Evaluating Large Language Models on Multiword Expressions in Multilingual and Code-Switched Contexts

Frances A. Laureano De Leon[1] and **Asim Abbas**[1] and **Harish Tayyar Madabushi**[2] and **Mark Lee**[1]
[1]University of Birmingham, School of Computer Science, Edgbaston, Birmingham B15 2TT
[2]University of Bath, Department of Computer Science, Bath BA2 7AY, UK
{fxl846, axa2233, m.g.lee}@bham.ac.uk, htm43@bath.ac.uk

Abstract

Multiword expressions, characterised by non-compositional meanings and syntactic irregularities, are an example of nuanced language. These expressions can be used literally or idiomatically, leading to significant changes in meaning. Although large language models perform well on many tasks, their ability to handle subtle linguistic phenomena remains unclear. This study examines how state-of-the-art models process the ambiguity of potentially idiomatic multiword expressions, particularly in less frequent contexts where memorisation is less likely to help. By evaluating models in Portuguese, Galician, and English, and introducing a new code-switched dataset and task, we show that large language models, despite their strengths, have difficulty handling nuanced language. In particular, we find that the latest models, including GPT-4, fail to outperform the xlm-roBERTa-base baselines in both detection and semantic tasks, with especially poor performance on the novel tasks we introduce, despite its similarity to existing tasks. Overall, our results demonstrate that multiword expressions, especially those that are ambiguous, continue to be a challenge to models. We provide open access to our datasets and prompts, https://github.com/francesita/CS-and-Multilingual-MWEs

1 Introduction

Multiword expressions (MWEs) present significant challenges in Natural Language Processing (NLP) due to their linguistic complexities. MWEs consist of multiple words that convey a specific meaning, which may be non-compositional and exhibit syntactic irregularities (Baldwin and Kim, 2010; Constant et al., 2017). These expressions are widespread across languages and domains, from the arts to the sciences (Villavicencio and Idiart, 2019). The number of MWEs in a speaker's lexicon is estimated to be comparable to that of single words (Sag et al., 2002). MWEs are often rooted in cultural context and communication rather than a straightforward lexical meaning, making their interpretation in NLP systems more challenging, particularly in idiomatic contexts (Masini, 2019). Additionally, MWEs can be ambiguous, functioning either idiomatically or literally, which can cause potential confusion (Kurfal and Östling, 2020). For example, the English MWE *break the ice* can refer to its literal meaning or describe making unfamiliar individuals feel more at ease with one another. Such expressions, known as potentially idiomatic expressions (PIEs) (Haagsma et al., 2020), require an understanding of both literal and figurative meanings. Given the idiosyncratic nature of MWEs, it is essential that state-of-the-art NLP models accurately capture their meanings.

Large language models (LLMs) have rapidly become the dominant approach in NLP, demonstrating strong language understanding and generation abilities across a range of tasks in zero-shot and few-shot settings (Liu et al., 2022). However, despite their performance, LLMs exhibit clear limitations, including hallucinations, fluent but inaccurate or fabricated content, and a reliance on statistical co-occurrence patterns from pre-training data (Razeghi et al., 2022; Kang and Choi, 2023). *This study investigates whether LLMs can truly understand nuanced language, using MWEs as a test case.* MWEs are well-suited for this purpose, as they can be interpreted either idiomatically or literally, requiring sensitivity to both social and linguistic context. As MWEs occur across languages and vary in frequency and usage, their interpretation may be even more difficult in multilingual settings. In particular, idiomatic MWEs present a distinctive challenge to LLMs: their meanings are often non-compositional and cannot be reliably inferred from the meanings of their individual words. Unlike named entities (e.g., New York City) or compound nouns (e.g., credit card), idiomatic ex-

Proceedings of Recent Advances in Natural Language Processing,pages 644–653
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-075

pressions such as *kick the bucket* or *spill the beans* require models to distinguish between literal and figurative meanings based on context. This ambiguity makes idiomatic MWEs a particularly effective probe for evaluating whether LLMs understand context, nuance, and figurative language, abilities typically associated with human-like language understanding. Since LLMs rely heavily on surface-level patterns, idiomatic MWEs offer insight into whether their apparent success reflects genuine semantic generalisation or simply statistical association. One such challenge arises in code-switching (CS), where speakers alternate between languages within a conversation or sentence (Joshi, 1982; Dogruoz et al., 2021). Code-switching is common in multilingual communities and presents additional difficulties for NLP systems, as MWEs may be split across languages. To evaluate how LLMs handle this complexity, we use data in English, Portuguese, and Galician, as well as CS examples. By including both high-resource and low-resource languages alongside CS data, we aim to assess whether models rely primarily on statistical patterns in training data or if they can generalise effectively across multilingual and code-switched contexts.

We conduct experiments using both open-source and closed-source LLMs to examine how these models handle MWEs across different languages and contexts. Specifically, we investigate:

- the ability of LLMs to detect whether MWEs are used as idioms or literally,
- the extent to which LLMs effectively capture the non-compositional meanings of MWEs,
- the ability of LLMs to acquire information from the prompt, and
- the relationship between LLM size and performance on text containing MWEs with non-compositional meaning.

Our experiments show that, despite strong performance on many tasks, LLMs continue to struggle with MWEs, particularly in multilingual contexts, while smaller fine-tuned models like XLM-RoBERTa perform more reliably. Although results for English MWEs are sometimes comparable, performance declines in other languages, highlighting limitations in handling nuanced language. Section 2 reviews related work, Section 3 outlines the datasets and synthetic CS data, Section 4 describes the setup, and Section 5 presents our findings. Section 6 concludes with a summary and future directions.

1.1 Contributions

We conduct experiments to evaluate generative models' ability to detect and interpret MWEs in both idiomatic and compositional contexts. MWEs are selected for their lower frequency, reducing the likelihood of memorisation. To further explore this, we introduce a new code-switched test set covering Galician–Spanish, English–Spanish, and Portuguese–Spanish, which we release publicly.

Models are evaluated on standard MWE detection (Section 4.3) and new tasks assessing semantic understanding (Section 4.5), using both synthetic MWEs (Section 4.4) and the CS data (Section 3.1). These tasks are designed to minimise memorisation and assess how well models interpret non-compositional meaning. Our results show that LLMs struggle with MWE semantics and suggest that prompt-based learning alone is insufficient for handling unseen vocabulary in challenging linguistic settings.

2 Related work

Studies have shown that popular transformer models generally struggle to effectively handle figurative language. Research on encoder-only models shows that PLMs do not adequately capture and represent the meanings of idiomatic expressions (Garcia et al., 2021). Additionally, generative models like GPT-2 have been found to perform poorly in handling figurative language without the use of mitigation strategies (Jhamtani et al., 2021). Miletić and Walde (2024) reveal that transformer models capture MWE semantics inconsistently, and find that they are reliant on surface patterns. However, research has shown that encoder-only PLMs can effectively represent MWEs when fine-tuned with appropriate data (Tayyar Madabushi et al., 2022). Studies exploring how generative models handle figurative language are few, especially studies focusing on idioms. Liu et al. (2022) test BERT, and RoBERTa, as well as generative models, namely, GPT-2, GPT-3 and GPT-NEO on their ability to reason about figurative language, with a focus on metaphors. They find that all models need to be fine-tuned to do well on interpreting figurative language, and that the capabilities of these models remain far from reaching human performance. De Luca Fornaciari et al. (2024) conduct experiments on an English language dataset they create to assess model abilities in detecting idiomatic expressions in a zero-shot setting. They

conduct their experiments on the 7B versions of Llama-2, Vicuna, and Mistral models. Phelps et al. (2024) conduct detection experiments in English, Portuguese and Galician on different state-of-the-art generative models and find that although they give competitive results, they fail to reach the performance of fine-tuned PLMs. In this work, we will also explore model abilities in detecting MWEs in English, Portuguese, and Galician. However, our work also examines how generative models represent the *meaning* of sentences where MWEs are used idiomatically in these languages. Additionally, we investigate the impact of CS text on model performance and assess the models' ability to manage unseen MWEs using synthetic examples.

3 Datasets

To assess the capabilities of LLMs in understanding nuanced language, we use the SemEval 2022 Task 2 dataset (Tayyar Madabushi et al., 2022) for all experiments. This dataset contains MWEs in both literal and idiomatic contexts in English, Portuguese, and Galician, and includes adversarial examples to test model consistency in interpreting idiomatic meaning. The SemEval task comprises two subtasks: *subtask a*, an MWE interpretation task, and *subtask b*, an MWE paraphrase similarity task adapted from a semantic text similarity (STS) task for use with generative models. Both tasks are designed to assess whether models can capture the meaning of sentences containing MWEs, whether used compositionally or not. We adapt these subtasks and create new tasks based on the same data to test whether familiarity with task format (likely present in instruction tuning) affects performance. This also motivates our use of variations on a single dataset rather than multiple datasets. A further advantage of this dataset is its multilingual nature and the fact that the test labels have not been released, reducing the chance that models have seen the answers during instruction tuning or pre-training.

3.1 Code-Switched Dataset

To evaluate model performance on MWEs in mixed-language settings, we create a synthetic CS dataset by combining Spanish with monolingual English, Portuguese, and Galician sentences. CS is common in multilingual communities and contributes to varieties such as Spanglish.

We generate the data using OpenAI's GPT-4-turbo-0125 model with a temperature of 0.65 and a seed of 42, balancing coherence and variation. Two base prompts guide generation: one for MWE-containing examples and another for adversarial, non-MWE paraphrases. Spanish is mixed separately with each language to reduce output errors. Prompts are available on the project GitHub [1].

This dataset tests how well LLMs handle nuanced language in CS contexts, an area often overlooked in monolingual evaluations. While synthetic data may lack full naturalness, it remains coherent and helps reduce models' reliance on surface-level co-occurrence patterns. It also offers a practical alternative to real CS data, which is scarce and costly to collect.

4 Experimental Setup

In this section, we introduce our experiments and the models used as part of this work. We run three different experiments: an **MWE interpretation task**, a **synthetic MWE interpretation task**, and **MWE paraphrase similarity task**. These experiments aim to determine whether the models can recognise the presence and meaning of text that contains nuanced language. Most of the experiments are conducted in a zero-shot and few-shot setting.

4.1 Models

We use four models in our experiments, including both open-source and proprietary systems: (1) gpt-3.5-turbo-0125, (2) gpt-4-0125-preview, (3) meta-llama-3-70b-instruct, and (4) meta-llama-3-8b-instruct. The GPT models are accessed via the OpenAI API[2], while the Llama models are run using the Replicate API[3]. All models are tested using the same settings: temperature set to 0 and seed value of 42. We specify a maximum token count of 7 for the Llama models and 5 for the GPT models. As a baseline, we use XLM-RoBERTa (Conneau et al., 2020), selected as the sole pre-trained language model in our evaluation due to its strong performance in SemEval-2022 Task 2 on idiomaticity detection (Tayyar Madabushi et al., 2022).

4.2 Prompts

For all experiments in this section, we use three different prompts to generate outputs from the LLMs. The prompts are based on strategies outlined on

[1] https://github.com/francesita/CS-and-Multilingual-MWEs

[2] https://openai.com/

[3] https://replicate.com/

the OpenAI website [4]. Specifically, we employ three approaches: the 'persona' prompt, where the model adopts the role of a linguist; the 'think' prompt, where the model is instructed to reflect on the meaning of a sentence before assigning a label; the 'ChatGPT' prompt, generated using ChatGPT-turbo-3.5. Each prompt is adapted according to the specific task. We opt to use three prompts because model results may vary given different inputs (Liu et al., 2024), and this allows to explore possible variations in the model outputs given variations in the input. The prompts are available on the project GitHub [5].

4.3 MWE Interpretation Task

We conduct MWE sense disambiguation experiments, in which we ask an LLM to determine whether a given MWE in a sentence is used idiomatically or literally. We carry out these experiments in both zero-shot and few-shot settings using data from *subtask a* of the SemEval competition, which includes a total of 2,342 examples: 916 in English, 713 in Portuguese, and 713 in Galician. For the few-shot experiments, the model is provided with five examples: three in English and two in Portuguese. At least one instance of an MWE being used literally and idiomatically is included for each of the above languages. The original test dataset for SemEval *subtask a*, along with the synthetic CS dataset, is used for this task. The languages of the few-shot examples is not altered for the CS experiments, that is to say, the examples provided to the LLMs is just in English and Portuguese. The model receives input which consists of text containing an MWE, along with the MWE itself. We prompt the model to generate a label, either 'idiom' or 'literal', in response to the question, 'Is the multiword expression [MWE] an idiom or literal in this sentence?'. The three base prompts previously mentioned are used to generate answers from each model. The results of these experiments are presented in Table 1 and Table 2.

4.4 Synthetic MWE Interpretation Task

The objective of the synthetic MWE experiments is to assess whether providing additional information in the prompt can improve model performance.

[4] https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results

[5] https://github.com/francesita/CS-and-Multilingual-MWEs

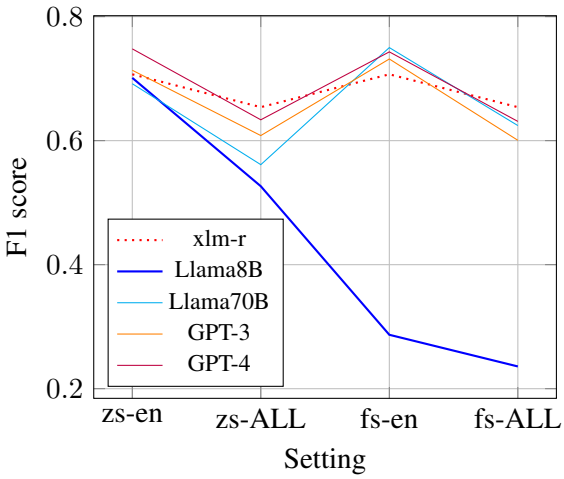

Figure 1: F1 scores for each model using the original SemEval task 2 dataset. This graph shows the F1 score for the English language subset of the dataset, as well as the total macro F1 score for all languages combined in the zero-shot (zs) and few-shot settings (fs).

We specifically aim to test whether models can understand nuanced language when given more context. This approach is based on (Eisenschlos et al., 2023), where synthetic words were introduced and defined as existing concepts to evaluate an LLM's ability to learn new vocabulary through prompting. It is likely that an LLM has not encountered all the MWEs in our datasets, particularly those in Portuguese and Galician. Therefore, it is important to evaluate whether a model can acquire knowledge of new words or MWEs through prompt-based learning. To investigate this, we use the development set of subtask a, which includes examples in English and Portuguese (but not in Galician), and replace the MWEs in each example with synthetic ones. The model is prompted using the same three base prompts from the MWE interpretation experiments, with an added definition for the synthetic MWE, taken from the original MWE it replaced. The results of these experiments are shown in Table 5.

4.5 MWE Paraphrase Similarity Task

As part of this work, we aim to evaluate the ability of LLMs to understand the meaning of text containing MWEs. We provide each model with sentence pairs: some include an MWE, while others contain correct or incorrect paraphrases of MWEs. To run these experiments, we adapt *subtask b* from the SemEval Task 2 competition, which was originally designed as a STS task. We construct three distinct prompts, as described in Section 4.2, each

Model	F-1 Score EN	F-1 Score PT	F-1 Score GL	F-1 Score ALL	Code-switched
GPT-4	**0.7480** $\pm$ 0.23	0.5701 $\pm$ 0.005	**0.5395** $\pm$ 0.012	0.6336 $\pm$ 0.013	No
GPT-3	0.7135 $\pm$ 0.030	0.5359 $\pm$ 0.015	0.5243 $\pm$ 0.101	0.6082 $\pm$ 0.053	
Llama 3-70B	0.6918 $\pm$ 0.025	0.5097 $\pm$ 0.32	0.4420 $\pm$ 0.044	0.5613 $\pm$ 0.034	
Llama 3-8B	0.7013 $\pm$ 0.022	0.4628 $\pm$ 0.047	0.3636 $\pm$ 0.061	0.5265 $\pm$ 0.032	
xlm-roBERTta	0.7070	**0.6803**	0.5065	**0.6540**	
GPT-4	**0.7595** $\pm$ 0.025	0.5759 $\pm$ 0.002	0.5411 $\pm$ 0.007	0.6398 $\pm$ 0.012	Yes
GPT-3	0.7191 $\pm$ 0.021	0.5417 $\pm$ 0.012	**0.5575** $\pm$ 0.063	0.6246 $\pm$ 0.030	
Llama 3-70B	0.7114 $\pm$ 0.020	0.5496 $\pm$ 0.033	0.4726 $\pm$ 0.047	0.5915 $\pm$ 0.034	
Llama 3-8B	0.7072 $\pm$ 0.024	0.5276 $\pm$ 0.051	0.4076 $\pm$ 0.040	0.5619 $\pm$ 0.034	

Table 1: These are the scores for the detection experiments in the **zero-shot** setting. Reported F-1 scores are for each of the individual languages tested, as well as the combined Macro-F-1 score (ALL).

Model	F-1 Score EN	F-1 Score PT	F-1 Score GL	F-1 Score ALL	Code-switched
GPT-4	0.7432 $\pm$ 0.006	0.5705 $\pm$ 0.008	**0.5356** $\pm$ 0.008	0.6310 $\pm$ 0.003	No
GPT-3	0.7317 $\pm$ 0.015	0.5455 $\pm$ 0.003	0.4755 $\pm$ 0.092	0.6006 $\pm$ 0.034	
Llama 3-70B	**0.7501** $\pm$ 0.017	0.5638 $\pm$ 0.007	0.5161 $\pm$ 0.016	0.6243 $\pm$ 0.014	
Llama 3-8B	0.2869 $\pm$ 0.397	0.1864 $\pm$ 0.305	0.1996 $\pm$ 0.341	0.2359 $\pm$ 0.360	
xlm-roBERTta	0.7070	**0.6803**	0.5065	**0.6540**	
GPT-4	0.7433 $\pm$ 0.014	0.5831 $\pm$ 0.012	**0.5342** $\pm$ 0.002	0.6341 $\pm$ 0.009	Yes
GPT-3	0.7308 $\pm$ 0.009	0.5433 $\pm$ 0.009	0.4988 $\pm$ 0.064	0.6074 $\pm$ 0.024	
Llama 3-70B	**0.7549** $\pm$ 0.011	0.5666 $\pm$ 0.007	0.53 $\pm$ 0.014	0.6321 $\pm$ 0.008	
Llama 3-8B	0.2653 $\pm$ 0.391	0.1737 $\pm$ 0.261	0.2107 $\pm$ 0.327	0.2291 $\pm$ 0.346	

Table 2: These are the scores for the detection experiments in the **few-shot** setting. Reported F-1 scores are for each of the individual languages tested, as well as the combined Macro-F-1 score (ALL).

asking the model if the two sentences have similar meanings. Each prompt notes that some sentences may include MWEs used idiomatically. The model is instructed to output 'true' if the sentences are similar in meaning and 'false' if they differ, turning the task into a binary classification problem. These experiments are conducted in both zero-shot and few-shot settings. We use the development split of *subtask b* from the original SemEval competition, which contains examples in English and Portuguese only. This dataset includes sentence pairs with MWEs and their paraphrases, as well as examples from a standard STS benchmark. The gold labels in the dataset are either a Spearman rank correlation score (for STS benchmark examples), a score of 1 (for identical meanings), or a label of NONE (for differing meanings). After adapting the task and removing examples with similarity scores other than 1 or NONE, we retained 974 examples: 521 in English and 454 in Portuguese. Since the task requires binary outputs, we exclude all examples with a Spearman score other than 1 and use only examples labelled as 1 (similar) or NONE (not similar). This required access to the gold labels, and as a result, we were unable to use the original competition evaluator, which assumes unmodified scoring. In addition to testing on the original development data, we run experiments on our generated CS examples, where each language is mixed with Spanish. The results for these experiments are shown in Table 3 and Table 4. We also experiment with the same dataset used for the semantic task. We substitute the real MWEs in each example with a synthetic MWE and supply the model with the original MWE's definition as the meaning of the synthetic one. The results of these experiments are on Table 6.

5 Results and Discussion

Are LLMs effective in interpreting idiomatic MWEs? In general, LLM abilities in detecting idiomatic expressions fall short of smaller fine-tuned PLMs. This is especially true for Galician and Portuguese experiments using the original dataset. Given that most of these models are predominantly trained on English data, it is unsurprising that they

Model	F-1 Score EN	F-1 Score PT	F-1 Score ALL	Code-switched
GPT-4	0.7178 ± 0.040	0.6142 ± 0.019	0.6684 ± 0.029	No
GPT-3	0.6396 ± 0.044	0.6574 ± 0.035	0.6490 ± 0.039	
Llama 3-70B	0.6802 ± 0.071	0.5065 ± 0.050	0.5969 ± 0.061	
Llama 3-8B	0.6142 ± 0.108	0.4872 ± 0.075	0.5538 ± 0.093	
xlm-roBERTta	**0.7590**	**0.7658**	**0.7712**	
GPT-4	0.6201 ± 0.023	0.5538 ± 0.003	0.5889 ± 0.012	Yes
GPT-3	0.5373 ± 0.072	0.5690 ± 0.016	0.5525 ± 0.039	
Llama 3-70B	0.5012 ± 0.060	0.4189 ± 0.032	0.4618 ± 0.045	
Llama 3-8B	0.5541 ± 0.019	0.4401 ± 0.018	0.4993 ± 0.018	
xlm-roBERTta	**0.6752**	**0.7180**	**0.7020**	

Table 3: Results for semantic experiments in the **zero-shot** setting.

Model	F-1 Score EN	F-1 Score PT	F-1 Score ALL	Code-switched
GPT-4	**0.7628** ± 0.030	0.6304 ± 0.013	0.7000 ± 0.021	No
GPT-3	0.6497 ± 0.015	0.6296 ± 0.109	0.6404 ± 0.060	
Llama 3-70B	0.7204 ± 0.006	0.5570 ± 0.006	0.6425 ± 0.006	
Llama 3-8B	0.7045 ± 0.009	0.5771 ± 0.010	0.6440 ± 0.008	
xlm-roBERTta	0.7590	**0.7658**	**0.7712**	
GPT-4	0.5455 ± 0.008	0.5209 ± 0.022	0.5362 ± 0.014	Yes
GPT-3	0.5634 ± 0.051	0.5841 ± 0.011	0.5732 ± 0.026	
Llama 3-70B	0.5670 ± 0.010	0.4434 ± 0.004	0.5071 ± 0.007	
Llama 3-8B	0.6172 ± 0.011	0.5073 ± 0.030	0.5651 ± 0.020	
xlm-roBERTta	**0.6752**	**0.7180**	**0.7020**	

Table 4: Results for semantic experiments in a **few-shot** setting.

Model	F-1 Score EN	F-1 Score PT	F-1 Score ALL
GPT-4	**0.6552** ± 0.031	0.5955 ± 0.031	**0.6392** ± 0.025
GPT-3	0.6231 ± 0.061	**0.6056** ± 0.041	0.6225 ± 0.050
Llama 3-70B	0.5514 ± 0.037	0.5600 ± 0.018	0.4289 ± 0.030
Llama 3-8B	0.5696 ± 0.076	0.5677 ± 0.077	0.5689 ± 0.026
Random baseline	0.5441	0.5170	0.5387

Table 5: Experiment results for *detection experiments* in the **zero-shot** setting using **synthetic MWEs.**

Model	F-1 Score EN	F-1 Score PT	F-1 Score ALL
GPT-4	0.5454 ± 0.016	0.4855 ± 0.025	0.5183 ± 0.017
GPT-3	**0.5655** ± 0.004	**0.5204** ± 0.071	**0.5444** ± 0.036
Llama 3-70B	0.4655 ± 0.028	0.3885 ± 0.030	0.4289 ± 0.030
Llama 3-8B	0.5455 ± 0.029	0.4711 ± 0.090	0.5108 ± 0.060
Random baseline	0.5092	0.4985	0.5025

Table 6: Experiment results for *semantic experiments* in the **zero-shot** setting using **synthetic MWEs.**

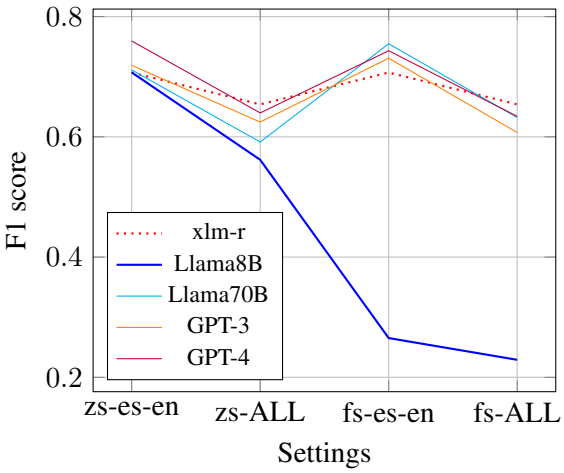

Figure 2: F1 scores for each model using the CS dataset. This graph shows the F1 score for the Spanish-English (es-en) language subset of the dataset, as well as the total macro F1 score for all CS examples combined in the zero-shot (zs) and few-shot (fs) settings.

perform best on the English portion of the dataset, with most models reaching an F1 score of 0.70+ in the zero-shot and few-shot settings. Interestingly, the overall macro F1 score shows a slight improvement for all models in the CS experiments. This improvement is likely due to the inclusion of Spanish in all examples, a language that is likely more prevalent in the pre-training data compared to Galician and Portuguese. Some models exceed xlm-roBERTa's F1 score for the English partition of the data, 0.7070, but none are able to reach a score higher than 0.76. However, all models fall short of the baseline for Portuguese, 0.6803, and the overall F1 score, 0.6540, when compared to a fine-tuned xlm-RoBERTa-base model in both zero-shot and few-shot settings. Notably, the Llama3-8B-instruct model failed to produce valid outputs in the few-shot setting for both the original and CS datasets. It seems that the combination of few-shot examples along with the input text confounded the model, to the degree that most of the model outputs were not a single word answer containing the label. Figures 1 and 2 demonstrate how Llama3-8B struggled in few-shot experiments, evidenced by the high variance in the scores in the three prompts used to generate model output. Although LLMs have been competitive in detecting nuanced language when compared to a much smaller PLM, in general, they struggle in the other lower-resourced languages.

To what extent do LLMs capture the non-compositional meanings of MWEs? LLMs seem to struggle to capture the meaning of text containing MWEs. In general, unlike in the MWE interpretation task, the few-shot examples here helped most models reach higher F1 scores for both Portuguese and English text. GPT-4 was able to perform comparably to the xlm-roBERTa baseline for English in the few-shot setting, with a score of 0.7638, marginally beating xlm-roBERTa's score of 0.7590. All models fall short of the combined baseline score and F1 score in Portuguese in zero-shot and few-shot settings. This may indicate that models are not capturing the true meaning of text containing MWEs and may be relying on spurious correlations to make decisions. Overall, LLMs appear to struggle with nuanced language, as they often fail to outperform a much smaller fine-tuned PLM.

What is the impact of providing information in the prompt? The results indicate LLMs generally perform above random on synthetic MWE interpretation task, were we provided the definition of the MWE in the prompt, with GPT-4 achieving the highest combined F1 score of 0.6392. However, their performance is notably lower compared to when real MWEs are used. This suggests that models may have memorised some real MWEs during training. For the *semantic experiments* involving synthetic MWEs, model performance remains close to the random baseline. Across all languages, GPT-3 achieves the highest scores, 0.5444, but these results still do not significantly exceed the random baseline of 0.5025. This demonstrates that providing additional information in the prompt does not improve model performance when the vocabulary differs from patterns encountered during pre-training. The models appear to rely heavily on statistical patterns from pre-training data, and when faced with unfamiliar patterns, they fail to generate correct responses, even with contextual cues. Overall, the models have difficulty applying the meanings of synthetic MWEs to complete the task, suggesting that additional context does not help them address complex linguistic issues or unfamiliar vocabulary.

Is there a relationship between LLM size and performance on text containing non-compositional phrases? The smaller Llama 8B model performed unexpectedly well on the zero-

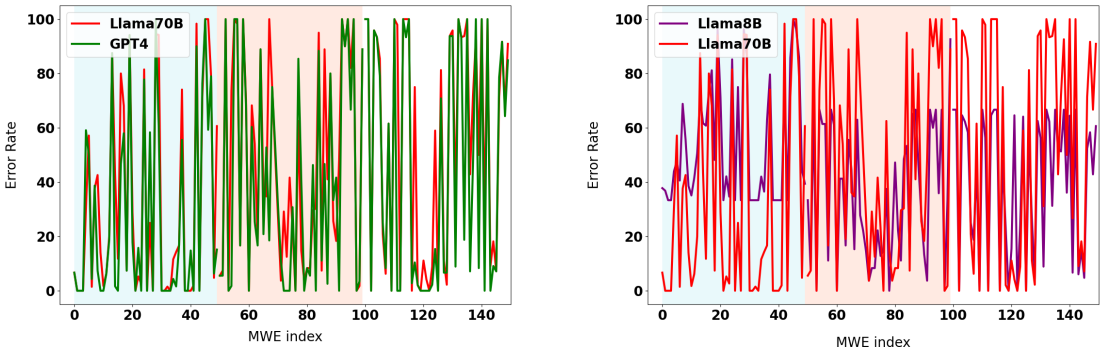

Figure 3: The error rates of all MWEs in the test set. These are in alphabetical order per language. The MWEs in English are in the light blue background, Portuguese are in the light orange background and Galician are in a white background. The figure on the left show the errors for LLama 70B and GPT-4 and the figure on the right shows the errors for the Llama 70B and Llama 8B.

shot detection task, achieving results comparable to Llama 70B and GPT-3 for English text. However, it showed a slight decline in performance for Portuguese and Galician. Notably, the Llama 8B model was unable to process both the original text as well as the CS text in the few-shot detection task and was more sensitive to prompting than the other models, evidenced by the high standard deviation seen in Table 2. GPT-4 outperformed other LLMs in English for the semantic task and was the only LLM capable of handling English-Spanish CS text in this context. While further experiments are necessary, these findings suggest that a smaller model, like Llama 8B, may suffice to achieve results comparable to much larger models for MWE detection in English. However, the results also show that GPT-4 is better than other models when confronted with unknown vocabulary and representing the semantics of MWEs compared to the other LLMs.

5.1 Error Analysis

Certain MWEs were consistently classified correctly by all models when using the original dataset for the MWE interpretation task, as detailed in the project GitHub. Figures 1 and 2 illustrate the sensitivity of models to different prompts. Notably, the standard deviations, especially for Llama 8B, show how different some model outputs may be depending on the prompt. A subset of MWEs was correctly classified by all models in the detection task, which are also available in the project GitHub. It lists correctly identified MWEs by all models, those correctly classified by specific model families, and the overlap between the two largest models. We also examined correlations between the error rates of individual MWEs across models. There is a moderate correlation of 0.64 between the error rates of Llama models, and a strong correlation of 0.70 between Llama 70B and GPT-4. However, the correlation between the two GPT models is weaker, at 0.32, possibly indicating differences in their training data. Figure 3 shows the errors between Llama 70B and GPT-4 as well as the two Llama models. The curves are overlaid for comparison. The strong correlation in errors between certain models may indicate shared limitations in tasks requiring nuanced language understanding, potentially stemming from their architectural design or training data.

6 Conclusion

This study examined whether state-of-the-art LLMs can detect and represent nuanced language, focusing on MWEs in English, Portuguese, Galician, and code-switched text. Despite their scale, LLMs struggle with MWEs—especially in non-English contexts—and often underperform compared to smaller fine-tuned PLMs. While code-switching supports metalinguistic tasks, it appears to hinder semantic representation. Models are better at detecting idiomatic usage than capturing meaning, highlighting ongoing limitations in handling complex language. Future work should prioritise open models with improved semantic understanding in multilingual and code-switched settings.

References

Timothy Baldwin and Su Nam Kim. 2010. *Handbook of Natural Language Processing*, 2 edition.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A Hard Nut to Crack: Idiom Detection with Conversational Large Language Models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

A Seza Dogruoz, Sunayana Sitaram, Barbara E Bullock, and Almeida Jacqueline Toribio. 2021. A Survey of Code-switching: Linguistic and Social Perspectives for Language Technologies. In *ACL/IJCNLP*.

Julian Martin Eisenschlos, Jeremy R Cole, Fangyu Liu, and William W Cohen. 2023. WinoDict: Probing language models for in-context word acquisition. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102, Dubrovnik, Croatia. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating Robustness of Dialog Models to Popular Figurative Language Constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aravind K Joshi. 1982. Processing of Sentences With Intra-Sentential Code-Switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of Co-occurrence on Factual Knowledge of Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7721–7735, Singapore. Association for Computational Linguistics.

Murathan Kurfal and Robert Östling. 2020. Disambiguation of Potentially Idiomatic Expressions with Contextual Embeddings. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 85–94, online. Association for Computational Linguistics.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the Ability of Language Models to Interpret Figurative Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Francesca Masini. 2019. Multi-Word Expressions and Morphology.

Filip Miletić and Sabine Schulte im Walde. 2024. Semantics of Multiword Expressions in Transformer-Based Models: A Survey. *Transactions of the Association for Computational Linguistics*, 12:593–612.

Dylan Phelps, Thomas M R Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann A Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Conference on Intelligent Text Processing and Computational Linguistics*.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Aline Villavicencio and Marco Idiart. 2019. Discovering multiword expressions. *Natural Language Engineering*, 25(6):715–733.