HiTZ at ADoBo 2025: Few-Shot Anglicism Detection in Spanish

Maite Heredia, Jeremy Barnes and Aitor Soroa

HiTZ Center - Ixa, University of the Basque Country UPV/EHU

Abstract

We present our submission to the ADoBo 2025 Shared Task, part of the IberLEF shared evaluation campaign. The task focuses on detecting anglicisms in Spanish newswire texts. Our approach leverages the instruction-tuned language model LLAMA 3.3 70B to identify spans containing anglicisms. To address certain shortcomings observed in the model's behavior, we experiment with zero- and few-shot strategies and explore the integration of additional model-based modules. However, the best performing system on the test set is a 5-shot model without auxiliary modules. We conclude with an analysis of the strengths and limitations of using large language models for anglicism detection.

Keywords

linguistic borrowing, loanwords, anglicisms, loanword detection, LLM,

1. Introduction

As languages come in contact, it is frequent for speakers to import words from one language to another. This phenomenon, known as *loanwords* or *borrowings*, can be caused by a number of reasons; for example, words are borrowed to fill semantical gaps in a language, as neologisms or as a stylistic feature [1]. Among these borrowings, the widespread use of English as *lingua franca* makes anglicisms - borrowings that come from English - especially salient in many languages [2]. The study of borrowings in general and anglicisms in particular can be aided by computational techniques and advances in NLP. In this context, the 2025 edition of the ADoBo Shared Task [3], proposed as part of IberLEF [4], tackles the task of detecting unassimilated anglicisms in Spanish newswire.

Recently, there has been widespread interest in the emergent capabilities of Large Language Models (LLMs), their ability to perform tasks without task-specific fine-tuning. In this work, we investigate whether anglicism detection can be considered one of these emergent abilities and explore strategies to mitigate the errors made by LLMs in this task. We prompt the LLAMA 3.3 70B [5] model in both zero-shot and few-shot settings, providing it with a summary of the annotation guidelines used to construct the dataset. To better control the model's behaviour and output, we avoid free-form generation and instead opt for a *guided decoding* approach [6], driving the model to produce structured responses in JSON format.

We perform an error analysis on the development test set and identify that the main issue is a very low precision. This is due to the model's tendency to overgenerate even in sentences that do not present any anglicisms, a known limitation of LLMs, which lack abstention abilities [7]. To address this, we propose a detection module based on an encoder-only model trained on a binary classification task to determine whether a sentence contains anglicisms. While this approach significantly improves performance on the development set, the different distributions between the development and test sets result in a substantial difference in performance. We also experiment with a similar strategy that involves using a Named Entity Recognition (NER) model [8] or prompting the generative model to identify named entities (NEs), as the model often incorrectly labels NEs as anglicisms—even when explicitly prompted not to. Ultimately, our best submission on the test set is the few-shot model without any additional modules.

IberLEF 2025, September 2025, Zaragoza, Spain
maite.heredia@ehu.eus (M. Heredia); jeremy.barnes@ehu.eus (J. Barnes); asoroa@ehu.eus (A. Soroa)
0009-0005-6719-5433 (M. Heredia); 0000-0002-8043-8058 (J. Barnes); 0000-0001-8573-2654 (A. Soroa)

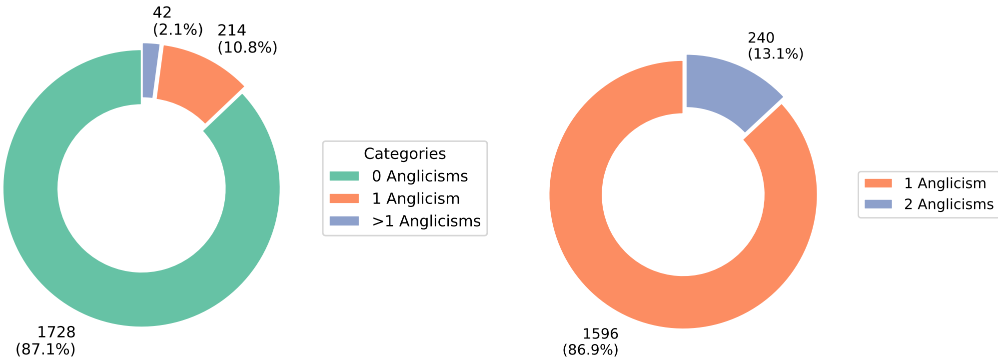

Figure 1: Proportion of anglicisms per sentence and in the development and test sets.

2. Related Work

The task of borrowing detection has been previously studied in several languages, e.g., German [9] and Norwegian [10]. Regarding borrowing detection in Spanish, the previous edition of this shared task, held in 2021, marked an important step toward borrowing detection in Spanish newswire texts. That edition focused particularly on borrowing detection with an emphasis on anglicisms [11]. In addition, there is other research that deals with the study of borrowings and anglicisms in Spanish from different perspectives [12].

Beyond monolingual studies, there has also been significant work on multilingual borrowing and anglicism detection. Recent approaches include those by Nath et al. [13] and Miller and List [14], which address the problem from a cross-linguistic perspective, highlighting the importance of generalized models for lexical borrowing detection.

Loanword detection is also relevant to sociolinguistic research. It provides valuable insights into language contact phenomena, lexical change, and linguistic influence [15].

3. Task Description

The proposed task focuses on the detection of unassimilated anglicisms in Spanish newswire texts. In the annotation guidelines, linguistic borrowings are defined as *"the incorporation of single lexical units from one language (the donor language) into another language (the recipient language) usually accompanied by morphological and phonological modification to conform with the patterns of the recipient language"*. In particular, for this edition of the shared task, the focus is on *unassimilated anglicisms* in Spanish, i.e., words from English origin that have not been assimilated ortographically nor morphologically into Spanish.

The task is framed as sequence labelling, and the output must contain the span(s) of text detected for each instance, instead of a classic BIO approach where there must be one tag per token, although both formats are easily compatible. There is no training set provided as part of the shared task, only a development and a test set, which was blind for the duration of the shared task. Nonetheless, participants are allowed and encouraged to use any previously available datasets, such as the COALAS dataset [16] which was provided as part of the previous edition of the shared task.

3.1. Development and Test Data

The development and test sets that have been provided as part of the shared task contain 1984 and 1836 total instances, respectively. Each instance is composed of a sentence and up to 5 annotated anglicisms, in a CSV format.

Figure 1 shows the distribution of anglicisms per sentence in both splits. As shown in the figure, the distribution of anglicisms differs significantly between the two sets. In the development set, the majority of instances (87.1%) do not contain any anglicisms. Among the remaining instances, most of them contain only 1 anglicism (10.8%), and only a small proportion (2.1%) have 2, 3 or 5 anglicisms. In contrast, all instances of the test set contain at least one anglicism, and 13.1% instances contain two. The development set includes a total of 304 anglicisms, comprising 269 unique forms. The test set includes 2,076 anglicisms across 373 unique ones. Only eight anglicisms are shared between the development and test sets.

Evaluation is computed using standard metrics: precision, recall and F1-score as the harmonic mean of both. For scoring purposes, differences in casing and the presence or absence of quotation marks are ignored.

4. System Description

In this section we present the different models and configurations that we fine-tune or prompt for the task of anglicism detection. When our experiments involve fine-tuning, we use the COALAS dataset [16] train split as training data. We test all models and settings with the development set provided for the shared task.

4.1. Encoder-Only Models

Given that the task can be formulated as a sequence labeling problem, it is appropriate to test the capabilities of encoder-only architectures as baselines. For this purpose, we have fine-tuned 5 encoder-only models: ModernBERT [17] and BETO [18], which are monolingual models in English and Spanish respectively; IXAmBERT [19], a multilingual model that focuses on Spanish, English and Basque; and XLM-RoBERTa large [20] and mdeBERTa v3 [21], massively multilingual state-of-the-art encoders. All models have been fine-tuned for five epochs with a default learning rate of $\eta = 2e{-}5$, and a batch size of 32.

These models are not part of our final submission, because they were not the focus of our experimentation. We have not performed an exhaustive hyperparameter tuning, nor are their results on the development set on par with those of the decoder-only models. Nonetheless, it can still be insightful to observe the performance of smaller models that are much faster to deploy and require less computational resources.

4.2. Decoder-Only Models

We leverage the model Llama 3.3 70B [5] using constrained decoding [6] for prompting, implemented using the vLLM library for LLM inference [22]. We constrain the output to follow a JSON structure, where the model can only fill the fields "text", "start" and "end". For instance, given the input sentence: *Receta para preparar una carrot cake vegan friendly*, the output should look like this:

```
{"anglicisms": [{"text": "carrot cake", "start": 26, "end": 36}, {"text": "vegan
    friendly", "start": 38, "end": 51}]}
```

```
{"NER": [{"text": "Google", "start": 26, "end": 36}, {"text": "Microsoft", "start": 38,
    "end": 51}]}
```

As per parameter- and prompt-tuning, the following variables have been tested against each other in different settings:

- Prompt: We have tried different levels of informativeness for the prompt, from a naive simple approach where the model is only asked to retrieve unassimilated anglicisms to the final prompt shown in Figure 2, which features a summary of the guidelines used to annotate the corpus and thus allows for a more accurate and task-appropriate detection.

Actúa como un experto lingüista especializado en detección de préstamos lingüísticos. Tu tarea es analizar un fragmento de texto en español y etiquetar todos los anglicismos no asimilados, según siguientes reglas: Solo marcas préstamos recientes del inglés que no hayan sido adaptados ortográfica ni morfológicamente al español (por ejemplo: smartphone, influencer, look, reality show, hype). Ignora los préstamos ya adaptados como tuit, líder, fútbol, espoiler, incluso si provienen del inglés. Excluye nombres propios, marcas, lugares, instituciones, fechas, eventos, hashtags, acrónimos y citas literales. Incluye expresiones multi-palabra si son préstamos completos como reality show, total look o tech bro. Si la palabra aparece en el Diccionario de la lengua española (DLE) sin comillas ni cursiva y con ese significado, no debe etiquetarse. No etiquetes calcos, traducciones literales ni palabras derivadas de raíces inglesas pero que siguen patrones del español como hacktivista, randomizar o shakespeariano. Etiqueta pseudoanglicismos como footing, balconing.

(a) Prompt in Spanish

Act as an expert linguist specializing in loanword detection. Your task is to analyze a fragment of text in Spanish and tag all unassimilated anglicisms, according to the following rules: Only tag recent English loanwords that have not been orthographically or morphologically adapted to Spanish (for example: smartphone, influencer, look, reality show, hype). Ignore already adapted loanwords such as tuit, líder, fútbol, espoiler, even if they come from English. Exclude proper nouns, brands, places, institutions, dates, events, hashtags, acronyms, and literal quotes. Include multi-word expressions if they are complete loanwords such as reality show, total look, or tech bro. If the word appears in the Diccionario de la Lengua Española (DLE) without quotation marks or italics and with that meaning, it should not be tagged. Don't tag calques, literal translations, or words derived from English roots that follow Spanish patterns, such as hacktivist, randomize, or Shakespearean. Tag pseudo-Anglicisms like footing and balconing.

(b) Prompt in English

Figure 2: Prompts used for anglicism detection.

- Examples: A zero-shot and 5-shot approach have been tested. The 5 examples have been manually selected to be representative of some common errors of the model on the zero-shot setting (detecting named entities, slogans or acronyms). Although they do not avoid these errors completely, the 5-shot strategy obtains better results.
- Language: We have tried to prompt the model in both English and Spanish, with the latter obtaining better results.
- Temperature: We have run the inference with temperature values of 0, 0.5 and 1. The value that yields the best results is 0.5.

4.2.1. Detection Module

The initial results show a reasonably high recall but a very low precision, indicating that the model is generating a large number of false positives. A manual inspection of its outputs confirms this trend: the model frequently overgenerates, attempting to identify at least one anglicism per sentence. This behavior appears to stem from the model's limited abstention capabilities [7], which prevents it from refraining from making a prediction when uncertain. As a result, the model's performance is significantly impacted, especially given the distribution of the development set described in Section 3.1. In many cases, it even misclassifies clearly Spanish words in sentences that are entirely in Spanish. Although we experimented with prompt-based strategies to mitigate this behaviour, they led to only marginal improvements.

To address this issue, we introduce a previous module to the inference step that performs a preliminary binary classification to determine whether a sentence contains any anglicisms. For this task, we fine-tuned several discriminative models on the COALAS training set, adapting the original labels to a binary format (0 for no anglicisms, 1 for presence of anglicisms). Among the models evaluated, which are the same in Section 4.1) mDeBERTa achieved the best performance, with an F1-score of 0.99 on the development set.

Model	Precision	Recall	F1-score
ModernBERT	33.83	53.56	41.47
Beto	08.96	16.27	11.55
IXAmBERT	65.94	72.20	68.93
XLM-RoBERTa	40.67	49.49	44.65
mDeBERTa	**73.24**	**74.24**	**73.74**

Table 1
Results of discriminative models on the development set

We integrate this binary classifier into our pipeline by first filtering sentences based on its predictions. Only those classified as containing anglicisms are passed to the LLM for fine-grained identification. This two-stage approach significantly improves precision (cf. Section 5.1) and also reduces inference time on the development set.

4.2.2. NER Module

Similarly to the previous strategy, we experimented with a pipeline that begins by detecting Named Entities, as we observed that the model frequently misclassifies them as anglicisms—even when explicitly instructed not to. Initially, we used a NER model [8] to identify and exclude Named Entities from the list of potential anglicisms. However, we ultimately opted to prompt the model to identify Named Entities directly as part of the anglicism detection task, as this approach yielded more accurate results.

5. Final Results

In this section, we report the results for the development and test sets provided for the shared task. We evaluate the encoder-only models and 4 different settings of the decoder-only models using the development set. Based on the results of the experiments, we submit 3 runs in total: (1) a few-shot decoder-only model, (2) a few-shot model with a detection module and (3) a few-shot model prompted to also detect named entities, and report the results of these 3 runs on the test set. Likely due to a difference in distribution between both splits, the performance of the models and the model ranking change drastically from one set to the other. For this reason, we first report the development set results, as they have guided some decisions taken for the experiments, and the results of the final submission on the test set.

5.1. Development Set

The results of the discriminative models on the development set can be seen in Table 1. The top-3 models are the multilingual models, and both monolingual models perform notably worse, suggesting that having knowledge of both Spanish and English is essential to be able to detect anglicisms in Spanish. The model that performs best for all metrics is mDeBERTa, even if it is not as large in size as XLM-RoBERTa, suggesting the importance of the pre-training architecture of the models for downstream-task performance.

The results of the different experiments performed with the Llama 3.3 70B model on the development set are presented in Table 2. These results highlight the importance of the detection module when there is a high proportion of sentences that do not contain any anglicisms. This module avoids over-detection, which is reflected in the precision. Enriching the prompt with 5-shot and NER both improve the results of the models, suggesting that prompt-tuning has a notable impact in the performance of the model.

The results of the best encoder-only model are not on par of those of the best decoder-only based pipeline, but they are more balanced than those of a 5-shot model with no additional modules, as well as faster to deploy.

Model	Precision	Recall	F1-score
Zero Shot	08.52	91.19	15.59
5-shot	11.25	**93.22**	20.08
5-shot + Detection	**78.79**	88.14	**83.20**
5-shot + NER	17.85	76.61	28.96

Table 2
Results of different settings of Llama 3.3 70B on the development set.

Strategy	Precision	Recall	F1-score
5-shot	92.60	**94.08**	**93.33**
5-shot + Detection	92.00	50.43	65.15
5-shot + NER	**93.97**	68.26	79.07

Table 3
Results of the final submissions on the test set. Best results and model with the highest F1 in bold.

5.2. Test Set

We have submitted a total of three systems for the task, whose results on the test set can be seen in Table 3. The best performing system is the 5-shot prompted model, without any of the modules. In both cases, adding the modules greatly decreases the recall.

It is clear that there is a large difference in performance between the development and test sets, which we hypothesize is due to the different distribution of both sets, which is likely why the recall is much lower with modules aimed at improving precision in the development set.

5.3. Error Analysis

The best-performing configuration on the test set—a 5-shot model without additional modules—achieves an F1-score of 93.33, with precision and recall at comparable levels. This indicates a balanced rate of false positives (non-anglicisms incorrectly identified as anglicisms) and false negatives (anglicisms that go undetected).

The test set includes multiple instances of the same anglicisms presented with variations in casing and quotation marks, likely to assess whether models rely on these formatting cues for detection. A manual analysis of the system's errors reveals that its misclassifications are consistent across different formats, suggesting that it does not rely on superficial format-based heuristics. Instead, the errors appear to stem from a conceptual misinterpretation of what constitutes a borrowing. Common mistakes include mislabeling named entities resembling English expressions (e.g., *Big Little Lies* or *Prision Break*) and incorrectly handling composition of anglicism phrases such as *look* and *total black*, that are treated as multiple anglicisms due to their syntactic integration into Spanish. These are often identified as a single span by the model but are annotated as separate spans in the gold standard, negatively impacting both precision and recall.

6. Conclusion & Discussion

In this paper, we report our experiments and submissions for the second edition of the ADoBo Shared Task, as part of the IberLEF 2025 evaluation campaign. The task at hand consists of unassimilated anglicism detection in Spanish newswire texts. We have based our contributions on the exploit of LLMs' capabilities and implicit knowledge aided with smaller models to make its results more robust. Although our best performing approach has consisted on 5-shot prompting, where the only tuning has been performed on the prompt for it to be as informative and rigorous as possible, it is still likely that the other approaches and findings that we have presented, namely, the use of smaller encoder-only models as a pre-classification step, can be useful for other corpora with different distributions, as proven with the evaluation performed on the development set. What is more, a few-shot prompted model has the advantage of avoiding overfitting on a training set or learning artifacts for classification, such as casing or quotation marks, as we show in the error analysis, which is sure to improve the results in unseen distributions.

Acknowledgements

This work is supported by the European Union under Horizon Europe (Project LUMINOUS, grant number 101135724) and by the Basque Government (IXA excellence research group IT1570-22). Maite Heredia is supported by the UPV/EHU PIF23/218 predoctoral grant.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

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

[2] C. Furiassi, V. Pulcini, F. R. González (Eds.), The Anglicization of European Lexis, John Benjamins Publishing Company, Netherlands, 2012.

[3] 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).

[4] 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.

[5] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[6] B. T. Willard, R. Louf, Efficient guided generation for large language models, 2023. URL: https://arxiv.org/abs/2307.09702. arXiv:2307.09702.

[7] N. Madhusudhan, S. T. Madhusudhan, V. Yadav, M. Hashemi, Do LLMs know when to NOT answer? investigating abstention abilities of large language models, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 9329–9345. URL: https://aclanthology.org/2025.coling-main.627/.

[8] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, GLiNER: Generalist model for named entity recognition using bidirectional transformer, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 5364–5376. URL: https://aclanthology.org/2024.naacl-long.300/. doi:10.18653/v1/2024.naacl-long.300.

[9] S. Leidig, T. Schlippe, T. Schultz, Automatic detection of anglicisms for the pronunciation dictionary generation: A case study on our german it corpus, 2014.

[10] G. Andersen, Semi-automatic approaches to anglicism detection in norwegian corpus data, in: C. Furiassi, V. Pulcini, F. R. González (Eds.), The Anglicization of European Lexis, John Benjamins, 2012, pp. 111–130.

[11] E. Álvarez Mellado, Extracting English lexical borrowings from Spanish newswire, in: A. Ettinger, E. Pavlick, B. Prickett (Eds.), Proceedings of the Society for Computation in Linguistics 2021, Association for Computational Linguistics, Online, 2021, pp. 384–386. URL: https://aclanthology.org/2021.scil-1.40/.

[12] E. Alvarez-Mellado, C. Lignos, Borrowing or codeswitching? annotating for finer-grained distinctions in language mixing, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3195–3201. URL: https://aclanthology.org/2022.lrec-1.342/.

[13] A. Nath, S. Mahdipour Saravani, I. Khebour, S. Mannan, Z. Li, N. Krishnaswamy, A generalized method for automated multilingual loanword detection, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 4996–5013. URL: https://aclanthology.org/2022.coling-1.442/.

[14] J. E. Miller, J.-M. List, Detecting lexical borrowings from dominant languages in multilingual wordlists, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 2599–2605. URL: https://aclanthology.org/2023.eacl-main.190/. doi:10.18653/v1/2023.eacl-main.190.

[15] I. Stewart, D. Yang, J. Eisenstein, Tuiteamos o pongamos un tuit? investigating the social constraints of loanword integration in Spanish social media, in: A. Ettinger, E. Pavlick, B. Prickett (Eds.), Proceedings of the Society for Computation in Linguistics 2021, Association for Computational Linguistics, Online, 2021, pp. 286–297. URL: https://aclanthology.org/2021.scil-1.26/.

[16] 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.

[17] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL: https://arxiv.org/abs/2412.13663. arXiv:2412.13663.

[18] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[19] A. Otegi, A. Agirre, J. A. Campos, A. Soroa, E. Agirre, Conversational question answering in low resource scenarios: A dataset and case study for basque, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 436–442.

[20] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[21] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.

[22] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with pagedattention, in: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023.