Multilingual Definition Modeling

Edison Marrese-Taylor[1,2], Erica K. Shimomoto[1], Alfredo Solano[2], Enrique Reid[2]
National Institute of Advanced Industrial Science and Technology[1]
The University of Tokyo[2]
edison.marrese@aist.go.jp, kidoshimomoto.e@aist.go.jp
asolano@weblab.t.u-tokyo.ac.jp, erarvin2007@gmail.com

Abstract

In this paper, we propose the first multilingual study on definition modeling. We use monolingual dictionary data for four new languages (Spanish, French, Portuguese, and German) and perform an in-depth empirical study to test the performance of pre-trained multilingual language models on definition modeling of monosemic words when finetuned on this data. Furthermore, we use a zero-shot approach to test the multilingual capabilities of two popular chat-based Large Language Models (LLMs) in the task. Results show that multilingual language models can perform on-pair with English but cannot leverage potential cross-lingual synergies, with LLMs generally offering better performance overall. A comprehensive human evaluation of the LLM-generated definition highlights the zero and few-shot capabilities of these models in this new task, also showing their shortcomings. Finally, we show that performance on our task via BERTScore strongly correlates to the performance on multilingual LLM benchmarks, suggesting that our task offers a viable compute-constrained, stable and natural alternative to these.

1 Introduction

Definitions play a key role in effective communication and language understanding. However, creating high-quality and precise definitions requires time and effort given the complexity of languages and how they are constantly evolving.

Aiming to suffice this necessity, the definition modeling task was first proposed by Noraset et al. (2017). The goal is to estimate the probability of a textual definition given a word being defined. This task can be applied to generate definitions for known and new words, assisting language learners, as well as serving as a tool for language preservation and description (Kabiri and Cook, 2020; Dimas Furtado et al., 2024).

Since its conception, several approaches have tackled this task, showing that it gives an arguably more transparent view of how much a model captures syntax and semantics. So far, existing approaches for this task are trained on a corpus of word-definition pairs, to be 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) are mainly encoder-decoder based, in which a contextual representation for a word/phrase, encoded using a variety of features, is used to generate the definition.

Despite the progress, previous work has predominantly focused on definition modeling in English. Concretely, we find just a handful of papers that work on definition modeling for languages other than English, namely Reid et al. (2020) for French, Dimas Furtado et al. (2024) for Portuguese and 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). Finally, while Kabiri and Cook (2020) has covered nine different languages in their study, they did not explore multilingual capabilities of models, focusing on performance of monolingual models on the polysemy aspect of words. Moreover, we note that these datasets are mostly based on crowdsourced dictionaries, where quality can be a concern.

In light of these issues, in this paper, we present the first *truly* multilingual study on definition modeling, introducing datasets obtained from not-crowdsourced dictionaries in French, Spanish, Portuguese, and German, with a total of approximately 460K new terms entries and 730K definitions, thus substantially extending language coverage for the task. We combine our collected data with existing corpora to train models and lay out the first set

Findings of the Association for Computational Linguistics: ACL 2025, pages 25888–25906

of experiments on multilingual definition modeling. As our main focus is to analyze the potential synergy between different languages, we focus our experiments on monosemic words, hoping to explore polysemy in future work.

Furthermore, we also test a selection of Large Language Models (LLMs) on our collected data, probing their multilingual abilities in this task through zero-shot and few-shot approaches (Brown et al., 2020; Kojima et al., 2022). We evaluate performance based on three different automatic metrics, i.e., BLEU, BERTScore and COMET, and through human evaluation.

Our results show that current multilingual language models such as mBART (Liu et al., 2020; Tang et al., 2020) and mT5 (Xue et al., 2021) can perform on-pair with English when fine-tuned on single-language dictionaries. They can also deliver consistent performance when trained in multiple languages, but are unable to leverage cross-lingual synergies to consistently improve performance. As for LLMs, we show that prompting techniques, either in the zero-shot or few-shot scenarios, offer generally better performance than multilingual language model fine-tuning, even in the case of LLMs. However, we also observe that the output language can be difficult to control and that techniques such as in-context learning do not lead to significantly better quality output.

Finally, we show empirical evidence suggesting that per-language performance in our task measured via BERTScore is strongly correlated with performance on multilingual benchmarks. Opposite to these, which are often translations of datasets in English, our task can offer a compute-constrained, static and natural alternative to assess the proficiency of LLMs on a given language. We release our code and data on GitHub[1].

2 Related Work

Our paper is primarily related to the seminal work by Noraset et al. (2017); Hill et al. (2016), in which a model is tasked with generating a definition for a word given its respective embedding or with mapping dictionary definitions to lexical representations of words, respectively. Later work has proposed improvements and extensions, introducing techniques and datasets to address shortcomings. For example, Gadetsky et al. (2018) address polysemy and presents a dataset from Oxford Dictionaries, where each definition is supplemented with context sentences. Ni and Wang (2017) proposed an approach for automatically explaining slang English terms in a sentence and introduced another dataset extracted from The Urban Dictionary. Ishiwatari et al. (2019) proposed to further rely on local and global contexts to help the model disambiguate and generate better definitions. Different from our approach, their work focused on unfamiliar words and phrases (many of which are polysemous words), rarely used idioms, or emerging entities, relying on Wikipedia as a source.

More recently, Huang et al. (2021) proposed to study the problem of definition specificity, creating a method for tuning a model to account for hyper-focused (over-specific) or highly general (under-specific) definitions. Finally, Chen and Zhao (2022) proposed to unify the seminal ideas of reverse dictionary and definition modeling in a single model, with the goal of helping better understand word sense and embeddings. Periti et al. (2024) proposed to directly finetune an LLM specifically for definition modeling.

Previous work discussed so far has mostly focused on definitions in the English language. There are a few exceptions, such as the work of Reid et al. (2020), who presented the first study on definition modeling for the French language with the release of a dataset collected from Le Petit Robert, and the work of Dimas Furtado et al. (2024), whose main goal was to collect a dataset for Portuguese definition generation, leveraging several models. Their approach is, however, still monolingual.

We also note that over the past few years, dictionary datasets in several languages derived from Wiktionary have been released, including English (Sajous et al., 2020), French (Hathout et al., 2014; Hathout and Sajous, 2016), Italian (Calderone et al., 2017). Ylonen (2022) collected data for all available Wiktionaries. These datasets, were not accompanied by models that leverage them, and we were unable to find other works using them to train or evaluate models on definition modeling. We think one critical point here is that Wiktionaries are built on the base of crowdsourcing, where quality could be a concern.

Finally, we find previous works aiming to generate definitions in specialized fields, e.g., scientific and medical terms (August et al., 2022), biomedical terminology definition (Liu et al., 2021; Huang et al., 2022), or financial terminology (Jhirad et al., 2023).

[1] https://github.com/epochx/defmod

3 Data

We are interested in assessing the capabilities of multilingual models to describe words in multiple languages. Therefore, we aim to collect high-quality data for "traditional" dictionary definitions, where most common words in a given language are contained. This approach deviates from recent previous work in definition modeling, where the focus has been on domains of "uncommon" words like slang terms (Ni and Wang, 2017) or rarely used idioms and emerging entities (Ishiwatari et al., 2019) and specific domains such as scientific or financial terms (August et al., 2022; Liu et al., 2021; Huang et al., 2021; Jhirad et al., 2023), and only in English.

To support our study, we utilize data for three of the Indo-European languages with the largest number of speakers, according to Wikipedia: Spanish (485 M native speakers), Portuguese (236 M native speakers), and French (80 M native speakers). As we are particularly interested in studying potential performance gains when training on multiple languages, we focus on these in particular, as they all derive from the Romance branch and thus share many similarities. Finally, we also consider German, another Indo-European language but from the Germanic family with 75 M native speakers. We deem German as the ideal candidate to experiment together with existing datasets for the English language.

Dataset (language)	Terms	Defs.	Mono.
OXFORD - en	36,767	122,319	44.07%
Le Petit Robert - fr	33,507	-	-
DRAE - es (ours)	82,386	159,863	60.85%
DICIO - pt (ours)	191,499	279,985	77.41%
LAROUSSE - fr (ours)	72,486	130,408	70.06%
DUDEN - de (ours)	110,444	164,086	73,17%

Table 1: Summary of our collected datasets, compared to prior relevant corpora. In the table, Mono. stands for Monosemic, i.e. terms with a single definition. We compare against OXFORD and Le Petit Robert (data from Reid et al. (2020)), which we regard as existing datasets extracted from "traditional" dictionaries.

For each language, we start by collecting word lists, which are put together from various online sources, including but not limited to MUSE (Lample et al., 2018). We found inflections of verbs (e.g., tenses) and adjectives (e.g., gender) are present and relied on spacy[2] to identify word lemmas, utilizing the *en_core_web_sm* model for English as well as the *core_news_sm* model for German, Spanish, Portuguese and French. After obtaining the lemmatized version of each entry in our data, we kept the example where the original term matches the lemmatized word, similarly to Ishiwatari et al. (2019).

In order to obtain good quality definitions for the terms collected, we work with traditional, well-known and readily available dictionaries. Concretely, we rely on the Dictionary of the Spanish language ("Diccionario de la lengua española - DLE") developed and maintained by the Royal Spanish Academy (Real Academia Española - RAE) for Spanish, Dicio for Portuguese, the Larousse Dictionary for French, and the Duden Dictionary for German. Collecting data from online dictionaries can involve accessing copyrighted information. In this sense, our approach follows previous work (Gadetsky et al., 2018; Ni and Wang, 2017; Ishiwatari et al., 2019; Reid et al., 2020; Dimas Furtado et al., 2024) in making sure that the data can indeed be downloaded/distributed. Based on our analysis of the Terms of Service for the consulted dictionaries, we will not directly distribute our datasets, but instead release recipes on how to reconstruct them based on our compiled word lists. Please see §A for further details about the elaboration of our datasets.

Table 1 summarizes the main characteristics of the data we collected, compared against existing resources in English and French, where we can see that our collected data is substantially richer, both in terms of the number of terms and definitions.

4 Experimental Setup

Data An important distinction between our corpus and recent prior work is that our collected data does not contain examples of word usage. While we acknowledge the importance of such context to disambiguate a specific meaning for a given term in the case of polysemy, we opt to explore *monosemic* words as a first step into the study of multilinguality for definition modeling. We thus select the subset of terms that exhibit only a single meaning, which in practice we achieve by simply selecting the terms with a single definition entry. The final word lists are split into the 80/10/10 ratio. Table 2 below shows the exact details of our splits.

In addition to our data, we also use the previ-

[2] https://spacy.io/

ously released dataset built from Oxford Dictionaries (Gadetsky et al., 2018) (OXFORD). Though we were also interested in using the French data collected by Reid et al. (2020), we found it was not available at the time of writing this paper.

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
LAROUSSE (fr)	31,224	3,903	3,904
DUDEN (de)	55,521	6,940	6,941

Table 2: Details of the size of each split for our collected data filtered for polysemy, compared against OXFORD.

Model Finetuning We train models based on mT5 (Xue et al., 2021). This model is a multilingual version of T5 (Raffel et al., 2020), a generative Transformer pre-trained on a Common Crawl-based dataset covering 101 languages and including a data for a variety of tasks which have been converted into a text-to-text format. We begin by training monolingual baselines for each language, finetuning each dataset. These models give us a set of robust baselines. We then train combining data for several languages. We utilize the "large" model (mT5) (approx. 1.5 B parameters) with the prefix "*Define in* {*language*}", where the variable {*language*} indicates the language in which we would like the model to generate (English, Spanish, Portuguese, French, German). During training, we randomly sample (word, definition) pairs from the data involved and use these examples to create multilingual batches. In preliminary experiments, we also tested alternative multilingual models, but found that mT5 performed better overall.

Model Prompting Given their success in solving several NLP tasks, we also evaluate the ability of instruction-tuned or instruction-tuned LLMs to provide definitions for words. We consider white-box models including Llama 2 (Touvron et al., 2023), Llama 3.1 (Grattafiori et al., 2024), and Mistral v1 models (Jiang et al., 2023). Specifically, we utilize *Llama-2-13b-chat-hf* (LLAMA-2), *Llama-3.1-8B-Instruct* (LLAMA-3), *Mistral-7B-Instruct-v0.1* (MISTRAL), respectively. Via preliminary experiments using 4-bit quantization with QLoRA (Dettmers et al., 2023) on a subset of languages and models, we observed that the effects of quantization were bounded $<1\%$ of the full model performance, and affected the datasets evenly, and thus adopted this strategy for all our LLM experiments.

We test two settings: (1) zero-shot, where the model is directly asked to generate the definition of the word, 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.

Prompting LLMs can lead to substantially different outputs depending on how instructions are crafted. For our experiments, we kept 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. We repeat each experiment 3 times with different random seeds. In the case of few-shot, we also sample a different set of (term, definition) tuples from the train portion of the dataset to use as context.

Evaluation Previous work in definition modeling has mainly utilized n-gram overall metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005). As the latter is language-specific, here we report BLEU, relying on the sacrebleu[3] implementation (Post, 2018). Previous work (Reid et al., 2020; Huang et al., 2021) has also shown that metrics based on n-gram overlap cannot capture nuance in the definitions generated by models. As an alternative, approaches have relied on metrics based on Machine Learning, such as BERTScore (Zhang et al., 2019). Following Jhirad et al. (2023) we also experiment with COMET (Rei et al., 2020), using the *wmt22-comet-da* model, which has multilingual support as it is built on top of XLM-R (Conneau et al., 2020).

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). Here, compare the desired target language with the highest probability language as predicted by this model. We report the percentage of examples where the prediction matches the desired target language (Compliance).

Finally, we also perform human evaluation on a subset of the definitions generated by our models. We split our study into two subsets: Romance and Germanic languages, with particular emphasis on the former, due to their linguistic similarities and relative higher number of first language speak-

[3]https://github.com/mjpost/sacrebleu

ers. We recruit three volunteer native speakers of Brazilian Portuguese, European French, and Spanish, and one native speaker of American English, and German. To make sure the evaluations are consistent across languages, we designed an evaluation scheme based on a Likert scale with 5 values (1 to 5), which we crafted based on a detailed examination of the outputs of the models.

Please see §B for more details on models, setup, training and inference, including the results of our study on quantization and perfomance the model used for language identification. For details about our evaluation scheme, please refer to §D.

5 Results

5.1 Automatic Evaluation

Table 3 summarizes the results of our multilingual experiments using mT5. We observe that this model can perform consistently when trained in multiple languages, as measured by COMET and BERTScore. When it comes to BLEU scores, although we observe little to no improvements over the monolingual baselines in the validation set, we see a more consistent behavior on the test portion. However, we note that the definitions generated by our models are often correct but expressed using different surface forms compared to the gold standard, which can lead to low BLEU scores.

Dataset	Training Data	BL	CMT	BS
DUDEN (de)	de	**0.939**	**0.346**	**0.664**
	de + en	0.507	0.345	0.656
OXFORD (en)	en	**1.330**	**0.345**	0.813
	en + fr	0.038	**0.345**	0.778
	en + de	0.863	0.344	**0.839**
	en + fr + es + pt	0.532	**0.345**	**0.839**
DRAE (es)	es	7.700	**0.384**	0.743
	es + fr	8.084	0.382	0.743
	es + pt	9.158	0.382	0.746
	es + pt + fr	**9.244**	0.381	**0.747**
	en + fr + es + pt	5.503	**0.384**	0.727
LAROUSSE (fr)	fr	0.490	**0.300**	0.685
	fr + es	0.936	0.293	0.686
	fr + pt	0.587	0.298	0.675
	fr + en	0.104	0.294	0.644
	fr + es + pt	**2.660**	0.292	**0.705**
	fr + en + es + pt	0.748	0.293	0.682
DICIO (pt)	pt	11.539	0.352	0.732
	pt + es	18.626	0.352	**0.748**
	pt + fr	4.969	**0.353**	0.682
	pt + fr + es	**18.856**	0.352	**0.748**
	en + fr + es + pt	4.542	0.351	0.685

Table 3: Results of our multilingual experiments based on mT5, where BL is short for BLEU, CMT is short for COMET, and BS is short for BERTScore.

Overall, our results show that even for languages from the same family, where we often find that similar words are associated to similar or related meanings (e.g. words sharing the same Latin root in the case of the Romance languages), exposing models such multilingual lexical data did not help them generalize across languages. However, training models in this fashion does not lead to degraded per-language performance either. A possible explanation for these observations is that a significant portion of the cross-lingual term-definition relatedness is rooted in polysemic terms.

Table 4 summarizes the results of our experiments with LLMs. Compared to our finetuning experiments, LLMs generally offer better performance in terms of COMET and BERTScore and are competitive in terms of BLEU. One important issue here is that LLMs are sometimes unable to follow instructions, which, in our case, often leads to the model generating outputs in English.

When it comes to differences due to in-context learning (ICL), although our results may suggest that this leads to improvements, these were not always statistically significant across metrics and languages, showing that ICL may not be as effective for this task as suggested by previous work.

Role of Prompt Language As explained in §4, the prompt utilized in our experiments with LLMs is provided in English. Since this may bias models toward producing output in this language, we performed a study using a prompt in Spanish on the DRAE dataset. In this case, we use the following prompt template: "Define la palabra '{term}'. Utiliza solo español en tu respuesta.", a loose translation of our English prompt, where '{term}' is a placeholder for the actual word. We follow the same methdology and test on LLAMA-2 and LLAMA-3, on the 0-shot and 5-shot scenarios. Table 5 summarizes our results and compares them against our baseline prompt in English. As shown in the table, results show that prompting models in English leads to consistent superior performance accross all metrics. This validates our choice to prompt models solely in English, and supports our observations during early experimentation. These results also suggest that more recent models offer increased abilities to follow instructions in languages other than English. However, a significant gap still exists.

Dataset	Model	S	BL	BS	CMT	Cmp.
DUDEN (de)	LLAMA-2	0	0.180	0.612	**0.439**	**0.685**
		5	**0.758**	**0.637**	0.378	0.604
	MISTRAL	0	0.159	0.621	**0.411**	**0.762**
		5	**0.233**	**0.623**	0.362	0.628
	LLAMA-3	0	3.261	0.676	0.580	**0.998**
		5	**7.967**	0.695	0.558	0.982
OXFORD (en)	LLAMA-2	0	0.949	0.826	**0.553**	-
		5	**3.906**	**0.851**	0.537	-
	MISTRAL	0	1.183	0.837	**0.557**	0.999
		5	**1.821**	**0.842**	0.529	-
	LLAMA-3	0	1.588	0.835	0.568	-
		5	**3.571**	0.847	0.537	-
DRAE (es)	LLAMA-2	0	0.271	0.617	**0.504**	**0.996**
		5	**1.360**	**0.675**	0.470	0.877
	MISTRAL	0	0.324	0.650	**0.458**	0.703
		5	**0.559**	**0.663**	0.440	**0.852**
	LLAMA-3	0	1.855	0.670	0.545	**0.999**
		5	**3.067**	0.683	0.530	0.974
LAROUSSE (fr)	LLAMA-2	0	0.326	0.631	**0.473**	**0.970**
		5	**1.350**	**0.676**	0.469	0.966
	MISTRAL	0	0.327	0.648	**0.441**	0.775
		5	**0.626**	**0.664**	0.425	**0.951**
	LLAMA-3	0	3.443	0.676	**0.543**	**0.999**
		5	**6.635**	**0.685**	0.513	0.950
DICIO (pt)	LLAMA-2	0	0.143	0.682	**0.513**	**0.969**
		5	**1.828**	**0.683**	0.462	0.883
	MISTRAL	0	0.146	0.689	0.475	0.630
		5	**4.030**	**0.709**	**0.501**	**0.926**
	LLAMA-3	0	0.457	0.683	0.542	**0.998**
		5	**2.298**	**0.715**	**0.549**	0.982

Table 4: Results of our experiments with LLMs, where S indicates the number of shots, and Cmp. is short for compliance, the % of cases where the answer is in the correct language. 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 statistically significant at level $\alpha = 0.05$.

Dataset Contamination A key aspect to consider when analizing our results is the the possibility that portions of our benchmark have appeared in the training data of the LLMs that we tested (Sainz et al., 2023; Deng et al., 2024; Li et al., 2024). An important point here is to first clarify what "contamination" should mean in the context of our task. We think exposing LLMs to actual word-definition pairs during training can and should be considered a kind of contamination. However, if models are simply exposed to examples of words/terms in context, this should not count as a case of contamination. Moreover, the degree of exposure of the LLMs to these contexts is highly relevant to our task and should be studied. One possibility is to work only with fully open-source LLMs, which would enable us to analyze the training data. This is left for future work.

Model	S	L	BL	BS	CMT	Cmp.
LLAMA-2	0	en	**0.271**	**0.617**	**0.504**	**0.996**
		es	0.186	0.601	0.502	0.814
	5	en	**1.360**	**0.675**	**0.470**	**0.877**
		es	0.019	0.526	0.409	0.078
LLAMA-3	0	en	**1.855**	**0.670**	**0.545**	-
		es	0.303	0.619	0.532	-
	5	en	**3.067**	0.683	**0.530**	**0.974**
		es	2.885	**0.711**	0.529	0.908

Table 5: Performance on Spanish definition modelling when using prompts in English (en) and Spanish (es), tested on DRAE, were S and L indicate the number of shots and prompt language respectively. BL, BS and CMT are short for BLEU, BERTScore and COMET, respectively. Numbers in bold indicate best results per model.

5.2 Human Evaluation

We further evaluate the output of the 5-shot experiments on all of our languages for LLAMA-2 and MISTRAL via a human study. We use 200 randomly-sampled terms, for one random seed across languages, totaling 400 definitions to be evaluated per language. We additionally require the annotators to provide us with an assessment of the quality of the gold standard definitions we collected via the same 1-5 Likert score defined in §4. During the whole evaluation procedure, we allow the annotators to see gold-standard definitions and consult external sources if needed. Table 6 summarizes our results.

Inter-annotation agreement We report Cohen's Kappa (κ) scores, which we regard as a lower bound in our case, as it assumes label independence. Since our Likert scores are inherently a ranking, we also report Kendall's Tau (τ), which accounts for this and offers a more realistic agreement measure. Finally, we also report the percentage of examples where the annotators have an exact match (**EM**) in their answers. For languages where we have more than two evaluators, we report the average value of all posible pairwise combinations.

As shown in Table 6 we first see a clear gap in performance between English and in the rest of the languages, which is in agreement with our automatic evaluation. These results further confirm

Dataset	Model	Score	IAA		
			κ	τ	EM
DUDEN (de)	LLAMA-2	2.050	-	-	-
	MISTRAL	1.950	-	-	-
	Data	3.825	-	-	-
OXFORD (en)	LLAMA-2	4.830	-	-	-
	MISTRAL	4.710	-	-	-
	Data	4.925	-	-	-
DRAE (es)	LLAMA-2	2.303	0.520	0.788	0.675
	MISTRAL	1.815	0.616	0.836	0.785
	Data	4.565	0.134	0.148	0.600
DICIO (pt)	LLAMA-2	1.812	0.337	0.594	0.633
	MISTRAL	1.352	0.281	0.446	0.777
	Data	2.967	0.164	0.920	0.355
LAROUSSE (fr)	LLAMA-2	2.720	0.449	0.652	0.563
	MISTRAL	2.133	0.581	0.731	0.700
	Data	4.972	0.083	0.125	0.965

Table 6: Results of our human evaluation on our best results with LLMs, in terms of average Likert scores, along with their IAA measured via average Cohen's Kappa (κ), Kendall's Tau (τ) and exact match (**EM**).

that all three models struggle with Portuguese and German, at least in comparison with the other European languages, despite reportedly being trained on a similar amount of text as is the case of LLAMA-2, where the training mixture contains 0.17% de, 0.16% fr, 0.13% es (Touvron et al., 2023).

Our human evaluation also shows substantial variation in terms of Likert scores for the quality of the gold standard definitions, suggesting that DICIO and DUDEN may offer comparatively lower quality data. In contrast, although results on LAROUSSE exhibit particularly low κ and τ values, we see an **EM** of 0.97. Given the high average score of 4.9, this shows that although the annotators may disagree about the exact score for a given definition, overall they think that definitions are of excellent quality.

Finally, we allowed the annotators to provide us with comments, and one concrete issue that was brought to our attention was that these dictionaries would often offer just a synonym as a definition for a given term. This phenomenon may respond to historical reasons (originally, dictionaries were printed) or be due to the editorial of each company. Nevertheless, if we regard these scores as an upper bound for the human evaluation, we find that in all cases there is still a significant gap between this and the performance of the LLMs. We performed an additional human study to gain insight into the effect of ICL and the effectiveness of COMET as a metric, please see §D for details.

5.3 Multilingual benchmark comparison

We study the relationship between our proposed task and existing datasets used to measure the multilingual performance of LLMs. Concretely, we compare per-language performance for each one of our metrics against multilingual versions of HellaSwag (Zellers et al., 2019), ARC-Challenge (Clark et al., 2018) and MMLU (Hendrycks et al., 2021), provided by Lai et al. (2023) via the *lm-evaluation-harness*[4] platform (Gao et al., 2023). Together, these datasets contain multiple choice questions spanning a wide variety of tasks, including adversarial commonsense natural language inference, abstraction and reasoning ability acquisition via demonstrations, and questions derived from diverse fields of knowledge.

We found that per-language performance in our task, measured via BERTScore, strongly correlates with all of our selected multilingual benchmarks, with average values of 0.94, 0.95 and 0.87 for LLAMA-2, MISTRAL and LLAMA-3, respectively. Please see Table 12 for the details of all the correlation values we obtained, and §C.3 for the exact benchmark performance we measured. Compared to these, which are machine-translations of the originals in English, we think the definition modeling task can offer a natural, stable alternative to assess the proficiency of LLMs in a given language.

Furthermore, to understand the computational trade-offs of using our task instead of our studied multilingual benchmarks, we compare the correlation between performance on the latter and in our datasets, versus subsamples of different sizes (from 0.1% to 5%). For each size, we sample 10 times and report average correlation and confidence intervals via Gaussian-based asymptotic approximation.

Figure 1 summarizes our findings. The computational budget required for our dataset is roughly equivalent to 0.5% of MMLU, 18% of ARC-Challenge or 36.7% or HellaSwag in number of tokens (more details available in Table 17). This means that definition modeling can offer an efficient alternative to measuring the LLM performance on a specific language, particularly when compared with computationally-intensive benchmarks like MMLU. Our datasets can also be combined with modern sampling approaches for bechmark score approximation such as IRT (Polo et al., 2024), leading to further efficiency gains.

[4] github.com/EleutherAI/lm-evaluation-harness

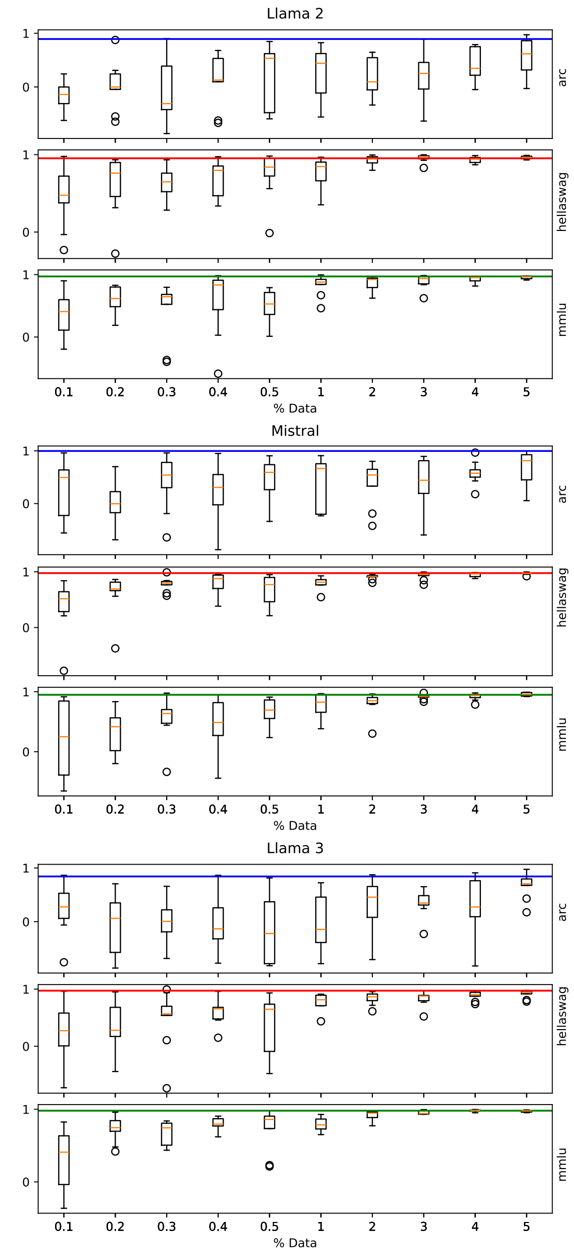

Figure 1: Pearson correlation between performance on multilingual benchmarks and on our task via BERTScore (solid lines), versus the correlation between the same benchmarks and random samples of different sizes for our three tested models LLAMA-2 (top), MISTRAL (center), and LLAMA-3 (bottom) [Improve this! –EMT].

5.4 Comparison with prior work

We compare our results against relevant prior work, focusing on the English language via OXFORD. When it comes to models finetuned on dictionary data, we consider the approaches by Ishiwatari et al. (2019), who proposed a local-and-global context model based on word embeddings, Reid et al. (2020), who leveraged BERT (Devlin et al., 2019) and combined it with a variational inference framework, Huang et al. (2021), who propose a specificity-sensitive approach with based on T5 finetunes and re-ranking (RR), and Periti et al. (2024), who finetuned LLMs to generate definitions of words in context. We point out that all of these approaches work with the complete OXFORD dataset and rely heavily on example sentences for disambiguation.

For zero-shot and few-shot approaches, we compare our results against the values reported by Jhirad et al. (2023). To perform an apples-to-apples comparison against Periti et al. (2024), we replicate their setup using our data by finetuning LLAMA-3 in the training portion of the concatenation of our dictionary data. We utilize the same prompt introduced in §4, which does not substantially differ from what the authors use, save from the fact that we do not provide example sentences. We follow their training procedure, using 4-bit quantization (Dettmers et al., 2023) via PEFT with their best hyperparameter configuration.

Table 7 summarizes our comparison against relevant prior work. We see a substantial gap of BLEU scores exists between models tested on

Model	BL	CMT	BS
Full OXFORD			
LoG-CAD (Ishiwatari et al.)	25.19	-	-
VCDM (Reid et al.)	27.38	0.570◇	-
T5 + RR (Huang et al.)	26.52	-	-
LLama3Dict. (Periti et al.)	21.98	-	0.889
Monosemic only OXFORD			
mT5 + ft (ours)	1.33	0.345	0.839
FLAN-UL2 (Jhirad et al.)◇	2.63	0.457	-
+ examples	4.03	0.501	-
+ 2 shots	3.40	0.475	-
+ 2 shots + examples	4.40	0.505	-
LLAMA-2 (ours)	0.95	0.553	0.826
+ 5 shots	3.91	0.537	0.851
MISTRAL (ours)	1.18	0.557	0.837
+ 5 shots	1.82	0.529	0.842
LLAMA-3 (ours)	1.59	0.568	0.835
+ 5 shots	3.57	0.847	0.847
+ ft	0.57	0.511	0.806
LLama3Dict (Periti et al.)†	0.87	0.455	0.799

Table 7: Comparison of our best models against state-of-the-art approaches on OXFORD. Note that for the "Full OXFORD" section of the table, we compare against models trained/tested on the entire dataset, which differs substantially from ours. BL, BS and CMT are short for BLEU, BERTScore and COMET, respectively; † observed results, ◇ reported by Jhirad et al. (2023).

the full OXFORD dataset, and its monosemic-only subset. These differences are not so significant when looking at COMET or BERTScore. We also note that LLama3Dict, a model trained on dictionary data with example sentences, performs significantly worse in the absence of these examples (BLEU of 21.98 decreased to 0.87). While our similar LLAMA-3 finetune is able to outperform LLama3Dict in this scenario in terms of BERTScore, we see both models perform worse than the base LLMs even when combined with ICL across all metrics. In fact, we observed (Table 13) that finetuning on dictionary data, with our without example sentences, decreased performance with respect to the base model in all our studied languages for the monosemic scenario. Finally, our finetuned mT5, although trained only on monosemic data, offers competitive results with the best models in terms of BERTScore.

Overall, we think these results show that the task of deriving the meaning of a word from example sentences differs substantially from the task of eliciting definitions purely by recalling concepts learned during pretraining. For the latter task, which is the focus of this paper, it is still unclear why LLMs finetuned on dictionary data without context sentences do not perform better than base LLMs with ICL. Based on the characterization recently introduced by Pan et al. (2023), we think it is possible that both finetuning and ICL lead to improvements in terms of Task Recognition (TR) but not in Task Learning (TL): by being exposed dictionary data, models learn to mimic the *style* in which definitions should be provided, but without actually leading to a better ability to recall definitions.

Our results can also be explained in light of the discussion by Agarwal et al. (2024), who point out important limitations of the next-token prediction loss as an indicator of downstream performance. It is possible that the low performance is due to a lack of exposure to lexically-rich data during pretraining, especially for languages other than English, which naive SFT cannot compensate for.

6 Conclusions

In this paper, we presented the first multilingual study on definition modeling, introducing monolingual dictionary data for German, Spanish, French and Portuguese. Results show that LLMs can provide generally better results compared to smaller multilingual LMs trained on monosemic dictionary data, but they can be unstable and generate outputs outside the target language. Furthermore, techniques such as ICL and finetuning did not lead to significantly better quality output. Finally, we show that our task can serve as a natural, stable alternative to assess the proficiency of LLMs in a language.

For future work, we would like to incorporate more languages and perform further human evaluation to better understand the relationship between human assesments and automatic metrics. We would also like to tackle the disambiguation problem by collecting usage examples to tackle polysemy. Our supplementary material offers preliminary evidence suggesting that using LLMs could be a viable alternative for this (§F).

Acknowledgements

This paper is based on results obtained from the by the National Institute of Advanced Industrial Science and Technology (AIST) policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain", and was also partially funded by project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). The computational resources of the AI Bridging Cloud Infrastructure (ABCI), provided by AIST, were used for most of the experiments in this paper. We are also grateful to the NVIDIA Corporation, which donated one of the GPUs we used. Finally, we would also like to thank our volunteer annotators Daniela Serien, Carolina Dias Alexiou, Mauricio Cordova, Lucie Kunitomo-Jacquin, Antoine Jacquin and Denis Higashi Imoto, for their williness to help and feedback.

Limitations

There are a few notable limitations of our work. First, while we tackled five different languages, there is no evidence that our results will generalize to other languages, especially those belonging to different linguistic branches. Furthermore, we recognize that despite their remarkable performance, LLMs may be challenging to access for many researchers, as white-box models require powerful computing resources to run them locally, and black-box models are behind a paywall. Finally, we demonstrated how LLMs can help generate use cases for polysemic words, but have not tackled polysemy.

Ethics Statement

Our main objective is to propose a new task to evaluate the abilities of LLMs. One potential use-case is to have a model generate fake definitions that may mislead users that interact with an LLM when deployed.

References

Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Luis Rosias, Stephanie C.Y. Chan, Biao Zhang, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. In *ICML 2024 Workshop on In-Context Learning*.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating Scientific Definitions with Controllable Complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237, Los Angeles, California. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. 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, and 12 others. 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.

Pinzhen Chen and Zheng Zhao. 2022. A Unified Model for Reverse Dictionary and Definition Modelling. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 8–13, Online only. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

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.

Chunyuan Deng, Yilun Zhao, Yuzhao Heng, Yitong Li, Jiannan Cao, Xiangru Tang, and Arman Cohan. 2024. Unveiling the spectrum of data contamination in language model: A survey from detection to remediation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16078–16092, Bangkok, Thailand. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anna Beatriz Dimas Furtado, Tharindu Ranasinghe, Frederic Blain, and Ruslan Mitkov. 2024. DORE: A dataset for Portuguese definition generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5315–5322, Torino, Italia. ELRA and ICCL.

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.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and

5 others. 2023. A framework for few-shot language model evaluation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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).

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

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.

Jie Huang, Hanyin Shao, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. 2022. Understanding Jargon: Combining Extraction and Generation for Definition Modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3994–4004, Abu Dhabi, United Arab Emirates. 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, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Arman Kabiri and Paul Cook. 2020. Evaluating a multi-sense definition generation model for multiple languages. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 153–161. Springer.

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. *Preprint*, arxiv:2010.05533.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024. An open-source data contamination report for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541, Miami, Florida, USA. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising

Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zequn Liu, Shukai Wang, Yiyang Gu, Ruiyi Zhang, Ming Zhang, and Sheng Wang. 2021. Graphine: A Dataset for Graph-aware Terminology Definition Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3453–3463, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

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.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What In-Context Learning "Learns" In-Context: Disentangling Task Recognition and Task Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8298–8319, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Francesco Periti, David Alfter, and Nina Tahmasebi. 2024. Automatically Generated Definitions and their utility for Modeling Word Meaning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14008–14026, Miami, Florida, USA. Association for Computational Linguistics.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinybenchmarks: evaluating LLMs with fewer examples. In *Forty-first International Conference on Machine Learning*.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

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.

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.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. *Preprint*, arxiv:2008.00401.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as Machine-Readable Structured Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1317–1325, Marseille, France. European Language Resources Association.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

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 Data Sources

A.1 Word Lists

- For German: https://gist.github.com/MarvinJWendt/2f4f4154b8ae218600eb091a5706b5f4
- For Spanish: https://github.com/keepassxreboot/keepassxc/discussions/9854 (Real Academia Española Corpus de Referencia del Español Actua (RAE, CREA))
- For French: https://github.com/kkrypt0nn/wordlists/blob/main/wordlists/languages/french.txt
- For Portuguese: https://natura.di.uminho.pt/download/sources/Dictionaries/wordlists/

A.2 Dictionaries

The list below contains links to the Terms of Service for each one of the dictionaries used in our work, as well as relevant data sources from previous work, together with the specific sections where access/distribution rights are stated.

- Oxford (Gadetsky et al., 2018)
  - https://www.oed.com/information/about-the-oed/legal-notice
- Urban Dictionary (Ni and Wang, 2017)
  - https://about.urbandictionary.com/tos/
- Cambridge (Ishiwatari et al., 2019)
  - https://www.cambridge.org/legal/copyright
- Le Petit Robert (Reid et al., 2020)
  - https://dictionnaire.lerobert.com/mentions-legales PROPRIÉTÉ INTELLECTUELLE
- Dicio (ours)
  - https://www.dicio.com.br/termos-de-uso.html
- DRAE (ours)
  - https://dle.rae.es/contenido/aviso-legal Point 4
- Duden (ours)
  - https://www.duden.de/service/agb Point 11.1
- Larousse (ours)
  - https://www.larousse.fr/infos/cgu Section PROPRIÉTÉ INTELLECTUELLE

In our paper, we utilize this data in a way that does not necessarily align with the intended use as defined by the providers. As mentioned in §3, in order to make sure we comply with the Terms of Services of our data sources, we will not distribute the definition data we collected, but only the wordlists we compiled. For the same reason, we do not run few-shot experiments with ChatGPT.

B Experimental Setup Details

Model Training We train our models with the AdamW optimizer, with an initial learning rate of 5e-5, using a batch size of 16 for mBART-50 or 8 for mT5 and for a maximum of 20 epochs. During inference, we sample from the output distribution with a maximum length of 256.

Prompting Regarding the system prompt, for LLAMA-2, we follow the approach by the original paper but replace the original one [5] 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. To generate definitions, we sample from the output distribution using *top_k=40*, *top_p=0.95*, using a repetition penalty of 1. We generate a maximum number of 256 tokens, which falls well below the length of our gold standard definitions.

Evaluation To compute the compliance of a language model, i.e. its ability to consistently generate definitions in the required language, we use

[5] https://github.com/huggingface/transformers/blob/ee2a3400f2a7038a23b83a39c5d0e24f7f699561/src/transformers/models/llama/tokenization_llama.py#L57, https://github.com/facebookresearch/llama/blob/main/UPDATES.md

fasttext-based language classification model (Grave et al., 2018). This is a linear model on top of said embeddings, which recognize 176 languages and was trained on 400 million tokens from Wikipedia, as well as sentences from the Tatoeba website. In terms of performance, this model is superior to previous work, including the readily available "langid.py" by Lui and Baldwin (2012), in the TCL, Wikipedia, and EuroGov benchmarks (Baldwin and Lui, 2010), with accuracies of 94.7, 93.0 and 98.7 respectively.

C Detailed Results

C.1 Model Training Detailed Results

In addition to experimenting with mT5, we also trained models based on mBART (Liu et al., 2020; Tang et al., 2020). The original mBART model was pre-trained in denoising full texts in multiple languages. We worked with the "large" variation (approx. 340 M parameters) pre-trained on 25 languages (mBART-CC25) (Liu et al., 2020). Since Portuguese was not included, we also use mBART-50 (mBART-50) (Tang et al., 2020), which was initialized with the former and trained to cover other 25 languages.

Table 8 summarizes the results of our monolingual experiments when training models. As we compare the performance of other languages against English, we see that our models can attain similar performance in terms of COMET, with Spanish obtaining the best results. In terms of BLEU scores, we observe an unusually high value for Portuguese, which is due to the limitations of our lemmatization-based data cleaning technique and leads to some data leakage problems. Based on our monolingual finetuning experiments, we utilize mT5 as our base model for the multilingual experiments. Table 9 summarized our results with mT5 on the validation split.

C.2 Model Prompting Details

Statistical Significance Details Table 10 below shows the details of the p-values for the experiments with LLMs.

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 the DRAE, LAROUSSE and DICIO, which we believe should serve as a reasonable estimate for the overall performance gap due to the effect of quantization. We run each experiments three times with different random seeds, and report average results, which we summarize in Table 11.

Dataset	Model	Valid BL	Valid CMT	Test BL	Test CMT
DUDEN (de)	mBART-CC25	4.06	0.338	1.732	**0.348**
	mBART-50	**6.47**	0.338	**2.477**	0.347
	mT5	4.64	**0.343**	0.939	0.346
OXFORD (en)	mBART-CC25	2.08	0.340	0.00	**0.345**
	mBART-50	**3.19**	0.340	**2.61**	0.341
	mT5	2.71	**0.341**	1.33	**0.345**
DRAE (es)	mBART-CC25	7.28	0.382	7.07	**0.385**
	mBART-50	6.11	0.378	6.10	0.383
	mT5	**9.19**	**0.379**	**7.70**	0.384
LAROUSSE (fr)	mBART-CC25	2.59	**0.295**	2.05	**0.300**
	mBART-50	3.01	**0.295**	**1.73**	0.298
	mT5	**3.68**	**0.295**	0.49	**0.300**
DICIO (pt)	mBART-50	34.02	0.316	**18.58**	0.351
	mT5	**44.77**	**0.317**	11.53	**0.352**

Table 8: Results of our monolingual experiments, finetuning multilingual language models on each of our datasets. In the table, CMT is short for COMET and BL is short for BLEU.

Hardware Our mBART-50 and mT5 finetuning experiments were performed on a large cluster, where we usually relied on a node with 4 NVIDIA V100 GPUs, or an instance with a single NVIDIA A100 GPU. For inference with LLMs and LLM finetuning, we used nodes with 1 NVIDIA H200 GPU. We spent a total of approximately 3,000 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 and to running the multilingual LLM benchmarks.

C.3 Correlation with Benchmarks

C.4 Finetuninig LLMs

D Human Evaluation Details

Table 14 shows the details of our Likert evaluation scheme, which includes example generations we used as guidance to define our score categories. Participants were verbally instructed about the task and informed that their answers would only be used for the purposes of this study, and they all agreed to participate on their on volition. They are all members of an Engineering department of a University, with ages ranging between 20 and 40

Dataset	Training Data	BLEU	COMET
DUDEN (de)	de	**4.640**	0.343
	de + en	0.497	**0.345**
OXFORD (en)	en	**2.706**	0.341
	en + fr	0.040	**0.347**
	en + de	0.775	**0.347**
	en + fr + es + pt	0.539	**0.347**
DRAE (es)	es	9.19	0.379
	es + fr	8.589	0.378
	es + pt	**9.519**	**0.380**
	es + pt + fr	9.493	0.379
	en + fr + es + pt	6.365	**0.380**
LAROUSSE (fr)	fr	3.675	0.295
	fr + es	2.603	0.295
	fr + pt	0.904	**0.299**
	fr + en	0.040	0.298
	fr + es + pt	**4.353**	0.294
	fr + en + es + pt	1.829	0.292
DICIO (pt)	pt	**44.772**	0.317
	pt + es	41.538	0.317
	pt + fr	7.464	**0.318**
	pt + fr + es	43.229	0.317
	en + fr + es + pt	12.050	0.316

Table 9: Results of our multilingual experiments based on mT5 on the validation set, where BL is short for BLEU and CMT is short for COMET.

Dataset	Model	BL	BS	CMT	Cmp.
DUDEN (de)	LLAMA-2	0.003	0.000	0.002	0.382
	MISTRAL	0.420	0.212	0.004	0.372
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
LAROUSSE (fr)	LLAMA-2	0.001	0.000	0.263	0.970
	MISTRAL	0.003	0.000	0.040	0.775
DICIO (pt)	LLAMA-2	0.185	0.602	0.597	0.295
	MISTRAL	0.027	0.012	0.035	0.000

Table 10: 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
LAROUSSE (fr)	LLAMA-2	4bit	0.326	0.473
		full	0.380	0.483
	MISTRAL	4bit	0.327	0.441
		full	0.432	0.463
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 11: Results of our study on the impact of low precision (4 bit). In the table, BL and CMT are short for BLEU and COMET, respectively.

years old.

As mentioned above, in order to further understand the effect of in-context learning in our task, and the effectiveness of COMET as a metric, we perform an additional experiment by using this metric to select the 50 best and 50 worst generations, again resulting in a total of 400 definitions to be evaluated per annotator. We choose the same random seed for all models. We exclude English from this study as this language has been studied before for our task, and because our initial evaluation shows that performance substantially superior.

Table 15 summarizes the results of this second experiment. We first notice that the effects of in context learning are again not consistent across models and languages. We often find that results for "Worse" sets contain rare terms, which we assume should also have relative lower frequency in the training data. Given the nature of our task, we think this shows that although few-shot learning may help the model achieve a more dictionary-like style when generating, it does not elicit better task-solving capabilities for those cases.

We also see observe that Likert scores on the "Best" sets are higher than on the "Worst" sets, which suggests that COMET is indeed able to capture the semantic quality of the generated definitions. However, from a qualitative point of few, feedback from our annotators suggest that while LLAMA-2 generations were more precise and correct, MISTRAL generations were rather "Very good", with rich details, or "Very Poor", containing wrong information; however, COMET could not detect such differences. These findings are in agreement with previous observations by Jhirad et al. (2023), and provide additional evidence to show that the performance of COMET when capturing semantic similarity decreases substantially with length of the inputs.

E Performance on multilingual benchmarks

Table 16 summarizes the result we obtained when running our chosen LLMs on the multilingual benchmarks HellaSwag, MMLU and ARC-Challenge. Following (Lai et al., 2023), experiments on HellaSwag are zero-shot, and we use 25 shots for MMLU and ARC in languages other than

Model	Score	BL	BS	CMT
LLAMA-2	HellaSwag	0.49	0.95	0.74
	MMLU	0.55	0.97	0.72
	ARC-Challenge	0.61	0.88	0.87
MISTRAL	HellaSwag	0.13	0.95	0.73
	MMLU	0.04	0.91	0.63
	ARC-Challenge	0.32	0.99	0.84
LLAMA-3	HellaSwag	-0.44	0.93	-0.18
	MMLU	-0.24	0.97	-0.10
	ARC-Challenge	-0.83	0.71	-0.25

Table 12: Pearson correlation between performance in our task and multilingual benchmarks for LLMs.

Lang.	Model	BL	BS	CMT
DUDEN (de)	LLAMA-3 (ours)	3.261	0.676	0.580
	+ ft (ours)	0.794	0.585	0.411
	+ ft (Periti et al.)	0.147	0.574	0.360
DRAE (es)	LLAMA-3 (ours)	1.855	0.670	0.545
	+ ft (ours)	0.634	0.584	0.460
	+ ft (Periti et al.)	0.186	0.570	0.379
OXFORD (en)	LLAMA-3 (ours)	1.588	0.568	0.835
	+ ft (ours)	0.568	0.511	0.806
	+ ft (Periti et al.)	0.867	0.455	0.799
LAROUSSE (fr)	LLAMA-3 (ours)	3.443	0.676	0.543
	+ ft (ours)	0.809	0.608	0.496
	+ ft (Periti et al.)	0.266	0.595	0.388
DICIO (pt)	LLAMA-3 (ours)	0.457	0.683	0.542
	+ ft (ours)	1.648	0.598	0.467
	+ ft (Periti et al.)	0.219	0.592	0.393

Table 13: Summary of the performance of our LLAMA-3 model finetuned on monosemic multilingual dictionary data on our datasets, compared against LLamaDict (Periti et al., 2024) [Highlight best values –EMT].

English.

F Tackling Polysemy

Our main results are limited to the set of terms for which we only have one definition. Though our results suggest that LLMs can, to some extent, generate plausible definitions for words, our empirical study offers no insight into the more challenging scenario of polysemy.

Given the success of our studied LLMs in generating definitions for our selected languages, we propose to utilize LLMs to obtain example sentences for polysemic terms in our data. We focus on Portuguese given the annotators' availability.

We run a pilot study with to generate such example sentences. Concretely, we prompt the model as follows: *In the Portuguese Language, the word '{term}' can mean "{definition}". Please give me a sentence in that language where this word is used.*, where {term} and {definition} are placeholders for variables denoting a given pair of a term and its corresponding definition.

For our study, we sample a subset of 59 terms from our data, with a total of 156 different meanings. We utilize the official OpenAI API to feed this data into the model. To assess the viability of this approach, we subject these sentences to a thorough human evaluation. Specifically, we ask a native Brazilian Portuguese speaker to evaluate the generated sentences according to the following simplified Likert scale: 1 - Poor: The sentence uses the term with a different meaning; 2 - Acceptable: The sentence uses the term with the specified meaning, but in the wrong context/sounds unnatural; 3 - Good: The sentence uses the term with the specified meaning.

Our results showed that 75.64% of the examples we obtained were labeled as "Good" by our human evaluator, while only 7.69% and 16.67% were labeled as "Poor" and "Acceptable", respectively.

The results of this evaluation are in Table 18. We can see that most of the generated examples received the "Good" score, with an average of 2.68, which indicates that using this specific LLM could be an alternative to obtain such example sentences for Brazilian Portuguese.

Category	Details	Example
Very poor	A completely wrong definition, one that contains only the word, or its English translation.	**Waterspout (*Tromba d'agua*)** 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.	**To sharpen (*Nitidificar*)** 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	**Supralegal (*Supralegal*)** 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)	**To oversize (*Superdimensionar*)** 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.	**To sanitize (*Sanitizar*)** 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 14: Explanation of our human evaluation system for generated definitions. Examples are taken from DICIO and manually translated to English by the annotator.

Model	Likert (b/w)	COMET (b/w)
DUDEN (de)		
LLAMA-2	3.060 / 1.100	0.670 / 0.241
+ 5 shots	1.400 / 1.040	0.767 / 0.194
MISTRAL	2.560 / 1.000	0.699 / 0.211
+ 5 shots	1.920 / 1.000	0.700 / 0.197
DRAE (es)		
LLAMA-2	4.680 / 1.100	0.700 / 0.317
+ 5 shots	4.900 / 1.100	0.751 / 0.235
MISTRAL	4.600 / 1.080	0.727 / 0.258
+ 5 shots	4.440 / 1.140	0.710 / 0.236
LAROUSSE (fr)		
LLAMA-2	4.280 / 1.122	0.697 / 0.281
+ 5 shots	4.460 / 1.260	0.736 / 0.239
MISTRAL	4.061 / 1.000	0.705 / 0.233
+ 5 shots	4.040 / 1.000	0.689 / 0.228
DICIO (pt)		
LLAMA-2	3.860 / 1.240	0.639 / 0.385
+ 5 shots	4.000 / 1.460	0.623 / 0.339
MISTRAL	3.260 / 1.140	0.577 / 0.320
+ 5 shots	1.780 / 1.180	0.530 / 0.306

Table 15: Results of our second human evaluation experiment in terms of average Likert scores for the best 50 / worst 50 denoted as (b/w); and compared to COMET.

Language	Model	HellaSwag	MMLU	ARC
English (en)	LLAMA-2	0.607	0.533	0.462
	MISTRAL	0.563	0.534	0.501
	LLAMA-3	0.591	0.681	0.515
French (fr)	LLAMA-2	0.460	0.439	0.418
	MISTRAL	0.426	0.413	0.387
	LLAMA-3	0.488	0.544	0.467
German (de)	LLAMA-2	0.431	0.431	0.408
	MISTRAL	0.395	0.401	0.358
	LLAMA-3	0.468	0.529	0.442
Portuguese (pt)	LLAMA-2	0.437	0.434	0.427
	MISTRAL	0.415	0.396	0.402
	LLAMA-3	0.487	0.541	0.487
Spanish (es)	LLAMA-2	0.474	0.441	0.441
	MISTRAL	0.433	0.414	0.393
	LLAMA-3	0.506	0.548	0.492

Table 16: Observed performance of LLAMA-2 and MISTRAL on our three selected multilingual benchmarks, separated by language.

Lang.	Existing Benchmark			Ours	
	MMLU	ARC	HellaSwag	Input	Exp. Output
fr	53,431,809	1,489,775	735,614	60,629	105,753
es	54,083,040	1,460,328	707,524	62,971	64,925
pt	54,044,351	1,459,704	707,201	231,425	285,676
de	54,805,912	1,459,108	735,190	139,289	108,314
Total	216,365,112	5,868,915	2,885,529	494,314	564,668

Table 17: Length of input (and expected output, where available), in number of tokens (via the LLAMA-2 tokenizer).

Poor	Acceptable	Good
7.69	16.67	75.64

Table 18: Results of the human evaluation of the generated example sentences. The numbers indicate the percentage of sentences evaluated in each category