WiC Evaluation in Galician and Spanish: Effects of Dataset Quality and Composition

Marta Vázquez Abuín, Marcos Garcia
CiTIUS – Research Center in Intelligent Technologies
Universidade de Santiago de Compostela
{martavazquez.abuin,marcos.garcia.gonzalez}@usc.gal

Abstract

This work explores the impact of dataset quality and composition on Word-in-Context performance for Galician and Spanish. We assess existing datasets, validate their test sets, and create new manually constructed evaluation data. Across five experiments with controlled variations in training and test data, we find that while the validation of test data tends to yield better model performance, evaluations on manually created datasets suggest that contextual embeddings are not sufficient on their own to reliably capture word meaning variation. Regarding training data, our results suggest that performance is influenced not only by size and human validation but also by deeper factors related to the semantic properties of the datasets. All new resources will be freely released.

1 Introduction

Lexical ambiguity (e.g., polysemous words conveying different senses depending on the context) is a central feature of natural languages, and its resolution remains a challenge for computational models, as distinguishing between different senses of a word can be difficult even for humans (Bevilacqua et al., 2021). In NLP, one of the most widely used tasks to evaluate model performance in lexical disambiguation is Word-in-Context (WiC) and its extensions to other languages (Pilehvar and Camacho-Collados, 2019; Raganato et al., 2020), where the goal is to determine whether a target word used in two sentences has the same meaning or not.

These datasets are typically built using existing lexical resources, primarily WordNet and Wiktionaries. As a result, their quality depends heavily on the coverage and reliability of these underlying sources, as well as on the specific methodology used to construct the WiC instances. In this context, human performance on WiC datasets varies considerably, e.g., around 75% in Japanese and 76% in Korean, compared to 97% in Farsi, where sentence pairs were manually grouped by an annotator (Raganato et al., 2020).[1]

In the case of languages such as Galician and Spanish, existing datasets have been automatically constructed using lexical resources with limited coverage, often relying on machine translation both in the development of the WordNets and in the generation of sentence examples (Vázquez Abuín and Garcia, 2025). Moreover, these datasets have neither been validated nor evaluated with respect to human performance, which makes it difficult to assess their overall quality and reliability.

This work investigates the impact of dataset composition and quality on model performance in the WiC task, focusing on both training and evaluation data. We begin by assessing the quality of existing test sets for Galician and Spanish, followed by a validation process to remove instances that are ambiguous for human annotators. Additionally, we construct two new evaluation datasets [2] in which all sentences are manually authored and validated by experts—resources that are also useful for analyzing potential data contamination issues (Sainz et al., 2023). Using these resources, we design five experiments that show that while validating evaluation data tends to improve model performance, results on manually created datasets suggest that contextual embeddings alone are not sufficient to reliably capture word meaning variation. However, during training, data size and human validation seem to have a more limited effect, while other factors related to the semantic properties of the data (such as the well-known effect of word not seen during training) may have a more substantial impact.

[1] This variation is also reflected in model performance, where for instance, zero-shot models perform better on Farsi than on Japanese or Korean.

[2] https://github.com/mrtva/wic-eval-starsem25

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 172–178
November 8-9, 2025 ©2025 Association for Computational Linguistics

2 Datasets and human annotation

Original datasets: We start by employing, to the best of our knowledge, the only available WiC datasets for Galician and Spanish (Vázquez Abuín and Garcia, 2025). These datasets were constructed following the same methodology as the original WiC paper (Pilehvar and Camacho-Collados, 2019), using WordNets from the Multilingual Central Repository (MCR) (Gonzalez-Agirre et al., 2012).[3] The Galician dataset comprises 1500 training, 400 development, and 1400 test instances. For Spanish, only two splits are available: 200 instances for training or development, and 800 for testing. In all cases, each instance consists of a pair of sentences containing the same word form, along with a binary label indicating whether the word has the same meaning in both contexts ('true') or not ('false'). It is worth mentioning that some instances of the Galician datasets were translated from English or Spanish using machine translation. While the translations were validated by native speakers, the original paper only mentions a manual review of the test data, but does not report any human evaluation of the final datasets.

Validation of the test sets: To assess the quality of the original automatically created dataset, we randomly selected 150 instances from the test sets, which were independently annotated by three experts, bilingual speakers of both languages. Annotators performed the WiC task by determining whether a target word carried the same meaning in two different contexts, assigning a binary label (0 for 'false', 1 for 'true'). No external resources were made available during the annotation process.

The average agreement between each annotator was 71% and 63% for Galician and Spanish, respectively.[4] Table 1 presents the inter-annotator agreement results, both for individual annotator pairs and across all three experts. The agreement levels range from fair to moderate for Galician, and remain fair for Spanish, highlighting potential limitations in automatically generated datasets.

Annotators	Gal	Spa
Annotator 1 vs. 2	0.444	0.323
Annotator 2 vs. 3	0.419	0.399
Annotator 1 vs. 3	0.443	0.455
Fleiss' κ	0.435	0.389
Krippendorff' α	0.436	0.390

Table 1: Inter-annotator agreement on the original Galician and Spanish test sets. Top rows are Cohen's κ scores between pairs of annotators, while bottom rows show the $kappa$ and α values for the three annotators.

2.1 New human validated test sets

The inter-annotator agreement results indicate that, although automatic dataset creation offers scalability, it may introduce ambiguities or inconsistencies that compromise data quality. To address this, we developed a revised version of the test sets by conducting large-scale human validation and retaining only those instances for which the original and human-assigned labels were in agreement.

To this end, we randomly selected 950 instances for Galician and 650 for Spanish from the original test sets, which were then validated by a bilingual language expert. The observed agreement between the expert and the original labels was 71% for Galician and 68% for Spanish. Following the original WiC setup, we ensured a balanced distribution of 'true' and 'false' instances by selecting 450 and 370 sentence pairs for the new validated Galician and Spanish datasets, respectively.

2.2 New manually created test sets

In addition to the potential ambiguities introduced by WordNet examples, automatically constructed datasets also present a risk of data contamination, as many of the sentences may have been seen by language models during pretraining. To assess the impact of these factors, we created an additional test set for each language, composed entirely of manually written sentences. The process is exemplified in Table 2, while Table 3 provides an overview of the three datasets (original, validated, and manually created) for Galician and Spanish.

We randomly selected 50 instances (100 sentences) from the original test sets for each language (*Test set* row in Table 2). Two language experts were then asked to carefully read each sentence, consider the meaning of the target word in its specific context, and compose a new sentence (between 5 and 15 words) in which the target word conveys the same meaning (*Expert* rows). Then, we combined the new sentences produced by each linguist to construct new instances labeled

[3] https://adimen.ehu.eus/web/MCR

[4] Furthermore, we computed a majority vote accuracy, where each instance was assigned by the majority of the three annotators, with an accuracy of 71.33% for Galician and 62% for Spanish.

Source	Sentences		Label
Test set	Sentence 1	Sentence 2	False
Expert 1	New_A1_S1	New_A1_S2	—
Expert 2	New_A2_S1	New_A2_S2	—
Comb.	New_A1_S1	New_A2_S1	True
Comb.	New_A1_S2	New_A2_S2	True
Comb.	New_A2_S1	New_A1_S2	False
Comb.	New_A2_S2	New_A1_S1	False

Table 2: Example of the process to create the new dataset. We use sentences from the test set (Sentences 1 and 2) to manually create new contexts with the target word conveying the same meaning (New_A1_S1, etc.).

Dataset	Train	TWs	Test	TWs
Original	1500	1187	1400	905
Validated	—	—	450	374
Manual	—	—	172	50
Original	200	190	800	641
Validated	—	—	370	322
Manual	—	—	174	50

Table 3: Summary of the datasets employed in our experiments for Galician (top) and Spanish (bottom), including where each dataset came from (original, validated subset, or manually created), the number of instances per split (train/test) and the number of unique Target Words (TWs). With the exception of the original datasets, both the validated and manually created datasets were reviewed by humans.

as 'true'—when both sentences were derived from the same original context—and 'false'—by pairing sentences that originated from different contexts in which the target word had distinct meanings—as can be seen in the *Comb.* rows of Table 2. As before, we maintained a balance between true and false instances by selecting 200 instances per language—100 for each class. Finally, two language experts validated half of the newly created instances, yielding an observed agreement of 86% for Galician and 87% for Spanish. We retained only those instances where there was agreement, resulting in final manually created test sets consisting of 172 sentence pairs for Galician and 174 for Spanish.[5]

[5]Although other experimental settings with less strict criteria could be explored, in this work we focus on minimizing ambiguities by selecting only agreed cases.

3 Experiments

We conducted five experiments to evaluate different combinations of training and test data, including original, human-validated, and fully human-created datasets for Galician and Spanish. These experiments were designed to analyze the impact of composition and quality of the datasets on model performance. The first three experiments use the same original training data while varying the test sets. **Exp1** evaluates performance on the original automatically constructed test set. **Exp2** replaces it with the manually validated version to assess the effect of validation. **Exp3** tests the models on the manually created datasets. Together with the third one, the final two experiments explore the influence of training data on the manually created test sets. In **Exp4**, we augment the original training set with validated data to evaluate the benefit of incorporating human-verified examples. **Exp5** trains models exclusively on the validated datasets to investigate whether smaller, high-quality training sets can outperform larger automatically created ones.

Models: We evaluated base-size encoder models for both Galician—Bertinho (Vilares et al., 2021) and BERT (Garcia, 2021)— and Spanish—Bertin-RoBERTa (De la Rosa et al., 2022) and RoBERTa-BNE (Fandiño et al., 2022).[6] We compared these encoder models to the multilingual XLM-RoBERTa-base (Conneau et al., 2020) in both languages, and included LLaMA 3.2 3B (Grattafiori et al., 2024) as an example of a state-of-the-art multilingual decoder model.

Method: We follow the standard WiC approach proposed by Wang et al. (2019), which involves training a logistic regression classifier on the concatenation of the contextualized representations of the target word in both sentences. For each model, we train a separate classifier for each layer and report the best performance across layers.[7] Word representations are extracted using transformer-based models via the minicons library (Misra, 2022), which is built on HuggingFace's Transformers.[8]

[6]These models were selected because they demonstrate state-of-the-art performance as encoder models for Spanish and Galician across most evaluations.

[7]For each experiment, we evaluated both the original and z-score normalized embeddings, using the standardization method from Timkey and van Schijndel (2021), and report the best results.

[8]https://github.com/huggingface/transformers

Baseline: As baselines, we implemented two cosine similarity-based methods: one using the contextualized representations from the transformer models, and another using sentence-level embeddings obtained by averaging FastText embeddings. For both methods, we varied the classification threshold in increments of 0.02, labeling a pair as 'true' if the similarity exceeded the threshold.

4 Results and discussion

The best results for each model can be found at Tables 4 and 5, including the highest results for the baselines.[9]

Focusing on the impact of the different test sets, we observe a slight performance improvement when using the human-validated test set in most of the models (Exp1 vs. Exp2), suggesting that human review contributes to the reduction of potential ambiguities and errors of the automatically generated test sets. However, when evaluating on the manually created and validated test sets, we observe a notable performance drop in both languages, particularly for the logistic regression classifiers, and to a lesser extent for the baselines (Exp2 vs. Exp3). This suggests that, despite being less ambiguous, the manually constructed datasets pose greater challenges for models relying on contextualized representations, as simpler cosine similarity-based methods outperform the classifiers.[10] Notably, the baseline results remain largely consistent across different training configurations (Exp3–Exp5 for Galician and Exp4–Exp5 for Spanish), as they share the same test set and the baselines are only minimally influenced by the training data. Contributing factors may include the fact that these sentences were not seen during pretraining, unlike those derived from WordNet and other public resources, which were likely included in the models' pre-training data, and that they may also differ in nature from examples originating in such lexical resources.

Concerning the impact of incorporating human-validated data into the training (Exp3 vs Exp4), no remarkable changes in the overall performance were observed. However, a deeper analysis reveals that increasing the amount of training data improves generalization, as evidenced by higher accuracy on words not seen during training rises (yielding average gains of 2% for Galician and 3% for Spanish in monolingual models) while only causing minor decreases in performance on seen words.

Finally, although the results from training models exclusively on a small validated dataset (which, in the case of Spanish, is larger than the initial one) are not conclusive, it is noteworthy that in some cases this setup outperforms training on a larger corpus (Exp3), even when including the validated data itself (Exp4). In this respect, and given that the datasets include a range of semantic phenomena (e.g., homonymy, different types of polysemy), these results suggest that generalization may be hindered not just by data size or quality, but also by other factors such as the semantic relatedness between training and test instances or the presence or absence of regular polysemy patterns.

5 Conclusions and further work

This paper presented an evaluation of the impact of dataset composition and quality on WiC performance for Galician and Spanish. We began by assessing the quality of publicly available datasets for these languages, followed by a validation process to enhance the reliability of the test sets. In addition, we constructed new manually created datasets for both languages, also verified by expert annotators. To systematically examine the effects of data quality and composition, we conducted five experiments involving controlled variations in both training and test data.

While models appear to handle many ambiguous cases in the automatically constructed datasets, they often struggle when evaluated on the manually created ones. This suggest that contextualized representations may not fully capture fine-grained sense distinctions, and that simpler methods based on cosine similarity can sometimes be more reliable. Regarding training data, our findings suggest that performance depends not only on the amount and validation of data but also on deeper factors that deserve further analysis, such as the proportion of target words shared between training and test sets, or the semantic relatedness and distribution of polysemy patterns across datasets.

It is worth noting that the models used in this study are general-purpose pretrained models, and

[9] The complete results of the baselines are shown in Tables 6 and 7 in Appendix A while the full results by layers and cosine thresholds are reported in Tables 8 and 9 in Appendix B.

[10] Additional evidence for this hypothesis comes from follow-up experiments (not reported here), in which incorporating cosine similarity as a feature into the logistic regression models led to substantial improvements in performance.

Exp	Train	Size	Test	Bas	BERT	Bertinho	XLM	Llama
1	Original	1500	Original	66.4	78.7	78.3	79.6	**81.4**
2	Original	1500	Valid	72.2	79.6	82.2	81.6	**84.0**
3	Original	1500	Manual	**75.7**	53.8	56.7	54.3	52.0
4	Orig+Valid	1950	Manual	**75.7**	56.1	56.1	56.7	55.5
5	Valid	450	Manual	**75.7**	57.2	53.2	56.1	55.5

Table 4: Summary of the best results for each model in Galician across the five experiments. *Bas* is the best baseline (see Table 6 for the complete results).

Exp	Train	Size	Test	Bas	Bertin	RoBERTa	XLM	Llama
1	Original	200	Original	63.1	60.4	61.9	60.5	**64.1**
2	Original	200	Valid	**72.2**	63.2	61.6	61.6	62.7
3	Original	200	Manual	**70.9**	52.0	52.6	53.7	51.4
4	Orig+Valid	570	Manual	**71.4**	54.3	54.3	54.9	55.4
5	Valid	370	Manual	**71.4**	54.9	56.0	56.6	54.2

Table 5: Summary of the best results for each model in Spanish across the five experiments. *Bas* is the best baseline (see Table 7 for the complete results).

not specifically fine-tuned for the WiC task. For future work, we aim to investigate strategies to enhance performance on WiC tasks, ranging from unsupervised methods, such as the WiC-targeted fine-tuning of MirrorWiC (Liu et al., 2021), to supervised fine-tuning approaches exemplified by XL-LEXEME (Cassotti et al., 2023).

Limitations

Models: Regarding the models, our experiments were limited to encoder-based architectures of 'base' and 3B decoder models. As such, the conclusions may not generalize to other types of models, including smaller monolingual decoders or significantly larger multilingual models. Furthermore, all models under consideration are generic pretrained and have not been adapted or fine-tuned specifically for the WiC task.

Data: With respect to the data, some conclusions should be further validated in other languages and with larger datasets. This applies to both training data (which remains limited for Spanish), and evaluation data, especially the manually constructed test sets, which are comparatively small.

Method: As for the evaluation methodology, we rely on standard WiC setups using simple classifiers that operate over concatenated contextual embeddings of the target word. More complex modeling approaches may yield improved results and provide additional insights not captured in this setup.

Analysis: Finally, a more fine-grained analysis would be required to draw robust conclusions about the relationship between training and test corpora. Such an investigation goes beyond the scope of this short paper.

Acknowledgments

This work was funded by MCIU/AEI/10.13039/501100011033 (grants with references PID 2021-128811OA-I00, and CNS2024-154902), by the Galician Government (ED481A-2024-070, ED431F 2021/01, ED431G 2023/04, and ED431B 2025/16), and by a Ramón y Cajal grant (RYC2019-028473-I).

References

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.

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.

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.

Javier De la Rosa, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Marcos Garcia. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3625–3640. Association for Computational Linguistics.

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).

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.

Qianchu Liu, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulic. 2021. Mirrorwic: On eliciting word-in-context representations from pretrained language models. *CoRR*, abs/2109.09237.

Kanishka Misra. 2022. minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint arXiv:2203.13112*.

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, José Camacho-Collados, and Mohammad Taher Pilehvar. 2020. Xl-wic: A multilingual benchmark for evaluating semantic contextualization. *CoRR*, abs/2010.06478.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Vilares, Marcos Garcia, and Carlos Gómez-Rodríguez. 2021. Bertinho: Galician BERT Representations. *Procesamiento del Lenguaje Natural*, 66:13–26.

Marta Vázquez Abuín and Marcos Garcia. 2025. Assessing lexical ambiguity resolution in language models with new WiC datasets in Galician and Spanish. *Procesamiento del Lenguaje Natural*, 74:305–319.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Appendix

A Baseline results

Exp	Train	Size	Test	FastT	BERT	Bertinho	XLM	Llama
1	Original	1500	Original	55.6	62.4	**66.4**	60.0	62.1
2	Original	1500	Valid	58.2	69.8	**72.2**	67.6	68.4
3	Original	1500	Manual	57.8	**75.7**	72.8	74.0	68.2
4	Orig+Valid	1950	Manual	57.8	**75.7**	72.8	74.0	67.6
5	Valid	450	Manual	57.8	**75.7**	74.0	74.0	68.8

Table 6: Baseline results for the five experiments (*Exp*) in Galician. *Size* is the number of instances in the training data, while *Train* and *Test* indicate the datasets used.

Exp	Train	Size	Test	FastT	Bertin	RoBERTa	XLM	Llama
1	Original	200	Original	54.1	56.3	60.5	**63.1**	58.6
2	Original	200	Valid	54.3	61.4	69.2	**72.2**	68.9
3	Original	200	Manual	59.4	58.9	**70.9**	64.0	68.6
4	Orig+Valid	570	Manual	59.4	58.9	**71.4**	64.0	68.6
5	Valid	370	Manual	59.4	58.9	**71.4**	64.6	68.6

Table 7: Baseline results for the five experiments (*Exp*) in Spanish. *Size* is the number of instances in the training data, while *Train* and *Test* indicate the datasets used.

B Layers and cosine thresholds for the best results

Exp	Bas		BERT			Bertinho			XLM			Llama		
	Acc.	Cos.	Acc.	L.	Cos.	Acc.	L.	Cos.	Acc.	L.	Cos.	Acc.	L.	Cos.
1	66.4	0.60	78.7	9	0.54	78.3	9	0.36	79.6	9*	0.46*	**81.4**	11	0.34
2	72.2	0.58	79.6	9	0.52	82.2	9	0.36	81.6	9	0.52	**84.0**	8	0.34
3	**75.7**	0.32	53.8	10	0.36*	56.7	9	0.30	54.3	11	0.54	52.0	21	0.62
4	**75.7**	0.32	56.1	10	0.54*	56.1	9	0.30	56.7	11	0.52*	55.5	16*	0.52*
5	**75.7**	0.32	57.2	10	0.34*	53.2	7	0.40	56.1	11*	0.52*	55.5	22	0.58

Table 8: Summary of the best results for each model in Galician across the five experiments with their layer(s) and cosine threshold(s). Cells marked with * indicate that multiple layer–cosine combinations yielded the same score; in such cases, we report the configuration with the lowest layer index.

Exp	Bas		Bertin			RoBERTa			XLM			Llama		
	Acc.	Cos.	Acc.	L.	Cos.	Acc.	L.	Cos.	Acc.	L.	Cos.	Acc.	L.	Cos.
1	63.1	0.74	60.4	2	0.74	61.9	10	0.58	60.5	12	0.56*	**64.1**	10	0.30
2	**72.2**	0.52	63.2	3	0.68	61.6	6	0.60	61.6	12	0.62	62.7	8	0.34
3	**70.9**	0.58	52.0	8	0.38*	52.6	9	0.38	53.7	11	0.42	51.4	10	0.32
4	**71.4**	0.58	54.3	8	0.38*	54.3	9	0.38	54.9	11	0.42*	55.4	10	0.32
5	**71.4**	0.58	54.9	8	0.38	56.0	9	0.38	56.6	11	0.44	54.2	10	0.32

Table 9: Summary of the best results for Spanish across the five experiments with their corresponding optimal layer(s) and cosine threshold(s). Cells marked with * indicate that multiple layer–cosine combinations yielded the same score; in such cases, we report the configuration with the lowest layer index.