

# An empirical study of Definition Modeling with LLMs for the main languages of Latin America

Erica K. Shimomoto<sup>1\*</sup>, Edison Marrese-Taylor<sup>1\*</sup>, Erique Reid<sup>2</sup>

National Institute of Advanced Industrial Science and Technology<sup>1</sup>

Graduate School of Engineering, The University of Tokyo<sup>2</sup>

{kidoshimomoto.e,edison.marrese}@aist.go.jp, erarvin2007@gmail.com

\*Authors contributed equally to this work.

## Abstract

In this paper, we propose the first multilingual study on definition modeling focusing on languages of Latin America. We collect monolingual dictionary data for Spanish and Portuguese and perform an in-depth empirical study to test the performance of large language models in the definition modeling of monosemic words. A comprehensive human evaluation of the LLM-generated definitions validates our choice of COMET as an automatic evaluation metric, highlighting the zero and few-shot capabilities of these models in this new task and also showing their shortcomings and the limitations of COMET for automatic evaluation.

## 1 Introduction

Large Language Models (LLMs) have recently caused a significant shift in our field by achieving excellent performance on a wide variety of downstream tasks without fine-tuning and instead relying on zero or few-shot approaches (Brown et al., 2020; Kojima et al., 2022). To better understand how these models capture linguistic information, several evaluation and probing tasks have been developed (Blevins et al., 2023; Starace et al., 2023).

Definition modeling is the task of estimating the probability of a textual definition given a word. It has been shown to give a transparent view of how well models capture syntax and semantics. First proposed by Noraset et al. (2017), it can be framed as a conditional generation, in which the definition or phrase is generated given a conditioning variable such as a word’s associated word embedding or other representations of context.

Existing approaches for this task train a model on a corpus of word-definition pairs that is later tested on how well they generate definitions for words not seen during training. Current approaches (Noraset et al., 2017; Gadetsky et al., 2018; Ni and Wang, 2017; Ishiwatari et al., 2019; Reid et al., 2020) are

mainly encoder-decoder based, in which one encodes a contextual representation for a word/phrase using a variety of features such as context or character composition, and then uses the contextual representation(s) to generate the definition.

Though several models and datasets have been proposed, previous work has predominantly focused on English. We find just a handful of papers that work on definitions for other languages, namely VCDM from Reid et al. (2020) for French and the work of Kong et al. (2020) for Chinese. We also find that dictionary data for French and Italian have been released in the past few years but that there are no approaches that leverage them so far (Hathout et al., 2014; Hathout and Sajous, 2016; Sajous et al., 2020; Calderone et al., 2017).

Although these first steps towards including other languages for this task are pivotal, exploring languages from relevant but usually neglected parts of the world, such as Latin America, is critical. In light of this issue, we study the multilingual abilities of instruction-tuned LLMs in two of the main languages of Latin America, Spanish and Portuguese. We evaluate how well LLMs can provide definitions for words in these languages, effectively using the task to probe their multilingual abilities. Also, we present a thorough empirical study based on automatic metrics, including BLEU, BERTScore, and COMET, as well as a comprehensive human evaluation.

Our results show that LLM prompting techniques, either in the zero-shot or few-shot scenarios, lead to worse results when compared to English. Furthermore, we also observe that the output language can be challenging to control and our human evaluation showed that many of the definitions mistakenly generated in English were correct, particularly in the case of Portuguese. Finally, techniques such as in-context learning do not lead to significantly better quality output.

## 2 Proposed Approach

**Data** For our study, we collect data for the two main languages spoken in Latin America, Spanish (485 M native speakers) and Portuguese (236 M native speakers). For each language, we choose a well-known and readily available dictionary to serve as a source: the Dictionary of the Spanish language for Spanish ([dle.rae.es](http://dle.rae.es)), and Dicio ([dicio.com.br](http://dicio.com.br)) for Portuguese.

During the data collection phase, we observed that these dictionaries only sometimes contain examples of word usage. This point presented an issue as previous research has shown that use case examples are critical in allowing models to disambiguate a specific meaning for a given term in the case of polysemy. Although it has been attempted to solve this task without the aid of context (Noraset et al., 2017), we believe this setting is of limited interest and therefore collect data for the terms that exhibit only a single meaning, thus experimenting in a *monosemic* scenario. We leave collecting this word usage information for future work.

During this phase, we also noted that we often collected multiple entries for the same term. A visual inspection of the data showed that inflections of verbs (e.g., tenses) and adjectives (e.g., gender) were sometimes present in the source dictionaries, particularly in DICIO. As this may lead to data leakage problems for our experiments, we utilized `spacy.io` to identify word lemmas, keeping the example where the original term matches the lemmatized word. The resulting datasets were split into the 80/10/10 ratio. Table 1 shows the exact details of our split sizes, compared to the English dataset built from Oxford Dictionaries (Gadetsky et al., 2018) (OXFORD), to which we apply the same preprocessing described above.

Dataset	Train	Valid	Test
OXFORD (en)	15,770	6,884	6,834
DRAE (es)	32,834	4,104	4,105
DICIO (pt)	118,591	14,824	14,824

Table 1: Details of the size of each split for our collected and filtered data, compared against OXFORD.

**Models** We consider two popular white-box LLMs, Llama2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023). For the former, we utilize the 13B-parameter *Llama-2-13b-chat-hf* (LLAMA-2), while for the latter, we use the 7B-parameter model

*Mistral-7B-Instruct-v0.1* (MISTRAL). We quantize these models to 4-bits using QLoRA (Dettmers et al., 2023) to fit into our GPU memory.

We test two settings: (1) zero-shot, where the model is directly asked to generate the word’s definition, and (2) few-shot, where we incorporate term-definition examples in the prompt before requesting the definition for the target term. These shots are randomly sampled from the training data and kept constant across examples. We keep the input to the model as simple as possible, always using the prompt “*Define the {language} word {term}*”. *Use only {language} to reply.*”, where {language} and {term} are variables denoting the target language and the term to define. More information regarding the prompts can be found on the supplementary material A.2 and A.3.

To increase the robustness of our reported results, we repeated each experiment 3 times with different random seeds. In the case of the few-shot scenario, we also sampled a different set of (term, definition) tuples from the training split to feed as context.

**Evaluation** We report BLEU, relying on the `sacrebleu` ([github.com/mjpost/sacrebleu](https://github.com/mjpost/sacrebleu)) implementation (Post, 2018). Since previous work (Reid et al., 2020; Huang et al., 2021; Jhirad et al., 2023) has shown that n-gram overlap metrics do not capture nuance in the definitions generated, we also report BERTScore (Zhang et al., 2019) and COMET (Rei et al., 2020), using the *wmt22-comet-da* model, which has multilingual support.

To evaluate the ability of the LLMs to generate definitions in the target language, we rely on a fasttext-based language classification model (Grave et al., 2018). We compare the target language with the highest probability language predicted, reporting the percentage of examples where the prediction matches the target, which we consider to measure the degree of instruction compliance.

Finally, we also performed a human evaluation on a subset of the definitions generated. We recruited one volunteer native speaker of Brazilian Portuguese and Spanish and asked them to evaluate a set of 100 generated definitions by our best two models. We choose one random seed and pick the 50 best/worst generations based on COMET. For evaluation, we designed an evaluation scheme based on a Likert scale with five values, which we crafted based on a detailed examination of the outputs of the models. For more details, please refer to Table 4 in the supplementary material.

Dataset	Model	S	BL	BS	CMT	Cmp.
OXFORD (en)	LLAMA-2	0	0.949	0.826	<b>0.553</b>	<b>1.000</b>
		5	<b>3.906</b>	<b>0.851</b>	0.537	<b>1.000</b>
	MISTRAL	0	1.183	0.837	<b>0.557</b>	0.999
		5	<b>1.821</b>	<b>0.842</b>	0.529	<b>1.000</b>
DRAE (es)	LLAMA-2	0	0.271	0.617	<b>0.504</b>	<b>0.996</b>
		5	<b>1.360</b>	<b>0.675</b>	0.470	0.877
	MISTRAL	0	0.324	0.650	<b>0.458</b>	0.703
		5	<b>0.559</b>	<b>0.663</b>	0.440	<b>0.852</b>
DICIO (pt)	LLAMA-2	0	0.143	0.682	<b>0.513</b>	<b>0.969</b>
		5	<b>1.828</b>	<b>0.683</b>	0.462	0.883
	MISTRAL	0	0.146	0.689	0.475	0.630
		5	<b>4.030</b>	<b>0.709</b>	<b>0.501</b>	<b>0.926</b>

Table 2: Results of our experiments, where S indicates the number of shots, Cmp. is short for compliance, and BL, BS and CMT are short for BLEU, BERTScore and COMET, respectively. Numbers in bold indicate best results for each language-model combination. Underlined results indicate differences against zero-shot are significant at  $\alpha = 0.05$ .

### 3 Results

Table 2 summarizes the results of our experiments with LLMs. Besides our two target languages, we also include English results as our baseline.

Overall, we observe that both models struggled to deliver a comparable performance in our target languages. However, MISTRAL performed better in Portuguese. One important issue we noticed here is that our LLMs often gave outputs in English.

Furthermore, although our results initially suggested that in-context learning leads to consistent improvements over zero-shot settings, these were not always statistically significant across metrics and languages. We did not perform prompt engineering for the in-context learning, so our results could be considered lower-bound. However, within this task, we believe that while giving examples might help the model learn the dictionary style of giving definitions, if the model does not know the word, it will not be able to give satisfactory results no matter how many examples we provide.

Table 3 summarizes the results of our human evaluation efforts. First, we observe that the Likert score on the Best sets is higher than on the Worst sets, which supports our choice of COMET as a metric for our task. However, we also observe that the correlation in the worst sets is overall smaller. This result is likely because while COMET varied between values of 0.2 to 0.4, these definitions were mostly tagged as “Very poor” across all languages.

Furthermore, we observe that the correlation in the Best set in the zero-shot scenario is almost

Model	Likert	COMET	Correlation
DRAE (es)			
LLAMA-2 + 5 shots	4.680 / 1.100	0.700 / 0.317	0.030 / 0.033
	4.900 / 1.100	0.751 / 0.235	0.056 / -0.100
MISTRAL + 5 shots	4.600 / 1.080	0.727 / 0.258	0.124 / 0.218
	4.440 / 1.140	0.710 / 0.236	0.270 / 0.237
DICIO (pt)			
LLAMA-2 + 5 shots	3.860 / 1.240	0.639 / 0.385	-0.224 / 0.118
	4.000 / 1.460	0.623 / 0.339	0.402 / 0.102
MISTRAL + 5 shots	3.260 / 1.140	0.577 / 0.320	0.089 / 0.181
	1.780 / 1.180	0.530 / 0.306	0.186 / -0.053

Table 3: Results of the human evaluation in terms of average Likert scores (best 50 / worst 50), compared to COMET. We also report the Pearson correlation between these two values among the given samples.

0 for LLAMA-2 in Spanish and negative for Portuguese. In this scenario, we noticed that models often generate very long definitions, which could have impacted COMET’s performance. Furthermore, while LLAMA-2 generations were more precise and correct, MISTRAL generations were rather “Very good”, with rich details, or “Very Poor”, with wrong information; however, COMET could not detect such differences. These findings agree with previous observations by Jhirad et al. (2023) and provide additional evidence to show that the performance of COMET decreases substantially with the length of the inputs.

Our human evaluation also revealed a significant challenge: many of the definitions mistakenly generated in English were correct, particularly in the case of Portuguese. These findings underscore the models’ difficulty in generating definitions outside English, indicating a clear need for improved mechanisms. Based on these insights, we propose a potentially fruitful avenue for further research: studying the task from the perspective of chain-of-thought reasoning (Wei et al., 2022; Kojima et al., 2022), for instance, by asking the models to solve the task using a define-then-translate approach.

### 4 Conclusions

In this paper, we presented a study on definition modeling for Spanish and Portuguese using LLMs. Our results indicate that models may have the correct information regarding a definition, but struggle to generate them in the target language, lacking performance when compared to English. Finally, techniques such as in-context learning did not lead to significantly better quality output.

## Limitations

Our work has a few notable limitations. First, the sources for our dataset forbid the distribution of the data, and therefore, we are not allowed to share our created datasets. Second, while we tackled three different languages, we do not have evidence that our results will generalize to other languages, especially those belonging to different linguistic branches. Third, we recognize that despite their remarkable performance, LLMs may be challenging for many researchers to access, as white-box models require powerful computing resources to run them locally, and black-box models are behind a paywall. Finally, while we demonstrated how LLMs can generate definitions for monosemic words, we have yet to tackle polysemy, an arguably more challenging scenario.

## Ethics Statement

Our main objective is to use the definition modeling task to evaluate the multilingual abilities of LLMs, introducing dictionary datasets for two of Latin America’s main languages. One potential use case is to have a model generate fake definitions that may mislead users who interact with an LLM when deployed. Through the insights of our work, we hope to minimize such risks.

## Acknowledgements

This paper is based on results obtained from the project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). For experiments, computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used.

## References

- Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. [Prompting language models for linguistic structure](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens

Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language Models are Few-Shot Learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Basilio Calderone, Matteo Pascoli, Franck Sajous, and Nabil Hathout. 2017. [Hybrid Method for Stress Prediction Applied to GLAFF-IT, a Large-Scale Italian Lexicon](#). In *Language, Data, and Knowledge*, Lecture Notes in Computer Science, pages 26–41, Cham. Springer International Publishing.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#).

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. [Conditional generators of words definitions](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271, Melbourne, Australia. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. [Learning word vectors for 157 languages](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nabil Hathout and Franck Sajous. 2016. [Wiktionnaire’s Wikicode GLAWified: A Workable French Machine-Readable Dictionary](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1369–1376, Portorož, Slovenia. European Language Resources Association (ELRA).

Nabil Hathout, Franck Sajous, and Basilio Calderone. 2014. [GLÀFF, a Large Versatile French Lexicon](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 1007–1012, Reykjavik, Iceland. European Language Resources Association (ELRA).

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. [Definition Modelling for Appropriate Specificity](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. [Learning to Describe Unknown Phrases with Local and Global Contexts](#). 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

- 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.
- James Jhirad, Edison Marrese-Taylor, and Yutaka Matsuo. 2023. Evaluating Large Language Models’ Understanding of Financial Terminology via Definition Modeling. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 93–100, Nusa Dua, Bali. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7b*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. *Large Language Models are Zero-Shot Reasoners*. In *Advances in Neural Information Processing Systems*.
- Cunliang Kong, Liner Yang, Tianzuo Zhang, Qinan Fan, Zhenghao Liu, Yun Chen, and Erhong Yang. 2020. *Toward Cross-Lingual Definition Generation for Language Learners*.
- Ke Ni and William Yang Wang. 2017. Learning to Explain Non-Standard English Words and Phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI’17*, pages 3259–3266, San Francisco, California, USA. AAAI Press.
- Matt Post. 2018. *A call for clarity in reporting BLEU scores*. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. *COMET: A Neural Framework for MT Evaluation*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2020. *VCDM: Leveraging Variational Bi-encoding and Deep Contextualized Word Representations for Improved Definition Modeling*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6331–6344, Online. Association for Computational Linguistics.
- Franck Sajous, Basilio Calderone, and Nabil Hathout. 2020. *ENGLAWI: From Human- to Machine-Readable Wiktionary*. In *12th International Conference on Language Resources and Evaluation (LREC 2020)*, page 3016.
- Giulio Starace, Konstantinos Papakostas, Rochelle Choenni, Apostolos Panagiotopoulos, Matteo Rosati, Alina Leidinger, and Ekaterina Shutova. 2023. *Probing LLMs for joint encoding of linguistic categories*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7158–7179, Singapore. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shriti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. *Llama 2: Open foundation and fine-tuned chat models*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. *Chain of thought prompting elicits reasoning in large language models*. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. *BERTScore: Evaluating Text Generation with BERT*. In *International Conference on Learning Representations*.

## A Detailed Results

### A.1 Human Evaluation Details

In Table 4, we show the criteria for each of the Likert categories used for the human evaluation of the generated definitions.

### A.2 System Prompts

Regarding the system prompt, for LLAMA-2, we follow the approach by the original paper but replace the original one <sup>1</sup> with the phrase “*You are a helpful assistant. Always answer as helpfully and concisely as possible.*”, to help reduce false refusal rates, i.e., when the model incorrectly refuses to answer a question that it should, for example, due to overly broad instructions to be cautious in how it provides responses. In our case, we observed that this often happened whenever we asked the model to define words that could be offensive in certain contexts. In the case of MISTRAL, we simply utilize the original system prompt setting.

### A.3 Few-shot setting

For the few-shot experiments, our prompt consisted of providing the zero-shot prompt described in Section 2 for a five different sampled word from the training set, each followed by its definition. An example can be seen below:

*[’role’: ’user’, ’content’: "Define the Portuguese word 'aldrave'. Use only Portuguese to reply.", ’role’: ’assistant’, ’content’: 'Aldrave vem do verbo aldravar. O mesmo que: tranque, aferrolhe.', ’role’: ’user’, ’content’: "Define the Portuguese word 'honorificai'. Use only Portuguese to reply.", ’role’: ’assistant’, ’content’: 'Honorificai vem do verbo honorificar.', ’role’: ’user’, ’content’: "Define the Portuguese word 'bicarboneto'. Use only Portuguese to reply.", ’role’: ’assistant’, ’content’: '[Química] Combinação em que o carbono é em quantidade dupla da que há no carboneto; bicarbureto.', ’role’: ’user’, ’content’: "Define the Portuguese word 'dicotomicamente'. Use only Portuguese to reply.", ’role’: ’assistant’, ’content’: 'Dicotomicamente é uma palavra derivada de dicotômico.', ’role’: ’user’, ’content’: "Define the Portuguese word 'desalbardar'. Use only Portuguese to reply.", ’role’: ’assistant’, ’content’:*

<sup>1</sup>[https://github.com/huggingface/transformers/blob/ee2a3400f2a7038a23b83a39c5d0e24f7f699561/src/transformers/models/llama/tokenization\\_llama.py#L57](https://github.com/huggingface/transformers/blob/ee2a3400f2a7038a23b83a39c5d0e24f7f699561/src/transformers/models/llama/tokenization_llama.py#L57), <https://github.com/facebookresearch/llama/blob/main/UPDATES.md>

*'Tirar a albarda a.', ’role’: ’user’, ’content’: "Define the Portuguese word 'abacatal'. Use only Portuguese to reply."]*

This prompt is then converted to the appropriate input format for each model.

### A.4 Statistical Significance Details

Table 5 below provides the details of the p-values for the experiments with LLMs.

### A.5 Role of Quantization

Quantization has been shown to be detrimental to performance in certain scenarios, we also evaluate the influence of this technique on our task. For this study, we limit ourselves to the zero-shot scenario and to DRAE and DICIO, which we believe should serve as a reasonable estimate for the overall performance gap due to the effect of quantization. Again, we run each experiments three times with different random seeds, and report average results, which we summarize in Table 6.

## B Hardware

All of our experiments were performed on a large cluster, where we usually rely on a node with 4 NVIDIA V100 GPUs, or an instance with a single NVIDIA A100 GPU. We spent a total of approximately 2,000 USD in our experimental setup, most of which is due to improving the robustness with of our study by repeating LLM experiments with multiple seeds.

Category	Details	Example
Very poor	A completely wrong definition, one that contains only the word, or its English translation.	<b>Waterspout (<i>Tromba d'agua</i>)</b> Definition: Phenomenon defined by the formation of a mass of vapors over rivers, lakes, seas, generally in the shape of a funnel, a cone with the base facing the clouds. Generated: Sure! "Waterspout" is a term used to designate a type of flooding caused by a river flood or storm surge, which can flood coastal or plain areas.
Poor	A definition of a related word.	<b>To sharpen (<i>Nitidificar</i>)</b> Definition: To make something clear, perceptible, sharp; To clear a substance; the concept became clear with your explanation. Generated: To sharpen is a verb that means "to enhance" or "to beautify." It is used to describe the action of improving or enriching something, it can be applied to objects, situations or people. Example: "The sun on the beach sharpened the sea with its rays".
Acceptable	It gives an idea of the broad sense of the word	<b>Supralegal (<i>Supralegal</i>)</b> Definition: [Legal] Above the law; that, hierarchically, occupies a superior position in relation to legislation. Generated: Of course! The term "supralegal" in Portuguese means "beyond the law" or "outside the law". It is used to describe acts or situations that are illegal or that are not subject to any type of legal regulation.
Good	Gives the correct word definition, but it has some imprecision (e.g. wrong use cases)	<b>To oversize (<i>Superdimensionar</i>)</b> Definition: Grant large dimensions to; give an exaggerated size, an exaggerated scope to something; to exaggerate the achievements of a government. Generated: Of course! The Portuguese "oversize" is a term that is not very common, but can be used to describe an action or process that increases or exaggerates a characteristic or property of something. Most of the time, it is used to criticize situations in which someone is exaggerating or increasing something disproportionately. For example, if someone is exaggerating their ability in a certain subject, it could be said that they are "oversizing" their ability.
Very good	Correct definition with no imprecision.	<b>To sanitize (<i>Sanitizar</i>)</b> Definition: To comply with hygiene requirements or recommendations; to clean; To sanitize an environment. Generation: Of course! The Portuguese "sanitize" refers to eliminating or removing anything that could be considered unclean or dangerous to health. For example, it may refer to disinfecting surfaces or objects, or eliminating bacteria or viruses from an environment. In short, "sanitize" means to purify or disinfect.

Table 4: Explanation of our human evaluation system for generated definitions. Examples are taken from DICIO and manually translated to English by the annotator.

Dataset	Model	BL	BS	CMT	Cmp.
OXFORD (en)	LLAMA-2	0.000	0.000	0.037	-
	MISTRAL	0.008	0.002	0.002	-
DRAE (es)	LLAMA-2	0.011	0.000	0.001	0.011
	MISTRAL	0.096	0.037	0.003	0.021
DICIO (pt)	LLAMA-2	0.185	0.602	0.597	0.295
	MISTRAL	0.027	0.012	0.035	0.000

Table 5: P-values of t-test comparing the model's performance in zero and five shot scenarios.

Dataset	Model	Type	BL	CMT
DRAE (es)	LLAMA-2	4bit	0.271	0.504
		full	0.295	0.510
	MISTRAL	4bit	0.324	0.458
		full	0.333	0.469
DICIO (pt)	LLAMA-2	4bit	0.143	0.513
		full	0.120	0.515
	MISTRAL	4bit	0.146	0.475
		full	0.138	0.489

Table 6: Results of our study on the impact of low precision (4 bit). In the table, Cmp. is short for compliance, the % of cases where the answer is in the correct language, while BL, BS and CMT are short for BLEU, BERTScore and COMET, respectively.