LBAD: Demonstrating the Effectiveness of Commercial Large Language Models for Anglicism Detection

Alex Lyman[1]

[1]Brigham Young University

Abstract

We present **LBAD**, an **L**LM-**B**ased **A**nglicism **D**etector, our submission to the ADoBo 2025 shared task at IberLEF 2025. We show that, with careful prompting, commercially available LLMs achieve state-of-the-art performance on the task of Anglicism detection. We systematically evaluate the impact of model selection and prompting, including levels of instruction, chain-of-thought reasoning, few-shot learning, and self-refinement. Our experiments reveal that prompt specificity and model selection are critical, with F1 validation scores varying by up to 75 percentage points depending on these two factors. These findings highlight the practical effectiveness of commercial LLMs for lexical borrowing detection and provide a generalizable framework for leveraging LLMs in similar natural language processing tasks.

Keywords

Anglicisms, Large Language Models, Prompting

1. Introduction

Lexical borrowing, the process by which words from one language are incorporated into another, is a pervasive linguistic phenomenon driven by cultural, social, and technological interactions between linguistic communities [1, 2]. In recent times, the influence of English as a global lingua franca has led to a significant influx of Anglicisms into many languages worldwide [3]. The automatic detection of these borrowed terms, particularly unassimilated ones that retain their original orthography and are not yet fully integrated into the recipient language's lexicon, presents a challenge for Natural Language Processing (NLP). Identifying such borrowings is important for downstream applications, including lexicography, machine translation [4], text-to-speech synthesis, [5], and linguistic parsing [6].

The ADoBo (Automatic Detection of Borrowings) shared task series aims to foster research and development in this area. The second edition of the ADoBo shared task at IberLEF 2025 [7] specifically focuses on retrieving Anglicisms from Spanish text [8]. Participants are tasked with developing systems to identify spans of text in Spanish documents that correspond to words recently imported from English, such as "running," "smartwatch," "fake news," or "youtuber" [9].

IberLEF 2025, September 2025, Zaragoza, Spain

alexlyman@byu.edu (A. Lyman)
https://alexlyman.org/ (A. Lyman)
0009-0004-1210-2538 (A. Lyman)

The challenge of automatically detecting unassimilated borrowings, especially Anglicisms, in Spanish has been approached from various perspectives. Early computational methods often relied on dictionary lookups and rule-based systems [10], which, while useful, can struggle with neologisms and context-dependent borrowings [11]. More recent approaches have explored machine learning techniques including Conditional Random Fields, language models, and other neural sequence labeling systems [12, 13, 14].

In the years since the first ADoBo challenge, Large Language Models (LLMs) have become the de facto approach for a variety of research tasks across disciplines. Recent LLMs have demonstrated human-level (or superhuman) performance on data coding [15, 16] and data analysis [17], and are now being explored to replace research assistants altogether [18, 19].

While computational NLP tasks have typically been somewhat inaccessible to the general public, LLMs provide a powerful new tool for non-technical practitioners. LLMs are much more accessible than traditional machine learning methods. [20]. "Programming" an LLM takes place in the form of a prompt, as LLMs are typically explicitly trained to follow natural language instructions [21]. Additionally, commercial LLMs can be accessed by anyone with an internet connection, which means prototyping can begin in a matter of minutes. LLMs have vast stores of knowledge encompassing syntax, semantics, and commonsense world knowledge, usually in several languages [22]. When accessed through an API, querying these models is generally cost effective, with a typical query for our experiments costing a fraction of a cent. We leverage these strengths, guided by experimentation across prompts, to create our entry.

Built on commercial LLMs, our submission to the ADoBo 2025 shared task achieves state-of-the-art results. We provide a detailed description of our prompt creation process, as well as experiments across prompt variations and model sizes. While our main focus is on detailing our approach to prompting LLMs for the task of Anglicism detection in Spanish, we hope to provide a generalizable example of using commercial LLMs for NLP tasks.

2. Experiments

We perform experiments across models and prompt variations. Prompt variations include:

- Baseline prompting
- Prompt with detailed guidelines
- Chain-of-thought prompting
- Few Shot prompting (with a varying number of in-context examples)
- Self-refined prompt

In order to measure the effect of model size, we test on a full sweep of OpenAI's latest generation of commercially available models, including reasoning models [23, 24]:

- GPT 4.1 nano
- GPT 4.1 mini
- GPT 4.1
- o4 mini
- o3

The ADoBo-25 challenge has no official train set, only a development set and a final test set. For testing purposes, we adapt the ADoBo-25 dev set into our own train, test, and validation sets, which we call **LBAD**-train, **LBAD**-test, and **LBAD**-valid.[1] We do this by randomly shuffling the ADoBo-25 development set and splitting it into thirds. We draw from **LBAD**-train to find example Anglicisms for our prompts. Because prompt creation is an iterative process, we test our prompts and iterate using the **LBAD**-test set. Finally, we report results on all experiments in this paper are on **LBAD**-valid. Because some training and testing occurs on parts of the ADoBo-25 dev set, we do not report results on the complete dev set. However, because **LBAD**-valid is randomly drawn from the ADoBo-25 dev set, we assert that results on **LBAD**-valid are a reasonable proxy for results on the complete dev set.

2.1. Prompt Variations

2.1.1. Baseline Prompt

All of the prompts in this experiment are built off of the following baseline prompt, adapted from the ADoBo 2025 task description:

```
Given the following Spanish sentence, identify all anglicisms.

An anglicism is a word or multi-word expression borrowed specifically from
English that has recently been imported into the Spanish language and is
used without orthographic adaptation.

Examples of anglicisms include 'running', 'smartwatch', 'influencer',
'country managers', 'marketing'.

This is the sentence to evaluate:
{sentence}

Output only the identified anglicisms, separated by semicolons. If no
anglicisms are found, output 'None'.
```

2.1.2. Detailed Guidelines

Because modern LLMs are explicitly trained to follow natural language instructions, properly formalizing and explaining the task is necessary to ensure peak performance.

To create a set of comprehensive guidelines, we pass the entire lexical borrowing annotation guidelines from the ADoBo 2021 challenge [11] to R1 1776 [25], a reasoning LLM fine-tuned from DeepSeek R1 [26]. We instruct R1 to comb through the annotation guidelines and summarize them into a step-by-step protocol for identifying Anglicisms in Spanish text. The full text of these detailed guidelines can be found in the appendix. We insert the guidelines into the baseline prompt before the sentence to evaluate.

[1] Replication materials are available at https://github.com/AlexMLyman/LBAD-ADoBo-2025

2.1.3. Chain-of-Thought

We perform experiments using Chain-of-thought prompting [27], a prompting technique that encourages LLMs to think step-by-step before outputting a final answer. Chain-of-thought prompting tends to increase model performance across a variety of domains, especially on complex tasks. We append the following instructions to the end of the prompt to elicit chain-of-thought reasoning:

```
First, think step-by-step about which words/phrases might be anglicisms and
why.
After thinking things through, on a new line, output only the identified
anglicisms, separated by semicolons. If no anglicisms are found, output
'None'.
```

2.1.4. Few-Shot

Language models are capable of in-context learning [22], (ICL) where the model is able to perform a task which is demonstrated in the prompt. Because models can learn from a few examples, this is sometimes referred to as few-shot learning. In-context learning can require variation in number and type of examples [28].

We test prompt variations with five, ten, and twenty-five ICL examples. ICL examples were selected from the **LBAD**-train dataset using R1 1776, which was instructed to reason through the entire **LBAD**-train dataset and select the five, ten, or twenty-five most representative examples of Anglicism detection. Examples were added to the baseline prompt with the following addition (right before the sentence to evaluate):

```
Here are some example sentences, with the correct output on the line
following the sentences:
```

2.1.5. Self-Refinement

LLMs are capable of refining their own outputs, as well as those of other LLMs. Several self-refinement frameworks allow LLMs to recursively refine their own prompts [29, 30].

Inspired by these frameworks, we begin by testing GPT 4.1 using the prompt with detailed guidelines on the **LBAD**-test dataset. We then pass those answers and the ground truth to R1 1776, and instruct it to classify GPT 4.1's failure modes and then write a series of helpful reminders which can be added to the prompt before validation. These reminders are added to the prompt after the detailed guidelines. The full text of the reminders can be found in the appendix.

2.2. Model Selection

LLM performance can vary wildly depending on the size of the model, with smaller models tending to perform worse than models with more trainable parameters [31]. To quantify this effect, we performed all of the above experiments using all three sizes of GPT 4.1 models (nano, mini, and standard).

Reasoning models [32] are a recent development. These models are explicitly trained to think through a reasoning process in a manner similar to chain-of-thought reasoning. We test both of OpenAI's reasoning models, o4-mini and o3 on our baseline prompt, prompt with instructions, and our prompt with instructions and reminders.

3. Results

We report F1 scores on **LBAD**-valid for all experiments. Following the convention from the task leaderboard, we round all scores to the nearest whole percent. To facilitate comparison with other entries, we also report precision, accuracy, and F1 scores on the ADoBo 2025 test set for our best performing prompt/model combination. Tables of all scores, including precision, recall, and F1 can be found in the appendix.

An examination of the detailed score breakdown tells us something about commercial LLMs' strengths. Across all models and prompts, recall scores are high. Most variation in F1 scores is due to poor precision. In other words, **LBAD** tends to find almost all anglicisms, but often misidentifies non-anglicism spans.

We see a gap in F1 scores between **LBAD**-valid and the ADoBo-2025 test set, with **LBAD**-valid scores being lower. This is attributable to the fact that roughly half of **LBAD**-valid contains no anglicisms, but the ADoBo-2025 test set is unbalanced in favor of sentences with Anglicisms. As a result, the ADoBo-2025 test set is more in line with **LBAD**'s strengths.

3.1. Qualitative Analysis of Failure Modes

Qualitative analysis of the best performance on ADoBo-2025 reveals where even the top-performing model struggles. The o3 model with Guidelines and Reminders fails 25 times. These failures comprise two main failure modes.

3.1.1. Multi-Word Spans

Several of the failures result from the model misidentifying a multi-word span as two single word spans, or identifying two single-word spans as one multi-word span. For example in the sentence `"Casual Looks" con bufanda y guantes para triunfar esta temporada` the model is supposed to identify `Casual Looks` as the Anglicism, but the model identifies `Casual` and `Looks` separately. Seven of the 25 errors (28%) have to do with incorrectly parsing span boundaries.

3.1.2. Confusing Spanish and English words.

Many words in this set are orthographically identical in English and Spanish. (e.g. `normal`, `total`, `error`) In these cases, the model will misidentify one or more Anglicisms as Spanish words. As an example, given the sentence `Durante su carrera profesional, también ha sido socio y director general de Ikea, global director de Apple y director`

Table 1
F1 Score Results on **LBAD**-valid. Highest score for each model is bolded.

	4.1 Nano	4.1 Mini	4.1	o4 mini	o3
Baseline	31	38	51	47	45
Baseline + Guidelines	24	49	70	81	**90**
Baseline + Guidelines + Reminders	21	64	65	**82**	89
Baseline + CoT	48	52	57	-	-
Baseline + CoT + Guidelines	52	74	**80**	-	-
Baseline + CoT + Guidelines + Reminders	**53**	**82**	**80**	-	-
Baseline + 5 ICL Examples	17	41	56	-	-
Baseline + Guidelines + 5 ICL Examples	14	47	62	-	-
Baseline + 10 ICL Examples	24	45	60	-	-
Baseline + Guidelines + 10 ICL Examples	20	49	70	-	-
Baseline + 25 ICL Examples	12	46	55	-	-
Baseline + Guidelines + 25 ICL Examples	12	45	64	-	-

`de Comunicación de Basket Market.` the model incorrectly guesses no Anglicisms are present, though the ground truth is `global director`. Because `global` and `director` are orthographically the same in Spanish and English, the model fails to detect this Anglicism although the syntax (`global` before `director` is English syntax) provides a clue that this is an Anglicism. Seventeen of the 25 errors (68%) are partial or complete misidentifications of words that share cross-lingual orthography.

These two failure modes explain almost all of LBAD's errors on the ADoBo-25 test set. It is likely that additional rounds of prompt refinement could help the model avoid these failure modes in the future.

3.2. Results by Prompting Technique

3.2.1. Detailed Guidelines

Anglicism detection is a challenging task with many particulars and precise rules. The baseline prompt relies heavily on models' inherent knowledge of Anglicisms, since it provides only an outline of the task. We see evidence of this in the fact that baseline performance tends to increase with larger models. Larger models have more parameters in which to store information about the world. Consequently, they have more world knowledge across many topics. This world knowledge likely includes information about NLP and Anglicisms.

We find that an in-depth task description, provided in the form of detailed guidelines, is the best way to teach the task to all LLMs tested. Across models, detailed guidelines tended to cause an increase in performance between 10 and 45 percentage points over the baseline prompt. This increase was larger for larger models, which are better able to think through the Anglicism identification process.

Table 2
Final Results on ADoBo-25 test set.

	Precision	Recall	F1
o3 Baseline + Guidelines + Reminders	99	99	99

3.2.2. Chain-of-Thought

By far, the most important prompt variation for improving performance was chain-of-thought reasoning. Whether prompt-based or (in the case of reasoning models) explicitly trained, models utilizing chain-of-thought reasoning performed best. This approach worked best in conjunction with detailed guidelines, as the detailed guidelines provided a prototype for model reasoning, and the chain-of-thought helped the models think through the process.

3.2.3. Few-Shot

Few-shot learning was less effective than chain-of-thought reasoning, regardless of the number of ICL examples. More examples were not necessarily better. Because Anglicism detection involves many subtleties, it is difficult to learn all of the rules from examples. We note that ICL examples do not appear to provide additional information when combined with detailed guidelines. To the contrary, ICL examples when combined with the detailed guidelines often caused a drop in performance as opposed to the baseline prompt with guidelines, probably by "confusing" the LLM.

3.2.4. Self-Refinement

Self-Refinement through the form of helpful reminders often increases performance across models, though the gains are modest. Three of the models improve with reminders, one achieves the same score, and one decreases by one percentage point. This may be due to a ceiling effect, as models with worse initial performance gain more from the reminders. These modest results should not discourage practitioners from trying iterative prompt refinement with LLMs. It is probable that additional human-based refinement of the reminders would result in greater performance gains, but that is beyond the scope of our experiments.

4. Conclusion

LBAD was the best performing system on the ADoBo-2025 challenge, and was built entirely off of commercial LLMs. Language models were used not only as the method of evaluation, but also to summarize the task guidelines, to analyze failure modes, and to select in-context examples. While LLMs are demonstrably capable of achieving state-of-the-art results, we emphasize that this capability depends heavily on model and prompt selection. In our experiments, F1 scores on **LBAD**-valid ranged from 12 to 90 simply by varying prompt and model choice. This 78 percentage point spread highlights the importance of testing prompts and iterating when using

LLMs for research. The fact that a carefully prompted LLM can achieve SOTA performance on this task should encourage researchers to consider using LLMs for NLP.

5. Declaration on Generative AI

The topic of this work is Generative AI (LLMs) and explores LLM capabilities in the context of Anglicism Detection. The only use of generative AI on this project was the LLM-based pipeline described in the paper. AI tools were not used in the drafting or editing of this manuscript.

References

[1] E. Haugen, The analysis of linguistic borrowing, Language 26 (1950) 210–231. doi:10.2307/410058.

[2] S. Poplack, D. Sankoff, C. Miller, The social correlates and linguistic processes of lexical borrowing and assimilation, Linguistics 26 (1988) 47–104. doi:10.1515/ling.1988.26.1.47.

[3] C. Furiassi, V. Pulcini, F. Rodríguez González (Eds.), The Anglicization of European Lexis, John Benjamins Publishing, 2012. doi:10.1075/z.174.

[4] Y. Tsvetkov, C. Dyer, Cross-lingual bridges with models of lexical borrowing, Journal of Artificial Intelligence Research 55 (2016) 63–93. doi:10.1613/jair.4786.

[5] S. Leidig, T. Schlippe, T. Schultz, Automatic detection of anglicisms for the pronunciation dictionary generation: A case study on our german it corpus, in: The 4th Workshop on Spoken Language Technologies for Under-resourced Languages, St. Petersburg, Russia, 2014. URL: https://www.isca-archive.org/sltu_2014/leidig14_sltu.html, sLTU 2014.

[6] B. Alex, Automatic Detection of English Inclusions in Mixed-Lingual Data with an Application to Parsing, Ph.D. thesis, University of Edinburgh, 2008. URL: https://homepages.inf.ed.ac.uk/balex/publications/thesis.pdf.

[7] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[8] E. Álvarez-Mellado, J. Porta-Zamorano, C. Lignos, J. Gonzalo, Overview of ADoBo at IberLEF 2025: Automatic Detection of Anglicisms in Spanish, Procesamiento del Lenguaje Natural 75 (2025).

[9] ADoBo Task Organizers, ADoBo — Automatic Detection of Borrowings, https://adobo-task.github.io/, n.d. Accessed: June 4, 2025.

[10] J. R. L. Serigos, Applying corpus and computational methods to loanword research: new approaches to Anglicisms in Spanish, Ph.D. thesis, The University of Texas at Austin, 2017. URL: http://hdl.handle.net/2152/63064.

[11] E. Álvarez Mellado, Annotation guidelines for lexical borrowings, https://adobo-task.github.io/docs/guidelines.pdf, 2021. ADoBo shared task at IberLEF 2021.

[12] S. Jiang, T. Cui, Y. Fu, N. Lin, J. Xiang, BERT4EVER at ADoBo 2021: Detection of Borrowings

in the Spanish Language Using Pseudo-label Technology, in: A. Montoyo, S. C. E. de la ingeniería, S. E. p. e. P. del Lenguaje Natural (Eds.), Proceedings of the IberLEF 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 278–283. URL: http://ceur-ws.org/Vol-2943/adobo_paper1.pdf.

[13] J. de la Rosa, ADoBo 2021: The futility of STILTs for the classification of lexical borrowings in Spanish, in: A. Montoyo, S. C. E. de la ingeniería, S. E. p. e. P. del Lenguaje Natural (Eds.), Proceedings of the IberLEF 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 284–292. URL: http://ceur-ws.org/Vol-2943/adobo_paper2.pdf.

[14] E. Álvarez-Mellado, C. Lignos, Detecting unassimilated borrowings in Spanish: An annotated corpus and approaches to modeling, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3868–3888. URL: https://aclanthology.org/2022.acl-long.268/. doi:10.18653/v1/2022.acl-long.268.

[15] P. Törnberg, Large language models outperform expert coders and supervised classifiers at annotating political social media messages, Social Science Computer Review 0 (0) 08944393241286471. doi:10.1177/08944393241286471.

[16] M. Heseltine, B. C. von Hohenberg, Large language models as a substitute for human experts in annotating political text, Research & Politics 11 (2024) 20531680241236239. doi:10.1177/20531680241236239.

[17] B. E. Perron, H. Luan, B. G. Victor, O. Hiltz-Perron, J. Ryan, Moving beyond chatgpt: Local large language models (llms) and the secure analysis of confidential unstructured text data in social work research, Research on Social Work Practice (2024). URL: https://doi.org/10.1177/10497315241280686. doi:10.1177/10497315241280686.

[18] S. Schmidgall, Y. Su, Z. Wang, X. Sun, J. Wu, X. Yu, J. Liu, Z. Liu, E. Barsoum, Agent laboratory: Using llm agents as research assistants, 2025. URL: https://arxiv.org/abs/2501.04227. arXiv:2501.04227.

[19] S. Agarwal, G. Sahu, A. Puri, I. H. Laradji, K. D. Dvijotham, J. Stanley, L. Charlin, C. Pal, Litllms, llms for literature review: Are we there yet?, 2025. URL: https://arxiv.org/abs/2412.15249. arXiv:2412.15249.

[20] L. P. Argyle, E. C. Busby, J. R. Gubler, B. Hepner, A. Lyman, D. Wingate, Arti-”fickle” intelligence: Using llms as a tool for inference in the political and social sciences, 2025. URL: https://arxiv.org/abs/2504.03822. arXiv:2504.03822.

[21] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, 2022. URL: https://arxiv.org/abs/2203.02155. arXiv:2203.02155.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2020. URL: https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf.

[23] OpenAI, Gpt-4.1, https://openai.com/index/gpt-4-1/, 2025. Large language model with 1M token context window, released April 14, 2025.

[24] OpenAI, Introducing openai o3 and o4-mini, https://openai.com/index/introducing-o3-and-o4-mini/, 2025. Accessed: 2025-05-14.

[25] P. AI, R1 1776: An uncensored version of deepseek-r1, https://huggingface.co/perplexity-ai/r1-1776, 2025. Open-source language model post-trained to remove censorship constraints. Accessed June 2025.

[26] DeepSeek-AI, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, arXiv:2501.12948 (2025). URL: https://arxiv.org/abs/2501.12948.

[27] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.

[28] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, X. Sun, L. Li, Z. Sui, A survey on in-context learning, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 1107–1128. URL: https://aclanthology.org/2024.emnlp-main.64/. doi:10.18653/v1/2024.emnlp-main.64.

[29] Y. Zhang, Y. Yuan, A. C.-C. Yao, Meta prompting for ai systems, 2025. URL: https://arxiv.org/abs/2311.11482. arXiv:2311.11482.

[30] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, P. Clark, Self-refine: Iterative refinement with self-feedback, 2023. URL: https://arxiv.org/abs/2303.17651. arXiv:2303.17651.

[31] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, 2020. URL: https://arxiv.org/abs/2001.08361. arXiv:2001.08361.

[32] M. Besta, J. Barth, E. Schreiber, A. Kubicek, A. Catarino, R. Gerstenberger, P. Nyczyk, P. Iff, Y. Li, S. Houliston, T. Sternal, M. Copik, G. Kwaśniewski, J. Müller, Łukasz Flis, H. Eberhard, H. Niewiadomski, T. Hoefler, Reasoning language models: A blueprint, 2025. URL: https://arxiv.org/abs/2501.11223. arXiv:2501.11223.

A. Full Text of Prompts

A.1. System Prompt

```
You are an expert evaluator.
```

A.2. Rules for Anglicism Identification

```
The following rules provide a robust, step-by-step protocol for identifying
emergent, unassimilated English lexical borrowings (anglicisms) in Spanish
text.

1. Initial Identification
- Determine if the word or multiword expression is of English origin or
mimics English word formation.
  - If the word is not of English origin, do not consider it an anglicism.
  - If it mimics English formation but does not exist in English
(pseudoanglicism, e.g., balconing), consider it an anglicism.

2. Exclusion of Proper Names and Code-Mixed Inclusions
- If the word is a proper noun (person, organization, product, location,
event, etc.) or a direct code-mixed quotation, do not consider it an
anglicism.
- Borrowings embedded within proper nouns or named entities do not count,
unless the proper noun is part of a multiword borrowing used grammatically
as in English (e.g., Google cooking is annotated).

3. Graphophonological Compliance
- If the word's spelling and pronunciation conform to Spanish
graphophonological rules (e.g., bar, club), proceed to dictionary checks.
- If not (e.g., show, look), generally consider it an anglicism unless it
is a long-registered realia word (see Rule 5).

4. Adaptation and Assimilation Status
- If the word has been morphologically or orthographically adapted to
Spanish (e.g., fútbol, tuit, líder), do not consider it an anglicism.
- If the word remains unadapted, continue to dictionary checks.

5. Dictionary Verification
- If the unadapted word is registered in the Diccionario de la Lengua
Española (DLE):
  - If it appears in italics, consider it an anglicism.
  - If it appears without italics and with the relevant meaning, do not
consider it an anglicism (it is considered assimilated).
```

```
  - If it is not registered or not with the relevant meaning, consider it
an anglicism.
- For multi-sense words (e.g., top), only consider them an anglicism when
used with unregistered meanings.

6. Realia and Long-Registered Borrowings
- If the word is a long-registered realia borrowing (cultural terms like
jazz, pizza, whisky, club), do not consider it an anglicism, even if
unadapted.
- If the word is a recent or emergent realia borrowing not yet registered,
consider it an anglicism.

7. Multiword Borrowings
- Do consider multiword expressions borrowed as a unit from English (e.g.,
reality show, best seller).
- For adjacent borrowings not forming a fixed English phrase (e.g., look
sporty), select each word separately.

8. Exclusions and Special Cases
- Do not consider an anglicism:
  - Latinisms, scientific units, species names, acronyms (unless part of a
multiword borrowing), or digits in isolation.
  - Metalinguistic usages, literal quotations, or code-switched expressions
not integrated into the sentence.
  - Names of peoples or languages, and words derived transparently from
proper nouns (e.g., un iPhone, un whatsapp).
- Do consider an anglicism:
  - Pseudoanglicisms (Spanish-coined words mimicking English, e.g.,
footing, balconing).
  - Unadapted names of fictitious creatures (e.g., hobbit, troll).
  - Borrowings embedded in compounds or prefixed forms, if the borrowed
element retains independence (e.g., ex influencer, nano influencers).
```

Because these rules refer to the DLE and dictionary checks, we add the following line to prevent model confusion:

```
The rules refer to dictionary checks. You don't have access to a
dictionary, so do the best you can.
```

A.3. Additional Reminders From Self Refinement

```
 Examples of words that are NOT anglicisms include:
```

- Words fully adapted to Spanish orthography (e.g., 'fútbol', 'líder', 'tuit')
- Long-established borrowings (e.g., 'bar', 'club', 'jazz', 'whisky')
- Proper names of people, places, or companies (e.g., 'Twitter' as a company name)
- Scientific terminology with Latin or Greek roots

Special rules for digital media and technology:
1. Names of online platforms (e.g., 'Facebook', 'Twitter') are NOT anglicisms when used as proper nouns to refer to the specific company or service.
2. However, these names ARE anglicisms when used generically (e.g., "Hizo un twitter" to mean "He made a tweet").
3. Generic terms related to digital culture ARE anglicisms when they maintain English form (e.g., 'post', 'blog', 'meme', 'podcast', 'streaming').
4. Terms for social media actions in their English form ARE anglicisms (e.g., 'like', 'share', 'tweet').

For multi-word expressions:
1. Analyze whether the entire expression functions as a unified borrowing from English.
2. Examples of multi-word anglicisms: 'fast food', 'big data', 'home office', 'fake news'.
3. Do NOT count individual English words that appear adjacent to each other but do not form a standard expression in English.
4. When in doubt about whether multiple words form a single anglicism, consider them separately.

For business, finance, and technical terminology:
1. Recent business terms that maintain English spelling ARE anglicisms (e.g., 'CEO', 'manager', 'marketing', 'branding', 'broker').
2. Technical computing terms that maintain English spelling ARE anglicisms (e.g., 'hardware', 'software', 'online', 'web').
3. Industry-specific English jargon ARE anglicisms (e.g., 'blockchain', 'big data', 'know-how').
4. International brand names and trademarks are NOT anglicisms unless used generically.

Follow this step-by-step process for identifying anglicisms:
1. First, identify all words and phrases that appear to have English origin.
2. For each candidate term, apply the exclusion criteria (proper names, adapted words, etc.).

```
3. For remaining terms, verify they maintain English spelling/pronunciation
patterns.
4. Group related words that form a single expression in English.
5. Verify each term against the specific rules for digital media, business
terms, etc.
```

B. Model Details

All models tested are April checkpoints of OpenAI models.

- `gpt-4.1-nano-2025-04-14`
- `gpt-4.1-mini-2025-04-14`
- `gpt-4.1-2025-04-14`
- `o4-mini-2025-04-16`
- `o3-2025-04-16`

C. Complete Score Reports

Scores reported in this section are rounded to the nearest whole number.

Table 3
GPT 4.1 Nano Results on **LBAD**-valid

	Precision	Recall	F1
Baseline	19	85	31
Baseline + Guidelines	14	88	24
Baseline + Guidelines + Reminders	12	88	21
Baseline + CoT	32	97	48
Baseline + CoT + Guidelines	36	94	52
Baseline + CoT + Guidelines + Reminders	37	96	53
Baseline + 5 ICL Examples	9	97	17
Baseline + Guidelines + 5 ICL Examples	7	91	14
Baseline + 10 ICL Examples	14	86	24
Baseline + Guidelines + 10 ICL Examples	11	89	20
Baseline + 25 ICL Examples	6	91	12
Baseline + Guidelines + 25 ICL Examples	6	94	12

Table 4
GPT 4.1 Mini Results on **LBAD**-valid

	Precision	Recall	F1
Baseline	24	93	38
Baseline + Guidelines	33	95	49
Baseline + Guidelines + Reminders	49	91	64
Baseline + CoT	35	98	52
Baseline + CoT + Guidelines	62	93	74
Baseline + CoT + Guidelines + Reminders	70	98	82
Baseline + 5 ICL Examples	26	97	41
Baseline + Guidelines + 5 ICL Examples	31	98	47
Baseline + 10 ICL Examples	30	94	45
Baseline + Guidelines + 10 ICL Examples	33	95	49
Baseline + 25 ICL Examples	30	95	46
Baseline + Guidelines + 25 ICL Examples	39	94	55

Table 5
GPT 4.1 Results on **LBAD**-valid

	Precision	Recall	F1
Baseline	37	85	51
Baseline + Guidelines	57	93	70
Baseline + Guidelines + Reminders	51	87	65
Baseline + CoT	39	100	57
Baseline + CoT + Guidelines	68	99	80
Baseline + CoT + Guidelines + Reminders	67	99	80
Baseline + 5 ICL Examples	38	100	56
Baseline + Guidelines + 5 ICL Examples	46	98	62
Baseline + 10 ICL Examples	42	99	60
Baseline + Guidelines + 10 ICL Examples	55	98	70
Baseline + 25 ICL Examples	39	99	55
Baseline + Guidelines + 25 ICL Examples	47	100	64

Table 6
o4-mini Results on **LBAD**-valid

	Precision	Recall	F1
Baseline	31	99	47
Baseline + Guidelines	70	97	81
Baseline + Guidelines + Reminders	71	98	82

Table 7
o3 Results on **LBAD**-valid

	Precision	Recall	F1
Baseline	29	100	45
Baseline + Guidelines	82	99	90
Baseline + Guidelines + Reminders	81	98	89