaLex: it requires a precise bidirectional match, whereas human evaluators typically assess only the forward direction and accept synonyms or partial descriptions as evidence of competence. Additionally, the uniform exact-match threshold removes the "benefit of the doubt" that manual assessors naturally apply, which is precisely what makes relative scores directly comparable across models

[5] https://huggingface.co/bertin-project/Gromenauer-7B-Instruct

Model	% CREA-10k-POS	% CREA-10k-Lemmas
Gromenauer-7B-Instruct	69.05%	81.30%
Llama-3.1-8B-Instruct	65.38%	83.96%
Mistral-7B-Instruct-v0.1	52.13%	70.64%
Llama-2-7b-chat-hf	44.06%	55.38%
Yi-6B-Chat	40.81%	51.18%
salamandra-7b-instruct	40.74%	42.58%
gemma-7b-it	28.22%	37.18%

Table 3: EVALEX scores on CREA-10k-POS (7,784 terms) and CREA-10k-Lemmas (4,594 terms).

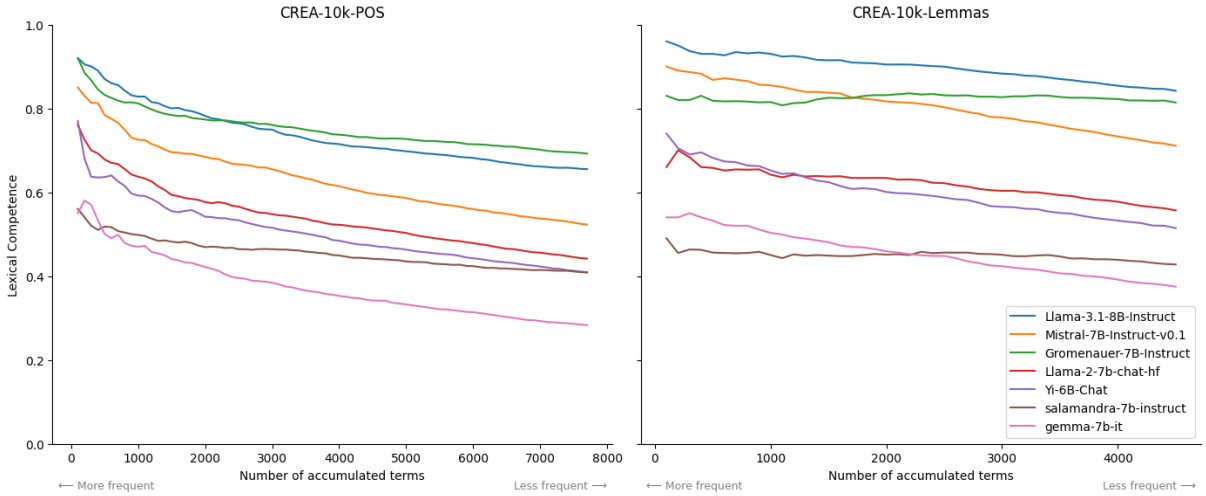

Figure 2: EVALEX results on both the CREA-10k-POS and CREA-10k-Lemmas datasets across terms accumulated by frequency in bins of 100 elements.

Model	Automatic evaluation	Manual evaluation
Mistral-7B-Instruct-v0.1	18	51
Yi-6B-Chat	13	27
Llama-2-7b-chat-hf	12	42
gemma-7b-it	5	20
Gromenauer-7B-Instruct	40	-
Llama-3.1-8B-Instruct	35	-
salamandra-7b-instruct	28	-

Table 4: Scores achieved by every model tested on the DEA dataset.

Model	TP	FP	TN	FN	Precision	Recall	F1 Score
Mistral-7B-Instruct-v0.1	16	2	47	35	0.8889	0.3137	0.4638
Yi-6B-Chat	8	5	68	19	0.6154	0.2963	0.4000
Llama-2-7b-chat-hf	9	3	55	33	0.7500	0.2143	0.3333
gemma-7b-it	3	2	78	17	0.6000	0.1500	0.2400

Table 5: Comparison of model predictions with manual annotations.

(see Limitations for a full discussion of these trade-offs). The metrics in Table 5 support this reading: high Precision relative to Recall confirms that when EVALEX classifies a word as known, it is highly likely to agree with human judgment (Precision 0.60–0.89), while it frequently rejects words a human would have accepted (Recall 0.15–0.31). The decline in F1 for lower-performing models points to a compounding effect, whereby limited lexical knowledge is exacerbated by weaker instruction-

following in older-generation architectures.

Beyond validation, our results highlight the role of the long-tail frequency distribution in shaping model performance. The sharp, consistent decline in lexical competence as term frequency decreases (Figure 2) confirms that the functional vocabulary of general-purpose LLMs is concentrated in high-frequency language. Compared to prior work on long-tail knowledge (Kandpal et al., 2023; Sun et al., 2024), EVALEX provides a word-level view of where this degradation occurs. The two Spanish-centric models (Gromenauer-7B-Instruct and salamandra-7b-instruct) appear to exhibit greater resilience at the low-frequency end of the distribution, even when they do not outperform general-purpose models on high-frequency terms. This pattern is consistent across both dataset variants and aligns with prior findings on the benefits of language-specific pretraining, though a formal statistical comparison would be needed to establish this difference conclusively. Broadening the annotated validation set beyond the current four-model comparison remains a clear priority for future work.

7 Conclusions

We introduced EVALEX, an automated, annotation-free framework for evaluating lexical competence in language models via a bidirectional definition-retrieval protocol. Validated on Spanish, EVALEX scores correlate strongly with expert human rankings (Pearson $r = 0.8539$, Spearman $\rho = 0.8$), a promising signal that nonetheless calls for validation on larger annotated sets before the framework can be considered a fully confirmed proxy for human judgment. Our results reveal a clear frequency bias across all tested models, with lexical competence declining sharply as word frequency decreases, and show that Spanish-centric models are more resilient on rare vocabulary despite not outperforming general-purpose models on high-frequency terms. These results are consistent with a potential benefit of language-specific pretraining for low-frequency robustness, though a causal interpretation would require controlled experiments beyond the scope of this benchmarking study.

Several directions follow naturally from this work. Beyond Spanish, applying EVALEX to morphologically diverse or lower-resource languages will test the framework's transferability and expose language-specific failure modes. Connecting lexical competence scores to downstream task performance would clarify how vocabulary mastery interacts with broader linguistic reasoning. Finally, developing methods to analyze not just whether a word is known, but the depth and nuance of understanding, could provide even richer insights into model capabilities such as epistemic humility that may help us build models able to recognize self-ignorance, a key competence to deal with hallucinations (Kalai et al., 2025).

Limitations

Conservative exact-match criterion. EVALEX uses exact string matching at the retrieval step, which means that semantically correct near-misses, such as synonyms, are penalised as failures. As a consequence, EVALEX systematically underestimates the absolute lexical competence of any individual model. This is a deliberate design choice that prioritises precision and comparability, but it implies that EVALEX scores should be interpreted as a lower bound rather than a point estimate of true vocabulary knowledge.

Binary aggregation and depth of knowledge. EVALEX classifies a word as known if the target term is recovered from any of the m generated definitions, and as unknown otherwise. This binary aggregation means that a model that retrieves the term from four out of five definitions is treated identically to one that retrieves it from a single definition. As a consequence, EVALEX does not currently capture the reliability of lexical knowledge, only its presence or absence.

Instruction-following as a confound. For older or less instruction-tuned models, low EVALEX scores may partially reflect a failure to adhere to the structured prompt format (producing malformed JSON or ignoring output constraints) rather than a genuine lexical deficit. Although our heuristic fallback parser recovers many such cases, some failures remain unrecoverable and are counted as unknown words.

Limited validation scope. The correlation between EVALEX scores and expert human judgments is established on a single 100-word Spanish sample. While the correlations are strong ($r = 0.8539$, $\rho = 0.8$), the small validation set and its restriction to contemporary Spanish mean that the reliability of EVALEX as a proxy for human judgment has not yet been confirmed in other

languages, domains, or annotation schemes.

Language and domain coverage. All experiments in this paper are conducted in Spanish using frequency-ranked general vocabulary. Whether EVALEX generalises to morphologically different languages, highly specialised technical domains, or lower-resource settings where model-generated definitions may be less reliable remains an open empirical question.

Acknowledgments

Removed for blind review.

References

01.AI. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Robin Algayres, Charles-Éric Saint-James, Mahi Luthra, Jiayi Shen, Youssef Benchekroun, Dongyan Lin, Rashel Moritz, Juan Pino, and Emmanuel Dupoux. 2025. LongTail-swap: benchmarking language models' abilities on rare words. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11231–11251, Suzhou, China. Association for Computational Linguistics.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *arXiv preprint arXiv:2309.12288*.

Françoise Boch and Julie Sorba. 2020. Tester la compétence lexicale des adultes francophones: Réflexion sur le choix des distracteurs dans un test à choix multiples. *Lidil. Revue de Linguistique et de Didactique des Langues*, 62.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3).

L. Chunakhova, N. Gerasimova, A. Podkovalnikova, and A. Kravtsova. 2023. Improving lexical skills as an important part of teaching english. In *XV International Scientific Conference "INTERAGROMASH 2022"*, volume 574 of *Lecture Notes in Networks and Systems*, pages 847–854, Cham. Springer.

Cristiano Ciaccio, Alessio Miaschi, and Felice Dell'Orletta. 2025. Evaluating lexical proficiency in neural language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1267–1286, Vienna, Austria. Association for Computational Linguistics.

Jaime Collado-Montañez, L. Alfonso Ureña-López, and Arturo Montejo-Ráez. 2025. Towards fundamental language models: Does linguistic competence scale with model size? *Preprint*, arXiv:2509.02225.

Javier Conde, Miguel González, Nina Melero, Raquel Ferrando, Gonzalo Martínez, Elena Merino-Gómez, José Alberto Hernández, and Pedro Reviriego. 2024. Open source conversational llms do not know most spanish words. *Preprint*, arXiv:2403.15491.

Gemma Team. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

José Ángel González, Ian Borrego Obrador, Álvaro Romo Herrero, Areg Mikael Sarvazyan, Mara Chinea-Ríos, Angelo Basile, and Marc Franco-Salvador. 2026. Iberbench: Llm evaluation on iberian languages. *Computer Speech & Language*, 96:101899.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Iñigo Pikabea, Jorge Palomar, Júlia Falcão, Lucía Tormo, and 4 others. 2025. Salamandra technical report. *Preprint*, arXiv:2502.08489.

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, Vienna, Austria. Association for Computational Linguistics.

Md. Arid Hasan, Prerona Tarannum, Krishno Dey, Imran Razzak, and Usman Naseem. 2024. Do large language models speak all languages equally? a comparative study in low-resource settings. *Preprint*, arXiv:2408.02237.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ravindu Jayakody and Gihan Dias. 2024. Performance of recent large language models for a low-resourced language. *Preprint*, arXiv:2407.21330.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego

de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *Preprint*, arXiv:2509.04664.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. *Preprint*, arXiv:2211.08411.

Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Zhannat Maigeldiyeva, Zukhra Abdrakhmanova, and Olga Pavlovskaya. 2023. Assessing the general and academic lexical competencies: Evidence from online russian vocabulary size test and word associations test (rl). *International Journal of Information and Education Technology*, 13(8).

Diego Marconi. 1997. *Lexical Competence*. A Bradford Book. MIT Press, Cambridge, MA.

Gonzalo Martínez, Javier Conde, Elena Merino-Gómez, Beatriz Bermúdez-Margaretto, José Alberto Hernández, Pedro Reviriego, and Marc Brysbaert. 2024. Establishing vocabulary tests as a benchmark for evaluating large language models. *PLOS ONE*, 19(12):1–17.

Viktor Moskvoretskii, Alexander Panchenko, and Irina Nikishina. 2024. Are large language models good at lexical semantics? a case of taxonomy learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1498–1510, Torino, Italia. ELRA and ICCL.

Mir Tafseer Nayeem, Sawsan Alqahtani, Md Tahmid Rahman Laskar, Tasnim Mohiuddin, and M Saiful Bari. 2025. Beyond fertility: Analyzing strr as a metric for multilingual tokenization evaluation. *Preprint*, arXiv:2510.09947.

Meral Ozturk. 2017. Multiple meanings in the efl lexicon. *International Journal of Curriculum and Instruction*, 9(2):1–10.

Kiriakí Palapanidi. 2019. La disponibilidad léxica como instrumento de evaluación léxica de aprendices griegos de ele. *Verba Hispanica*, 27(1):47–71.

Nikoleta Pantelidou, Evelina Leivada, and Paolo Morosi. 2025. Resource-sensitive but language-blind: Community size and not grammatical complexity better predicts the accuracy of large language models in a novel wug test. *Preprint*, arXiv:2510.12463.

Bach Pham, JuiHsuan Wong, Samuel Kim, Yunting Yin, and Steven Skiena. 2025. Word definitions from large language models. In *2025 19th International Conference on Semantic Computing (ICSC)*, pages 158–162. IEEE.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high- (but not low-) resource languages. *Preprint*, arXiv:2309.07423.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya K Dalmia, Abraham Diress, Sharad Duwal, and 38 others. 2025. Include: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations*, volume 2025, pages 83291–83322.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? 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 311–325, Mexico City, Mexico. Association for Computational Linguistics.

Simone Teglia, Simone Tedeschi, and Roberto Navigli. 2025. How much do encoder models know about word senses? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2266–2277, Vienna, Austria. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Da Wu, Jingye Yang, and Kai Wang. 2024. Exploring the reversal curse and other deductive logical reasoning in bert and gpt-based large language models. *Patterns*, 5(9):101030.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *Preprint*, arXiv:2504.15521.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. *Preprint*, arXiv:2309.03882.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

A Prompts

Prompts used to define terms (Table 6) and generate words (Table 7) are shown below in both the original Spanish and a translated English version to facilitate reader comprehension.

B Word list

Table 8 shows all the words used to evaluate the lexical competence of language models in Spanish.

C Generation sample

Table 9 shows a sample of the original results in Spanish extracted from the Llama-3.1-8B-Instruct model. The first two words are concluded to be known by the model, and the third and fourth are labeled as unknown.

```
{"role": "system", "content": "Eres un experto en vocabulario español. Tu tarea es dar una definición para una palabra
↪ dada.
    Instrucciones:
    - La definición debe ser objetiva y clara.
    - Usa entre 5 y 25 palabras.
    - La definición generada no debe incluir la palabra original ni derivados directos o familia léxica.
    IMPORTANTE:
    La respuesta debe ser un **objeto JSON**:
    Formato de respuesta:
    {
    "palabra": "(palabra original)",
    "definición": "(definición sin usar la palabra ni su familia léxica)"
    }"
    },
{"role": "user", "content": "Palabra: {word}"}
```

```
{"role": "system", "content": "You are an expert in Spanish vocabulary. Your task is to provide a definition for a given
↪ word.
    Instructions:
    - The definition must be objective and clear.
    - Use between 5 and 25 words.
    - The generated definition must not include the original word or any direct derivatives or lexical family members.
    IMPORTANT:
    The response must be a **JSON object**:
    Response format:
    {
    ¨word¨: ¨(original word)¨,
    ¨definition¨: ¨(definition without using the word or its lexical family)¨
    }"
    },
{"role": "user", "content": "Word: {word}"}
```

Table 6: Spanish and English versions of the prompt used to define every term.

```
{"role": "system", "content": "Eres un experto en vocabulario español. Tu tarea consiste en encontrar una lista de
↪ posibles palabras en español que pueden definirse con una definición dada.
    Instrucciones:
    - Las palabras posibles no están en la definición.
    - Genera todas las palabras que encajen con la definición.
    - Responde con una lista vacía si no encuentras ninguna palabra adecuada para la definición dada.
    La respuesta debe ser un **objeto JSON**.
    Formato de respuesta:
    {
    "palabras": ["(palabra 1)", "(palabra 2)", "(palabra n)"]
    }"
    },
{"role": "user", "content": "Definición: instrumento de escritura que utiliza tinta y tiene una punta"},
{"role": "assistant", "content": "{"palabras": ["bolígrafo", "pluma", "estilográfica"]}"},
{"role": "user", "content": "Definición: {definition}."}
```

```
{"role": "system", "content": "You are an expert in Spanish vocabulary. Your task is to find a list of possible
↪ Spanish words that can be defined by a given definition.
    Instructions:
    - The possible words must not appear in the definition.
    - Generate all words that fit the definition.
    - Respond with an empty list if no suitable words are found for the given definition.
    The response must be a **JSON object**.
    Response format:
    {
    "words": ["(word 1)", "(word 2)", "(word n)"]
    }"
    },
{"role": "user", "content": "Definition: writing instrument that uses ink and has a tip"},
{"role": "assistant", "content": "{"words": ["pen", "fountain pen", "stylus"]}"},
{"role": "user", "content": "Definition: {definition}."}
```

Table 7: Spanish and English versions of the prompt used to generate words from a given definition.

acelguilla wild chard	*agüista* spa visitor	*antidiarreico* antidiarrheal	*apealar* to lasso	*apparat* apparat
arante arante	*ardilla* squirrel	*bátavo* batavian	*bicicross* bmx	*bifocal* bifocal
cantautor singer-songwriter	*cantinero* bartender	*cartulina* cardstock	*centavo* cent	*cerebrotónico* cerebrotonic
chalaza chalaza	*chigüire* capybara	*coliseo* coliseum	*conspirar* to conspire	*corbata* necktie
corralero cowboy	*cosmotrón* cosmotron	*crístico* Christ-like	*cuentapartícipe* joint participant	*dabuti* great
dactiloscopia fingerprinting	*deformabilidad* deformability	*desinfección* disinfection	*desinsectación* pest control	*desmitificador* demystifier
diligentemente diligently	*emparejador* surveyor	*empurpurado* purple-clad	*epifito* epiphyte	*escorar* to heel
estadista statesman	*esteatosis* steatosis	*estuco* stucco	*exequias* funeral rites	*faisánido* phasianid
fétido fetid	*floración* flowering	*fotogramétrico* photogrammetric	*funcionario* civil servant	*gabato* fawn
garcilla little egret	*giroscopio* gyroscope	*helenizante* hellenizing	*hipogino* hypogynous	*incrustar* to encrust
intercadencia intermittence	*jaín* Jain	*lipotimia* fainting fit	*magnolia* magnolia	*manes* manes
meano meatus	*mediar* to mediate	*mensualizar* to monthly	*mesmerización* mesmerization	*mestizo* mestizo
minuto minute	*mochalero* fruit picker	*modisto* fashion designer	*monásticamente* monastically	*morra* morra
nefólogo nephologist	*novatada* hazing	*ovni* ufo	*pagar* to pay	*paleteo* wagging
palmítico palmitic	*paralogismo* paralogism	*pasarratos* pastime	*perrillo* small dog	*pezuña* hoof
pinabete silver fir	*pitahaya* dragon fruit	*postrar* to prostrate	*prédica* sermon	*prolongador* extender
provinciano provincial	*puzzle* puzzle	*quepis* kepi	*raor* razorfish	*reciclado* recycled
rememorar to remember	*ridiculización* ridicule	*sagum* sagum	*salbanda* salband	*socialmente* socially
speed speed	*standarizar* to standardize	*sublimado* sublimated	*superbomba* superbomb	*talgo* talgo
tornajo trough	*treinta* thirty	*tundidor* shearer	*vega* meadow	*verraquear* to bellow

Table 8: List of 100 words used to evaluate the lexical competence of language models in both Spanish and English. English translations have been automatically generated and might not be perfectly accurate for extremely unusual words.

Word	Definition	Word List	Known
personas	seres humanos con pensamiento, conciencia y capacidad para interactuar socialmente.	humanos, **personas**, seres humanos, individuos, gente	True
people	*human beings with thought, consciousness, and the capacity to interact socially*	*humans,* ***people****, human beings, individuals, people*	True
noche	período del día comprendido entre el atardecer y la aurora siguiente por la mañana.	**noche**, tardes, anochecer, crepúsculo	True
night	*period of the day between sunset and the following morning's dawn.*	***night****, evenings, dusk, twilight*	True
derecho	enseñanza académica que estudia las instituciones, leyes y principios sociales.	sociología, antropología, sociología, política	False
law	*academic education that studies institutions, laws, and social principles.*	*sociology, anthropology, sociology, politics*	False
pesetas	moneda oficial de españa hasta la introducción del euro en 2002.	peseta	False
pesetas	*official currency of Spain until the introduction of the euro in 2002.*	*peseta*	False

Table 9: Sample of the results of Llama-3.1-8B-Instruct on the CREA_10k_POS dataset, both in Spanish and its English translation.