False Friends or Cognates? A Cross-lingual Semantic Ambiguity Evaluation for Galician, Portuguese and Spanish

Marta Vázquez Abuín[1], Jose Camacho-Collados[2], Marcos Garcia[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela,
[2]School of Computer Science and Informatics, Cardiff University
martavazquez.abuin@usc.gal, camachocolladosj@cardiff.ac.uk,
marcos.garcia.gonzalez@usc.gal

Abstract

The linguistic proximity between Galician, Portuguese, and Spanish results in a lexical overlap that often conceals semantic interference. This is particularly evident in false friends, posing a challenge for NLP systems. In this work, we assess whether state-of-the-art language models can identify and process false friends among these languages. We introduce six cross-lingual datasets –created manually or using semi-automatic methods, with all instances being carefully verified– covering cognates and false friends. We evaluate a broad range of encoder and decoder models of varying sizes via zero-shot and few-shot settings. Our results highlight the challenging nature of the task, but also show the clear progress made by LLMs in recent years, particularly those of a larger size, with smaller language models struggling on the task. Notably, unlike other tasks where language distance poses additional challenges, we find that linguistic proximity itself introduces errors: closely related language pairs tend to perform worse, reflecting the challenge of semantic discrimination due to lexical overlap.

1 Introduction

Closely related languages exhibit a high degree of lexical and orthographic similarity, which can facilitate cross-lingual understanding but also give rise to systematic semantic ambiguity (Kallini et al., 2025). This issue is especially salient for closely related Romance languages such as Spanish, Portuguese, and Galician, which share a substantial portion of their vocabulary (Garcia et al., 2018), increasing the risk of semantic interference.

In bilingual and multilingual contexts, formally similar words are categorized as cognates:[1] words across different languages that share a common etymological origin and maintain the same meaning, for example the Galician and Portuguese form *ponte* with the meaning of 'bridge'. However, cognates may diverge semantically over time; such cases are commonly referred to as false cognates or false friends. These are pairs that, despite their shared origin, have developed distinct meanings (Dominguez and Nerlich, 2002; Allan, 2009; Chamizo-Domínguez, 2012), often leading to cross-lingual interference. For instance, the Portuguese adjective *esquisito* means 'strange' or 'odd', whereas the Spanish adjective *exquisito* denotes something 'refined' or 'excellent'. As similar words are often assumed to share meaning, these phenomena are problematic both for second-language learners and bilingual speakers (Durán Escribano, 2004; Brenders et al., 2011), and also for computational language models (Limisiewicz et al., 2023; Cahyawijaya et al., 2025).

The proliferation of multilingual Large Language Models (LLMs) raises the question of whether these models can handle cross-lingual semantic ambiguity. In this context, false friends and cognates constitute a valuable diagnostic of lexical processing. Analyzing whether language models can identify false friends, and determining which contexts and cases are more challenging, is crucial for assessing their cross-lingual capabilities.

Our study investigates the ability of state-of-the-art models to detect and process false friends and cognates in closely related varieties (Galician, Portuguese, and Spanish), which coexist in regions such as the Iberian Peninsula and Latin America, and are frequently considered jointly in evaluation efforts and multilingual modeling (Gonzalez-Agirre et al., 2025; Ángel González et al., 2026). To this end, we introduce new cross-lingual datasets of cognates and false friends in context for these three languages, combining automatically extracted sentences with manually created examples. We evaluate these cross-lingual phe-

[1]Throughout this paper, we use the term *cognates* to refer exclusively to *true cognates*.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 39199–39214

nomena using encoder and decoder models. First, we establish an unsupervised baseline and train a logistic regression classifier to assess whether cognates and false friends can be distinguished from contextualized embeddings. We then evaluate current LLMs on a contextual semantic judgment task, where models must decide whether two target words are semantically equivalent in context, evaluating both zero-shot and few-shot settings. Our results, complemented by qualitative analysis and a preliminary evaluation in machine translation, show that although closed-weight LLMs demonstrate strong disambiguation abilities, cross-lingual semantic ambiguity remains challenging for most models, particularly in closely related language pairs. All new resources are freely released.[2]

2 Related work

Beyond traditional Word Sense Disambiguation, which identifies the intended meaning of a word in context typically by linking it to predefined sense inventories (Navigli, 2009; Bevilacqua et al., 2021), Word-in-Context (WiC) (Pilehvar and Camacho-Collados, 2019) has become increasingly popular for evaluating contextualized word meaning in an inventory-free manner. WiC has reformulated lexical semantic evaluation as a binary classification task, determining whether a target word is used with the same meaning in two different contexts. This approach avoids reliance on explicit sense inventories and has been extended to multilingual and cross-lingual settings, including XL-WiC (Raganato et al., 2020) and MCL-WiC (Martelli et al., 2021), making it particularly suitable for analyzing multilingual LLMs.

In this regard, while multilingual LLMs provide coverage across dozens of languages, their performance often degrades in non-English contexts (Zhang et al., 2023), which can be attributed in large part to the predominance of English data in their training corpora (Xu et al., 2025). In cross-lingual scenarios, lexical overlap often introduces semantic interference in closely related languages, where orthographic similarity (categorized as cognates or false friends) may mask divergent senses.

Prior works in this area have focused on the automatic identification of cognates and false friends at the word level. Early approaches addressed the identification of cognates and false friends using bilingual corpora (Mitkov et al., 2007). Subsequent studies have relied on word embeddings and vector-space models for the automatic detection of false friends (Sepúlveda Torres and Aluísio, 2011; Ljubešić and Fišer, 2013; Castro et al., 2018; Periñán-Pascual and Fernández Martínez, 2025), as well as for cognate detection (Batsuren et al., 2019; Akavarapu and Bhattacharya, 2024). Alternative approaches have treated false friend identification as a classification task (Nikov et al., 2024). Other studies have proposed unsupervised methods for constructing multilingual lexicons of false friends across multiple languages Uban and Dinu (2020), or conduct linguistic analyses of semantic false friends and borrowings accompanied by automatic correction methods (Uban et al., 2025). Furthermore, the effect of vocabulary overlap in multilingual LLMs has also been examined (Kallini et al., 2025). Despite these advancements, recent studies have highlighted limitations in the contextual disambiguation capabilities of multilingual LLMs. Specifically, Cahyawijaya et al. (2025) demonstrates that state-of-the-art models continue to struggle with false friends, where shared orthography across languages triggers incorrect semantic transfer despite conflicting contextual cues.

Some studies on false friends and cognates have focused on descriptive analyses or vocabulary lists between Galician, Spanish, and Portuguese (Bragado Trigo, 2006; Figueroa, 1995), with small initial WiC datasets available for Galician and Spanish (Abuín and Garcia, 2025). For the Spanish–Portuguese language pair, recent computational approaches have been proposed (Castro et al., 2018; Uban et al., 2025). Including Galician constitutes a compelling challenge, as it is closely related to Portuguese while strongly influenced by Spanish. In this work, we introduce new WiC-like multilingual datasets of false friends and cognates and present a systematic evaluation of current LLMs using a range of methods.

3 Dataset construction

To evaluate the cross-lingual abilities of LLMs on lexical semantic understanding, we build a cross-lingual WiC-style dataset including cognates and false friends. Each instance consists of a word pair in two languages and two sentences containing each word, with information whether the meanings of both words in context are the same (cognate) or different (false friend) – Table 1 includes an example for each type.

[2]https://github.com/mrtva/false-friends-vs-cognates

Dataset	W1	S1	W2	S2	Label	POS
ES–PT	TALHER	Deve ser comido apenas com as mãos, jamais com **talheres**. *It should only be eaten with your hands, never with **cutlery**.*	TALLER	También tiene un moderno **taller** de restauración. *It also has a modern restoration **workshop**.*	False Friend	N
GL–PT	VOTO	Só se permite un **voto** por familia. *Only one **vote** is allowed per family.*	VOTO	Havia apenas dois **votos** a favor da proposta. *There were only two **votes** in favor of the proposal.*	Cognate	N

Table 1: Example instances from our dataset. Columns report the language pair (Dataset), the target words (W_1/W_2) together with their contextual sentences (S_1/S_2), the semantic relation label (False Friend or Cognate), and the Part-of-Speech (POS).

Languages: Galician, Portuguese and Spanish constitute a closely related group of Ibero-Romance languages. While Galician and Portuguese have historically been considered varieties of the same language, Galician has been influenced by Spanish due to its status as a lower-prestige language in Spain and an official orthography modeled after Spanish conventions (Samartim, 2012). Although Portuguese and Spanish also share similarities, they exhibit more pronounced orthographic differences. These asymmetric statuses, combined with the fact that in some cases the languages coexist in overlapping territories and are included in multilingual LLMs and related initiatives, make this triad a compelling testbed for assessing cross-lingual semantic disambiguation.

Semantic phenomena: The datasets incorporate both *cognates* and *false friends*, further categorized into *total* false friends, whose meanings diverged across languages (e.g., *carpeta*, ES 'folder' vs. *carpete*, PT 'carpet'), and *partial* false friends, where at least one sense differs while others may still overlap (e.g., *suspender*, PT 'suspend' vs. GL 'suspend' or 'fail'). The datasets include two different types of variation. First, semantic variation in strict homographs (e.g., *propina*, meaning 'tip' in ES vs. 'fee' in PT). Second, formal variations, which include morpho-phonological variations, such as the preservation of diphthongs in Galician and Portuguese in contrast to the Spanish simplification (e.g., GL/PT *touro* vs. ES *toro*, 'bull'), and orthographic correspondences (e.g., the representation of /ɲ/ and /ʎ/ respectively as 'nh' and 'lh' in Portuguese, and as 'ñ' and 'll' in Spanish and Galician).

Data compilation: The datasets were constructed using two complementary resources, one targeting false friends and the other cognates. False friends pairs were manually compiled from language learning materials and online pedagogical resources for the relevant language pairs (a complete list of resources is provided in Appendix B). These items correspond to words that are known to be problematic for second-language learners or bilinguals due to cross-linguistic similarity. Each pair was then annotated as either a total or partial false friend depending on the degree of semantic overlap. Regarding cognate pairs, they were automatically extracted from WordNet (Miller et al., 1990) by identifying synsets shared between the two languages using the Multilingual Central Repository (MCR) 3.0 (Gonzalez-Agirre et al., 2012). In order to maintain balance between categories, an equivalent number of cognate pairs was selected in each respective language pair. To obtain example sentences, we first collected 2024 Wikipedia dumps for each language, which were tokenized and lemmatized using FreeLing (Padró and Stanilovsky, 2012). We then extracted up to ten contextualized sentences per word for each language, and the most representative examples were selected by a linguist with native or near-native proficiency in the three languages.

Annotation task: The resulting sentences were validated by two annotators with a linguistic background.[3] The task consisted of verifying (i) whether the target word in each sentence correctly reflected the assigned sense, and (ii) whether the semantic relation remained valid in context. The process yielded an average raw agreement of 90.43% and a pooled Cohen's κ of 0.839 (ES–PT: 0.800, GL–ES: 0.809, GL–PT: 0.854). Disagreements were discussed, and when no consensus was reached, the pairs were discarded to maintain high-quality and balanced datasets.

[3]Both annotators are native speakers of Galician and Spanish with advanced academic proficiency in Portuguese.

Human-authored subset: In addition to the semi-automatically constructed data, we derived a human-authored subset from the validated instances. For each language pair, we randomly sampled between 50 and 60 word pairs and asked annotators to produce new sentences containing the same target words and preserving the original senses. These newly created sentences underwent the same validation procedure as described above.

Dataset statistics: Table 2 presents an overview of the dataset per language pair. In total, our data contain 704 validated cross-lingual word pairs, each one associated with two sentences, thus totaling 1,408 examples. Each set is balanced in number of cognates and false friends, the latter divided into total and partial, as detailed in Appendix A.[4]

Pair	Cogn.	FFs	Total	Sents.
ES–PT	182	182	364	728
GL–ES	75	75	150	300
GL–PT	95	95	190	380
Total	352	352	704	1,408

Table 2: Distribution of cognates (Cogn.) and false friends (FFs) in the dataset (semi-automatic and human-created subsets), and number of sentences (Sents).

4 Experimental Setup

Our cross-lingual evaluation includes two complementary paradigms: (i) WiC-based standard methods on encoder models (Section 4.1), and (ii) prompt-based approaches using LLMs, in zero-shot and few-shot settings (Section 4.2).

4.1 Encoder experiments

Unsupervised Baseline: We implement a baseline based on the cosine similarity between the contextualized embeddings of the target word pair. Following a binary classification approach, a pair is labeled as *same sense* if its similarity exceeds a given threshold, and as *different sense* otherwise. We explore thresholds from 0.00 to 1.00 in steps of 0.02 across all layers, and select the layer-threshold combination that yields the highest accuracy, which serves as a performance upperbound for this type of approach.

Logistic Regression: In line with the supervised framework described by Wang et al. (2019), we train logistic regression classifiers using the concatenated contextualized embeddings of the target word pair. To ensure a consistent comparison with the baseline, we use the representations from the same best-performing layer identified in the unsupervised experiment. We employ the `minicons` library[5] (Misra, 2022) to extract contextualized word representations from HuggingFace's Transformers (Wolf et al., 2020), and classifiers are implemented with *scikit-learn* (Pedregosa et al., 2011). We perform a grid search using different regularization strengths, solvers, and penalty types and training up to 1k iterations, and selected the best configuration based on the performance on the development set. The best configuration (C=1, *lbfgs*, l_2) was stable across settings, with accuracy varying by at most 1.2 points. For evaluation on a specific language pair, the model is trained using the combined data from the two remaining language pairs.[6]

Models: We use the following multilingual encoders: mBERT (cased and uncased) (Devlin et al., 2019), XLM-RoBERTa (base and large) (Conneau et al., 2020), and XL-Lexeme, an XLM-RoBERTa-large trained with a combination of various WiC-like datasets (Cassotti et al., 2023).

4.2 Prompt-based LLMs

We evaluate LLMs in zero-shot and few-shot ($k = 2$) configurations. Given target words w_1 and w_2, and their corresponding sentences S_1 and S_2 (each from a different language), the model must perform a binary decision: `YES` if the target words are semantically equivalent, or `NO` if they refer to distinct senses. To enable automated parsing, we constrain the output format and require a brief justification for each decision (see Appendix E and F for the prompts used in both settings). The justification is not used for scoring, and it is included only to support qualitative error analyses. For the few-shot setting, we provide two balanced examples specific to each language pair: one representing a cognate (`YES`), and one a false friend (`NO`). These examples are excluded from the test set. Prompts remain fixed across all models to ensure a fair comparison.

Models: For the prompt-based evaluation, we consider a broad range of open and closed-weight

[4]For these pairs, 534 contextual sentence pairs were obtained semi-automatically, while 170 were human-produced (see Table 8 for a more detailed explanation).

[5]https://github.com/kanishkamisra/minicons

[6]For instance, training on GL-ES and ES-PT pairs to evaluate the GL-PT pair.

Type	Size	Family	Models
Open-weight	*Small* [<10B]	Google	gemma-3-4b-it, gemma-2-9b-it
		Meta	llama-3.1-8b-instruct, llama-3.2-3b-instruct
		Microsoft	phi-3-mini-128k-instruct
		Mistral AI	ministral-3b, mixtral-8x7b-instruct
	Medium [10B - 50B]	Google	gemma-3-12b-it, gemma-3-27b-it
		Microsoft	phi-3-medium-128k-instruct, phi-4-reasoning-plus†, phi-4-multimodal-instruct
		Mistral AI	mistral-small-24b-instruct-2501, mistral-small-3.2-24b-instruct
		Qwen	qwen3-30b-a3b-instruct-2507
	Large [>50B]	DeepSeek	deepseek-chat-v3.1†
		Meta	llama-3.1-70b-instruct, llama-3.2-70b-instruct
		Qwen	qwen3-next-80b-a3b-instruct, qwen3-235b-a22b-2507
Closed-weight		Google	gemini-2.5-flash, gemini-2.5-flash-lite, gemini-2.0-flash-001
		OpenAI	gpt-4o-mini, gpt-4.1, gpt-4.1-mini
		xAI	grok-3-mini, grok-4-fast, grok-4

Table 3: Categorization of the evaluated LLMs by access type (Open vs. Closed), parameter scale (Small, <10B); Medium, 10B–50B; and Large, >50B), and architectural family. Parameter counts for closed-weight models are omitted due to proprietary restrictions. While the majority of our evaluated models follow the standard instruction-tuning paradigm, we include several reasoning models marked with †.

multilingual LLMs. The closed-weight group includes Gemini 2.0 and 2.5 (Comanici et al., 2025), GPT-4o-mini and GPT-4.1 (OpenAI et al., 2024), and Grok-3 and Grok-4. Among the open-weight models, we selected Gemma-2 (Team et al., 2024) and Gemma-3 (Team et al., 2025), Qwen-3 (Yang et al., 2025), Phi-3 (Abdin et al., 2024), Phi-4 (Abdin et al., 2025), Deepseek-V3 (DeepSeek-AI et al., 2025), Llama 3 series (Grattafiori et al., 2024) and Mistral (Jiang et al., 2023, 2024). These models span various scales, from 3B to over 200B parameters. For clarity, Table 3 summarizes the evaluated models, grouped by availability, size, and family.

5 Results

This section presents the experimental results, using the average of each language pair based on both subsets of the data (semi-automatic and manual).[7]

Encoder models: Among the encoder models, XL-Lexeme —an XLM-RoBERTa-large adapted to WiC-like tasks— achieved the best performance in the baseline method (see Table 9 in Appendix C for the complete baseline results). XL-Lexeme embeddings were subsequently used to train logistic regression classifiers, whose best results are reported in the first rows of Table 5. Despite optimizing both the layer and decision threshold for each language pair, this method underperformed the unsupervised baseline in all cases (except GL–PT, where it was one accuracy point higher), yielding an average accuracy of 70.16 versus the baseline's 77.80.

Zero vs. few-shot: For the LLMs, overall, few-shot prompting yields a consistent, but modest, improvement over zero-shot performance across the majority of the datasets (see Table 4). This stability suggests that the models' behavior relies more on pre-existing knowledge than on the specific examples in the prompt. Consequently, the remainder of our analysis focuses on the few-shot paradigm, as it represents the upper bound of the models' capabilities in this task.

Dataset	Zero-shot	Few-shot
ES–PT	82.80	**84.15**
GL–ES	80.65	**82.35**
GL–PT	74.45	**78.50**
Overall	79.30	**81.60**

Table 4: Comparison of zero-shot and few-shot mean performance (%) across datasets.

Open vs. closed-weight performance: While open-weight models like Gemma-3-27B, Qwen3-235B, and Llama-3.3-70B show competitive results, closed-weight models obtain the best performance, in particular, the Grok family achieving over 90% of accuracy (Table 5).

Model scale performance: Most models under 10B parameters struggle significantly with the task, particularly the Llama-3.2-3b (52.20%). Moreover, it should be noted that unsupervised encoder

[7]Overall, results tend to be slightly higher on the manual subset. Full results for both subsets are provided in the appendices.

baseline experiments outperform those obtained by models of this size. In the medium-sized category (10B–50B), performance stabilizes: Mistral-small-3.2-24b and Gemma-3-27b-it demonstrate high efficiency, with both exceeding the 80% accuracy. Notably, Gemma-3-27B surpasses much larger models, such as Llama-3.3-70B (85.70%) and DeepSeek-Chat-v3.1 (85.13%). This suggests that performance is driven not only by scale but also by architectural and data-driven factors. Specifically, these medium-sized models seem to have a more curated multilingual data mixture —as reflected in their Spanish-Portuguese performance— indicating that data quality and task-specific training are more influential than the total number of parameters. Our recall scores further support this: a larger scale does not guarantee semantic coverage. For instance, Gemma-3-27B achieves a higher recall in Galician subsets than Llama-3.3-70B (e.g., 0.87 vs. 0.78 in GL-ES and 0.87 vs. 0.81 in GL-PT). This suggests that massive-scale models may suffer from cross-lingual interference.

Language pair: Overall, linguistic distance tends to be inversely related to model performance: pairs with greater distance (e.g., ES–PT) are generally easier to resolve, while closely related pairs (e.g., GL–PT) tend to be more challenging, with ES–GL exhibiting intermediate performance. For instance, Gemini-2.0-flash exhibits a performance drop from 90.80% for the GL–ES pair to 81.75% in GL–PT. This degradation is expected, given the linguistic proximity between Galician and Portuguese, which increases the difficulty of identifying semantic divergences. Additionally, the substantially larger presence of Portuguese data in training corpora may introduce a bias toward Portuguese interpretations. Only the highest-performing closed models, such as Grok-4-fast, have demonstrated consistent efficiency on GL–PT (94.05%).

Analysis of classification bias: Beyond performance metrics, we analyze model behavior by measuring the deviation, defined as the absolute imbalance between `YES` and `NO` in predictions. Results in Table 12 (Appendix H which reports models with the highest deviation) show that performance drops correlate with systematic bias ($r = -0.73$, $\rho = -0.66$), rather than with random errors. This pattern is particularly prevalent in models under 10B parameters. Specifically, Gemma-2-9B is the only model exhibiting a consistent negative bias, with a tendency to over-label pairs as false friends. In contrast, models such as Llama-3.2-3b and Mixtral-8x7b show a strong positive bias, classifying the majority of instances as cognates.

6 Error Analysis

Following the general results from the previous section, we perform a more detailed error analysis to better understand model behaviour.

6.1 Error types

While the previous results included all 29 models, we now focus on a representative subset of five architectures for a more detailed analysis: Grok-4 (among the best performing models), Qwen-3-235B (high-capacity open-weight), Gemma-3-27B (high-efficiency), Llama-3.1-8B and Gemma-2-9B (representative of smaller open models). This selection allows us to determine how different models handle the linguistic phenomenon in our datasets.

Cognates and False Friends: As shown in the left section of Table 6, Gemma-2-9B emerges as an outlier, exhibiting a disproportionately high error rate for cognates. This pattern confirms a systematic bias toward over-predicting false friends, causing the model to fail at recognizing semantic equivalence.

Overall, the results indicate that performance varies across language pairs. For the ES-PT pair, partial false friends constitute the most challenging category. By contrast, for the pairs including Galician (GL-ES and GL-PT), total false friends dominate the error profile. This suggest that, in the presence of strong similarity, models have greater difficulty detecting diverging meanings. The full distribution per word type is detailed in Table 8 in Appendix A.

Part-of-Speech: A POS analysis was conducted to identify grammatical categories more prone to misclassifications, evaluated using category-specific error rates. Due to their dominance (91.47% of all instances, full distribution in Table 13, Appendix I), we focus on nouns (N), adjectives (A), and verbs (V). As shown in Table 6, no overall trend is observed across these categories, suggesting that performance is driven by model capacity and language pair, rather than by the specific POS.

6.2 Qualitative analysis

We complement the quantitative results with a qualitative analysis of individual model errors.

Encoder models						
Method	**Family**	**Model**	**ES–PT**	**GL–ES**	**GL–PT**	**AVG**
Unsup. Baseline	XLM (large)	XL-Lexeme	81.50	81.50	70.50	77.80
Log. Regression	XLM (large)	XL-Lexeme	75.50	63.50	71.50	70.16
Open-weight LLMs						
Size	**Family**	**Model**	**ES–PT**	**GL–ES**	**GL–PT**	**AVG**
Small [<10B]	Google	gemma-2-9b-it	72.35	69.00	68.85	70.00
		gemma-3-4b-it	73.55	78.00	71.35	74.30
	Meta	llama-3.1-8b-instruct	66.70	65.00	65.65	65.78
		llama-3.2-3b-instruct	52.10	50.00	54.50	52.20
	Microsoft	phi-3-mini-128k	68.50	66.00	72.90	70.15
	Mistral AI	ministral-3b	76.00	72.50	61.65	70.05
		mixtral-8x7b	68.85	76.00	62.10	68.98
Medium [10B–50B]	Google	gemma-3-12b-it	85.80	82.50	78.95	82.41
		gemma-3-27b-it	87.95	87.50	83.95	86.46
	Microsoft	phi-3-medium-128k	71.55	66.00	72.90	70.15
		phi-4-multimodal	86.80	86.00	82.50	85.10
		phi-4-reasoning-plus	84.30	83.50	79.50	82.43
	Qwen	qwen3-30b	89.40	81.50	78.95	83.28
	Mistral AI	mistral-small-24b	87.40	86.65	77.60	83.83
		mistral-small-3.2-24b	91.40	85.50	81.15	86.01
Large [>50B]		deepseek-chat-v3.1	91.05	83.50	80.55	85.13
	Qwen	qwen3-235b	90.40	90.00	89.05	89.81
		qwen3-next-80b	87.60	86.50	86.60	86.90
	Meta	llama-3.3-70b-instruct	91.75	84.50	80.85	85.70
Closed-weight LLMs						
	Google	gemini-2.0-flash	90.80	87.50	81.75	86.68
		gemini-2.5-flash	92.55	91.50	84.30	89.61
		gemini-2.5-flash-lite	82.90	91.50	78.55	84.28
	OpenAI	gpt-4o-mini	85.95	79.50	77.40	80.95
		gpt-4.1	88.40	93.00	87.87	89.55
		gpt-4.1-mini	91.25	89.00	81.75	87.33
	xAI	grok-3-mini	93.20	95.50	88.75	92.48
		grok-4	96.05	91.50	89.60	92.38
		grok-4-fast	95.40	96.00	94.05	95.15

Table 5: Accuracy (%) results across models and language pairs. LLM performance is based on few-shot prompting. The final column (AVG) denotes the macro-average accuracy across all datasets.

6.2.1 Most frequent misclassifications

To further investigate the qualitative nature of errors, we focus on word pairs most frequently misclassified across models, selecting those incorrectly classified by at least half of them. Table 7 reports the most problematic cases for each language pair. Overall, the analysis reveals that both cognates and false friends are major sources of error. For instance, models often treat *marmelada* (PT 'marmalade') and *mermelada* (ES 'jam') as equivalent, overlooking subtle lexical differences. Similarly, in the GL–ES pair, *almorzar* causes confusion due to mismatched meanings (GL 'breakfast' vs. ES 'lunch'). In GL–PT, *bacharelato* is misinterpreted, as it denotes different educational levels in each language. Many cases involve partial false friends, further complicating the task for the models.

6.2.2 Model reasoning and justification

To provide a more comprehensive analysis, we further examine models' justification and reasoning. Thus, we manually analyzed the rationales from the two smaller and generally underperforming models: Llama-3.1-8B and Gemma-2-9b.

Cross-lingual interference: Our analysis shows that a recurrent error arises from the models insufficient understanding of Galician, often confusing it with Spanish. For instance, on the mentioned *almorzar*, models justify their decision by stating that both sentences refer to lunch, projecting the Spanish meaning onto Galician term. Similar cross-lingual interference is observed in pairs such as *rato* (GL 'mouse' vs. ES 'a while') or *ano* (GL 'year' vs. ES 'anus'), where explanations consistently reflect the Spanish interpretation also in Galician.

Pair	Model	Error type (%)			POS errors (%)		
		COGN	PFFs	TFFs	A	N	V
ES–PT	llama-3.1-8b	10.99	65.71	46.75	23.33	43.22	13.95
	gemma-2-9b-it	45.60	8.57	1.30	50.00	16.58	37.21
	gemma-3-27b-it	6.59	25.71	16.88	12.22	12.56	25.58
	qwen3-235b	8.79	22.86	2.60	10.00	11.56	16.28
	grok-4	6.04	8.57	0.00	4.44	6.53	6.98
	AVERAGE	15.60	26.28	13.50	19.99	18.09	20.00
GL–ES	llama-3.1-8b	18.95	52.38	53.13	28.57	33.71	43.90
	gemma-2-9b-it	43.16	3.17	12.50	50.00	34.83	21.95
	gemma-3-27b-it	11.58	7.94	9.38	21.43	8.99	17.07
	qwen3-235b	2.11	11.11	21.88	7.14	10.11	14.63
	grok-4	3.16	3.17	3.13	7.14	3.37	4.88
	AVERAGE	15.79	15.55	20.00	22.85	18.20	20.48
GL–PT	llama-3.1-8b	18.95	47.62	56.25	12.20	38.61	42.31
	gemma-2-9b-it	54.74	4.76	15.63	48.78	30.69	34.62
	gemma-3-27b-it	13.68	14.29	25.00	19.51	17.82	15.38
	qwen3-235b	9.47	9.52	31.25	12.20	17.82	7.69
	grok-4	5.26	6.35	9.38	12.20	6.93	0.00
	AVERAGE	20.42	16.50	27.50	20.97	22.37	20.00

Table 6: Error analysis by error type (cognates, partial false friends, total false friends) and Part-of-Speech (adjectives, nouns, verbs).

Word-Pairs	Pair	Type	Err.
mermelada–marmelada	ES–PT	FF	27/29
individualizar–individualizar	ES–PT	COG	27/29
grasa–graxa	ES–PT	FF	24/29
presentar–apresentar	ES–PT	COG	23/29
batata-batata	ES–PT	FF	21/29
almorzar–almorzar	GL–ES	FF	25/29
almorzo–almuerzo	GL–ES	FF	23/29
costume–costumbre	GL–ES	COG	21/29
reserva–reserva	GL–ES	COG	20/29
bacharelato–bacharelato	GL–PT	FF	29/29
motorista–motorista	GL–PT	FF	27/29
presente–presente	GL–PT	COG	27/29
copo–copo	GL–PT	COG	27/29
sobrenome–sobrenome	GL–PT	FF	27/29

Table 7: Word pairs with the highest consensus misclassifications across models. Err. column are proportions of models (out of 29) that failed on each item.

Hallucinations: In addition, both models exhibit hallucinations when lexical knowledge is lacking. For example, *toro* (GL 'slice') is incorrectly described as a type of fish, suggesting that models do not know the actual meaning in Galician and they attempt to infer it from context.

Overgeneralization: We also observe cases where models try to justify similarity. For *oficina* (PT 'workshop' vs. ES 'office'), Llama-3.1-8B overgeneralized by claiming that both refer to 'a workplace where administrative tasks are carried out', ignoring the Portuguese sense. We also observed cases of semantic over-interpretation in Gemma-2-9B: For *húmido* (GL/PT 'humid'), the model incorrectly classified it as a false friend, arguing that 'the first refers to humid weather, while the second refers to a humid climate'. This finding suggests that, despite the model's ability to capture the overarching concept, its inherent bias may lead to the differentiation of identical concepts.

6.3 Translation analysis

In order to further investigate the potential implications of mishandling false friends, we perform an experiment on a machine translation task. We evaluate false friends retention by translating sentences using a subset of LLMs (see Appendix K). To automate interference detection, we lemmatize the output with FreeLing and automatically identify cases where the model preserves the false friend lemma in the target language, making the assumption that this involves an inadequate translation.

Our analysis shows that, while models generally handle frequent terms well, there are persistent cases of false friend interference. For instance, in Galician–Spanish translation, the Galician verb *chocar* (GL 'to brood') is incorrectly translated following its Spanish false friend (ES 'to collide'). Similarly, in PT–ES pairs, the term *salsa* (PT 'parsley') is mistranslated as *salsa* (ES 'sauce') instead of *perejil* (ES 'parsley'). These errors, while infrequent, can hamper the reliability of machine trans-

lation systems when dealing with closely-related languages.

7 Conclusion and Future Work

This paper presented a comprehensive evaluation of how current LLMs handle cross-lingual semantic ambiguity in related linguistic varieties: Galician, Portuguese and Spanish. For this purpose, we introduced a novel, WiC-style manually validated dataset of cognates and false friends. Our experimental results led to several findings. First, while specialized multilingual encoders like XL-Lexeme provide a strong baseline, they are often surpassed by medium and large generative models. Specifically, the few-shot setting provides a consistent improvement over zero-shot and embedding-based configurations.

Closed-weight models (e.g., Grok-4-fast) achieve superior performance, but large and medium-sized open-weight models (e.g., Gemma-3-27B or Qwen-235B) show competitive results. Notably, the results of models within this range (10B-50B) suggest that once a model reaches a certain size, performance is driven less by model scale and more by data-driven factors. However, a minimum scale is required to solve semantic interference in related varieties, as evidenced by the significant performance drop observed in smaller models (<10B).

The outcomes of a qualitative analysis show that greater linguistic distance between languages reduces semantic ambiguity, thereby facilitating disambiguation. Consequently, the GL–PT pair is the most challenging configuration across all experiments, indicating that languages with high lexical overlap pose difficulties for models both in distinguishing senses and in treating them as separate languages. More generally, false friends—both total and partial—remain a persistent challenge for current LLMs.

In future work, we plan to extend this study by expanding the datasets with more examples, richer contexts, and additional languages, and by transitioning from pairwise to language triplets. Future analyses may also consider the role of lexical frequency and degrees of ambiguity (e.g., as defined in WordNet) in model performance. Finally, extending the assessment to additional downstream tasks (expanding the analysis on machine translation) would help assess the impact in real-world scenarios.

Limitations

Data: In our study there are some limitations regarding the dataset. First, the use of sentences from Wikipedia may not fully reflect language in other domains. Second, while the datasets are balanced, the overall number of instances remains relatively small. Similarly, the fact that we deal with three specific languages pairs may prevent us from drawing more general conclusions about the cross-lingual capabilities of the evaluated models. Furthermore, the datasets are focused on nouns, adjectives, and verbs, with a limited representation of other lexical categories. Finally, there is an inherent risk of data contamination, as the false-friends sources and the contextual sentences might have been included in the model's pre-training corpora.

Models: In our evaluation we include closed-weight models for which information regarding the architecture and training data is not publicly available. Additionally, our prompt-based experiments are limited to specific zero and few-shot configurations; alternative strategies could yield different performance patterns.

Justification analysis: In this study, we ask models to provide a short justification for their predictions. However, we do not conduct a quantitative analysis of the logical consistency of these justifications. For that reason, there is the possibility of models hallucinating while still producing the correct answer. A rigorous human evaluation is required to confirm whether models understand these semantic distinctions.

Acknowledgments

This paper was funded by MCIU/AEI/10.13039/501100011033 (grants with references PID 2021-128811OA-I00, PID2024-161928OB-I00, CNS2024-154902, and AIA2025-163322-C62), by the Galician Government (ED481A-2024-070, ED431G 2023/04 and ED431B 2025/16), and by the *Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia* - Funded by EU — NextGenerationEU within the framework of the project *Desarrollo Modelos ALIA*. Jose Camacho-Collados was supported by a UKRI Future Leaders Fellowship.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, and 4 others. 2025. Phi-4-reasoning technical report. *Preprint*, arXiv:2504.21318.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Marta Vázquez Abuín and Marcos Garcia. 2025. WiC evaluation in Galician and Spanish: Effects of dataset quality and composition. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 172–178, Suzhou, China. Association for Computational Linguistics.

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2024. Automated cognate detection as a supervised link prediction task with cognate transformer. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 965–975, St. Julian's, Malta. Association for Computational Linguistics.

Keith Allan. 2009. *Concise encyclopedia of semantics*. Elsevier.

Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. 2019. CogNet: A large-scale cognate database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145, Florence, Italy. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Iago Bragado Trigo. 2006. Sobre a amizade (léxica) Galiza-Portugal: os falsos amigos Galego-Portugués-Español. *Madrygal. Revista de Estudios Gallegos*, 9:33–42.

Pascal Brenders, Janet G. van Hell, and Ton Dijkstra. 2011. Word recognition in child second language learners: Evidence from cognates and false friends. *Journal of Experimental Child Psychology*, 109(4):383–396.

Samuel Cahyawijaya, Ruochen Zhang, Jan Christian Blaise Cruz, Holy Lovenia, Elisa Gilbert, Hiroki Nomoto, and Alham Fikri Aji. 2025. Thank you, stingray: Multilingual large language models can not (yet) disambiguate cross-lingual word senses. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3228–3250, Albuquerque, New Mexico. Association for Computational Linguistics.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Santiago Castro, Jairo Bonanata, and Aiala Rosá. 2018. A high coverage method for automatic false Friends detection for Spanish and Portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 29–36, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pedro J Chamizo-Domínguez. 2012. *Semantics and pragmatics of false friends*. Routledge.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Pedro J Chamizo Dominguez and Brigitte Nerlich. 2002. False friends: their origin and semantics in some selected languages. *Journal of pragmatics*, 34(12):1833–1849.

María del Pilar Durán Escribano. 2004. Exploring cognition processes in second language acquisition: the case of cognates and false-friends in est. *Ibérica (Madrid)*, 7(1):87–106.

Tiago Vidal Figueroa. 1995. Presuntos falsos amigos entre portugués e galego. I. *Viceversa. Revista galega de tradución*, pages 145–152.

Marcos Garcia, Carlos Gómez-Rodríguez, and Miguel A Alonso. 2018. New treebank or repurposed? on the feasibility of cross-lingual parsing of romance languages with universal dependencies. *Natural Language Engineering*, 24(1):91–122.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2525–2529, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Julie Kallini, Dan Jurafsky, Christopher Potts, and Martijn Bartelds. 2025. False Friends are not foes: Investigating vocabulary overlap in multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21138–21154, Suzhou, China. Association for Computational Linguistics.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681, Toronto, Canada. Association for Computational Linguistics.

Nikola Ljubešić and Darja Fišer. 2013. Identifying false friends between closely related languages. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 69–77, Sofia, Bulgaria. Association for Computational Linguistics.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.

George A. Miller, Richard Beckwith, Christiane D. Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Ruslan Mitkov, Viktor Pekar, Dimitar Blagoev, and Andrea Mulloni. 2007. Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation*, 21(1):29–53.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Mitko Nikov, Žan Tomaž Šprajc, and Žan Bedrač. 2024. *Cross-Lingual False Friend Classification via LLM-based Vector Embedding Analysis*, page 33–36. Univerzitetna založba Univerze v Mariboru.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2473–2479, Istanbul, Turkey. European Language Resources Association (ELRA).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos,

D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Carlos Periñán-Pascual and Nicolás José Fernández Martínez. 2025. Detección de cognados verdaderos y falsos amigos con word embeddings. *Revista Signos. Estudios de Lingüística*, 58(118).

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.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Roberto Samartim. 2012. Língua somos: A construção da ideia de língua e da identidade coletiva na galiza (pré-) constitucional. In *Novas achegas ao estudo da cultura galega II: enfoques socio-históricos e lingüístico-literarios*, pages 27–36. Universidade de Santiago de Compostela, Santiago de Compostela.

Lianet Sepúlveda Torres and Sandra Maria Aluísio. 2011. Using machine learning methods to avoid the pitfall of cognates and false friends in Spanish-Portuguese word pairs. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Ana Sabina Uban and Liviu P. Dinu. 2020. Automatically building a multilingual lexicon of false Friends with no supervision. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3001–3007, Marseille, France. European Language Resources Association.

Ana Sabina Uban, Liviu P Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Claudia Vlad. 2025. Friend or Foe? A Computational Investigation of Semantic False Friends across Romance Languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15309–15323, Suzhou, China. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

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.

A Complete Dataset Statistics

Dataset	Pair	COG	TFFs	PFFs	Tot.
Auto	ES-PT	152	66	87	304
Auto	GL-ES	50	30	20	100
Auto	GL-PT	65	22	43	130
Human	ES-PT	30	11	19	60
Human	GL-ES	25	16	9	50
Human	GL-PT	30	10	20	60

Table 8: Distribution of cognates (COG), total false friends (TFF), and partial false friends (PFF) for each language pair and data source (automatic or human created). The table also reports the total number of word pairs in each subset (Tot.).

B False Friends Resources

Online Resources:

- https://ec.europa.eu/translation/portuguese/magazine/documents/folha47_lista_pt.pdf
- https://github.com/pln-fing-udelar/false-friends/tree/master
- https://www.lingua.gal/c/document_library/get_file?file_path=/portal-lingua/curso/medio-administrativo/_13_falsosamigos.pdf

Academic Works and Books:

- Ramos, E. *Portugalizar. Portugués para galegofalantes.*
- Vidal Figueiroa, T. *Presuntos falsos amigos entre portugués e galego.* (VOL. I-II-III)

C Baseline experiments

Model	Pair	L.	Thr.	Acc.	F1
bert-base-multilingual-uncased					
	ES–PT	12	0.58	0.73	0.75
	GL–ES	12	0.62	0.79	0.76
	GL–PT	8	0.70	0.58	0.68
	H–ES–PT	12	0.56	0.73	0.77
	H–GL–ES	9	0.76	0.74	0.78
	H–GL–PT	0	0.76	0.57	0.58
bert-base-multilingual-cased					
	ES–PT	9	0.74	0.69	0.72
	GL–ES	8	0.62	0.71	0.76
	GL–PT	10	0.74	0.53	0.66
	H–ES–PT	8	0.68	0.72	0.76
	H–GL–ES	12	0.52	0.76	0.76
	H–GL–PT	7	0.70	0.60	0.71
pierluigic/xl-lexeme					
	ES–PT	13	0.82	0.81	**0.83**
	GL–ES	14	0.82	0.83	**0.83**
	GL–PT	16	0.80	0.68	**0.72**
	H–ES–PT	20	0.92	0.82	**0.83**
	H–GL–ES	15	0.82	0.80	**0.82**
	H–GL–PT	14	0.84	0.73	**0.76**
xlm-roberta-base					
	ES–PT	6	0.84	0.69	0.74
	GL–ES	3	0.88	0.70	0.69
	GL–PT	12	0.98	0.54	0.68
	H–ES–PT	6	0.86	0.68	0.72
	H–GL–ES	0	0.22	0.66	0.74
	H–GL–PT	5	0.80	0.60	0.71
xlm-roberta-large					
	ES–PT	14	0.86	0.73	0.74
	GL–ES	13	0.86	0.72	0.75
	GL–PT	13	0.84	0.63	0.71
	H–ES–PT	13	0.88	0.75	0.78
	H–GL–ES	11	0.88	0.68	0.75
	H–GL–PT	13	0.88	0.68	0.75

Table 9: Results for baseline experiments for all the models. We report the optimal transformer layer *(L)* and the decision threshold *(Thr.)*, alongside classification Accuracy *(Acc.)* and F1-score for each language pairs. Bold values indicate the highest F1-score.

D Explained results

Encoder models									
Method	**Family**	**Model**	**ES–PT**	**GL–ES**	**GL–PT**	**H-ES–PT**	**H-GL–ES**	**H-GL–PT**	**AVG**
Unsup. Baseline	XLM (large)	XL-Lexeme	81.25	83.00	68.46	81.67	80.00	73.33	77.80
Log. Regression	XLM (large)	XL-Lexeme	73.36	67.00	70.00	78.30	64.00	73.30	70.16
Open-weight LLMs									
Size	**Family**	**Model**	**ES–PT**	**GL–ES**	**GL–PT**	**H-ES–PT**	**H-GL–ES**	**H-GL–PT**	**AVG**
Small [<10B]	Google	gemma-2-9b-it	71.40	68.00	67.70	73.30	70.00	70.00	70.00
		gemma-3-4b	70.40	84.00	67.70	76.70	72.00	75.00	74.30
	Meta	llama-3.1-8b	65.10	62.00	64.60	68.30	68.00	66.70	65.78
		llama-3.2-3b	55.90	52.00	52.30	48.30	48.00	56.70	52.20
	Microsoft	phi-3-mini-128k	64.50	67.00	48.50	65.00	68.00	70.00	63.83
	Mistral AI	ministral-3b	73.70	75.00	60.00	78.30	70.00	63.30	70.05
		mixtral-8x7b	69.40	76.00	59.20	68.30	76.00	65.00	68.98
Medium [10B–50B]	Google	gemma-3-12b-it	81.60	81.00	76.20	90.00	84.00	81.70	82.41
		gemma-3-27b-it	85.90	87.00	84.60	90.00	88.00	83.30	86.46
	Microsoft	phi-3-medium-128k	71.40	70.00	70.80	71.70	62.00	75.00	70.15
		phi-4-multimodal	80.30	84.00	80.00	93.30	88.00	85.00	85.10
		phi-4-reasoning-plus	81.90	83.00	72.30	86.70	84.00	86.70	82.43
	Qwen	qwen3-30b-a3b	85.50	83.00	76.20	93.30	80.00	81.70	83.28
	Mistral AI	mistral-small-24b	86.50	87.00	76.90	88.30	86.00	78.30	83.83
		mistral-small-3.2-24b	87.80	85.00	82.30	95.00	86.00	80.00	86.01
Large [>50B]	DeepSeek	deepseek-chat-v3.1	88.80	85.00	80.00	93.30	82.00	81.70	85.13
	Qwen	qwen3-235b	87.50	88.00	83.10	93.30	92.00	95.00	89.81
		qwen3-next-80b	85.20	87.00	81.50	90.00	86.00	91.70	86.90
	Meta	llama-3.3-70b	88.50	87.00	80.00	95.00	82.00	81.70	85.70
Closed-weight LLMs									
	Google	gemini-2.0-flash	84.90	91.00	78.50	96.70	84.00	85.00	86.68
		gemini-2.5-flash	91.80	92.00	86.90	93.30	92.00	81.70	89.61
		gemini-2.5-flash-lite	80.60	87.00	75.40	85.00	96.00	81.70	84.28
	OpenAI	gpt-4o-mini	85.20	79.00	73.10	86.70	80.00	81.70	80.95
		gpt-4.1	86.80	90.00	86.20	90.00	96.00	88.30	89.55
		gpt-4.1-mini	90.80	86.00	78.50	91.70	92.00	85.00	87.33
	xAI	grok-3-mini	93.10	93.00	89.20	93.30	98.00	88.30	92.48
		grok-4-fast	95.40	93.00	89.20	96.70	90.00	90.00	92.38
		grok-4	94.10	96.00	93.10	96.70	96.00	95.00	95.15

Table 10: Accuracy(%) results across models and language pairs for semi-automatic and manual constructed datasets (H). LLM performance is based on few-shot prompting. The final column (AVG) denotes the macro-average accuracy across all datasets.

E Zero-shot prompt

```
Task: Decide whether the target words have the same sense in both sentences.
TARGET WORD 1: {w1}   SENTENCE 1: {s1}
TARGET WORD 2: {w2}   SENTENCE 2: {s2}
Rules:
- Analyze only the meaning of the target words in their respective contexts.
- Ignore grammatical or syntactic differences unless they affect the meaning.
- If the meaning of the target words is the same, respond with: “Label: YES”.
- If the meaning of the target words is different, respond with: “Label: NO”.
- Provide a concise explanation for your decision in the format: “Reason: <short sentence>”.
- Ensure the output is strictly in plain text and follows the exact format specified.
```

F Few-shot prompt

```
Task: Decide whether the target words have the same sense in both sentences.
TARGET WORD 1: w1   SENTENCE 1: s1
TARGET WORD 2: w2   SENTENCE 2: s2
Examples:
Example 1:
TARGET WORD 1: fabricar SENTENCE 1: A súa empresa fabrica cadeiras de madeira
TARGET WORD 2: fabricar SENTENCE 2: Eles fabricam brinquedos para as crianças.
Label: YES Reason: Both refers to the action of making. Example 2:
TARGET WORD 1: brinco SENTENCE 1: María deu un brinco de alegría cando lle dixeron que ía ser nai.   TARGET WORD 2: brinco SENTENCE 2: Ela colocou um brinco dourado para combinar com o vestido.
Label: NO Reason: The first refers to a jump of joy, while the second refers to a piece of jewelry.
Rules:
Analyze only the meaning of the target words in their respective contexts.
- Ignore grammatical or syntactic differences unless they affect the meaning.
- If the meaning of the target words is the same, respond with: "Label: YES"
- If the meaning of the target words is different, respond with: "Label: NO"
- Provide a concise explanation for your decision in the format: "Reason: <short sentence>"
- Ensure the output is strictly in plain text and follows the exact format specified.
```

G Mean Zero vs. Few-shot

Dataset	ZERO mean	FEW mean
ES–PT	81.00	82.50
GL–ES	78.70	82.10
GL–PT	72.30	76.30
H–ES–PT	84.60	85.80
H–GL–ES	82.60	82.60
H–GL–PT	76.60	80.70
Overall	79.30	81.60

Table 11: Comparison between zero-shot and few-shot mean performance (%) across the six datasets.

H Deviation

Model	Pair	Dev.	Acc.
llama-3.2-3b-instruct	GL–ES	0.75	50.00
mixtral-8x7b-instruct	GL–PT	0.67	62.10
gemma-2-9b-it	GL–ES	0.64	69.00
llama-3.2-3b-instruct	GL–PT	0.63	54.50
mixtral-8x7b-instruct	ES–PT	0.57	68.85
ministral-3b	GL–PT	0.57	61.65
llama-3.2-3b-instruct	ES–PT	0.54	52.10
gemma-2-9b-it	GL–PT	0.48	68.85
gemma-2-9b-it	ES–PT	0.48	71.35
llama-3.1-8b-instruct	ES–PT	0.47	66.70
mixtral-8x7b-instruct	GL–ES	0.40	72.50
ministral-3b	GL–ES	0.37	76.00

Table 12: Models with the highest prediction deviation (Dev.) across language pairs (and corresponding overall accuracy). Higher values indicate a systematic bias toward one class.

I Part-of-Speech distribution

POS	ES–PT		GL–ES		GL–PT	
	n	%	*n*	%	*n*	%
A	90	24.7	14	9.3	41	21.6
A/N	18	4.9	4	2.7	17	8.9
ART/N	0	0.0	0	0.0	2	1.1
C	5	1.4	0	0.0	0	0.0
PRON	1	0.3	0	0.0	0	0.0
N	199	54.7	89	59.3	101	53.2
N/NUM	1	0.3	0	0.0	0	0.0
N/P	1	0.3	2	1.3	0	0.0
N/R	1	0.3	0	0.0	0	0.0
N/V	1	0.3	0	0.0	3	1.6
R	4	1.1	0	0.0	0	0.0
V	43	11.8	41	27.3	26	13.7
TOTAL	**364**		**150**		**190**	

Table 13: Distribution of Part-of-Speech (POS) tags across all the datasets, showing both raw counts (n) and percentages (%). Abbreviations include Adjectives (A), Articles (ART), Conjunctions (C), Pronouns (PRON), Nouns (N), Numerals (NUM), Preposition (P), Adverbs (R), and Verbs (V). Combined tags indicate lexical categories that vary between the two languages.

J Part-of-Speech error rates

Pair	Model	A (%)	N (%)	V (%)
ES–PT	llama-3.1-8b	23.33	43.22	13.95
	gemma-2-9b-it	50.00	16.58	37.21
	gemma-3-27b-it	12.22	12.56	25.58
	qwen3-235b	10.00	11.56	16.28
	grok-4	4.44	6.53	6.98
GL–ES	llama-3.1-8b	28.57	33.71	43.90
	gemma-2-9b-it	50.00	34.83	21.95
	gemma-3-27b-it	21.43	8.99	17.07
	qwen3-235b	7.14	10.11	14.63
	grok-4	7.14	3.37	4.88
GL–PT	llama-3.1-8b	12.20	38.61	42.31
	gemma-2-9b-it	48.78	30.69	34.62
	gemma-3-27b-it	19.51	17.82	15.38
	qwen3-235b	12.20	17.82	7.69
	grok-4	12.20	6.93	0.00

Table 14: Error rates (%) per Part-of-Speech category (A = adjectives, N = nouns, V = verbs). Each value represents the percentage of misclassified instances within each category for each model and language pair.

K Translation models

- Qwen3-235B-A22B-2507[8]
- GPT-4o-mini[9]
- Grok-4.1-Fast[10]

[8] https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507

[9] https://platform.openai.com/docs/models/gpt-4o-mini

[10] https://x.ai/news/grok-4-1-fast