A Tough Hoe to Row: Instruction Fine-Tuning LLaMA 3.2 for Multilingual Idiom Processing

Debora Ciminari*[1], Alberto Barrón-Cedeño[1]

[1]Università di Bologna, Corso della Repubblica, 136, 47121, Forlì, Italy

Abstract

Idiomatic expressions (IEs) are a core part of language but exhibit considerable complexity and heterogeneity, posing significant challenges to natural language processing (NLP). Effective automatic idiom processing could enhance our understanding of language and could benefit downstream tasks such as machine translation. However, previous research fails to adopt a comprehensive approach and struggles to consider languages different from English and the rich variety of idiom types. We thus aim to develop a version of LLaMA 3.2 that is instruction fine-tuned on data in three languages — English, Italian, and Portuguese — and covering a wide range of IE types. Specifically, we build on already annotated corpora to create our instruction-formatted dataset, and we employ instruction fine-tuning on two tasks — sentence disambiguation and idiom identification. We then investigate the effectiveness of this approach and assess the impact of the instruction language on the model's performance. We release a multilingual instruction-formatted dataset for automatic idiom processing. Additionally, we show that fine-tuning might help the model disambiguate between literal and idiomatic sentences, while gains in idiom identification are limited and require further investigation. The F_1-measure also suggests that the choice of the instruction language significantly affects the results.

Keywords

idiomatic expressions, multilinguality, sentence disambiguation, idiom identification, instruction fine-tuning

1. Introduction

Idiomatic expressions (IEs) are a prominent component of language and constitute a broad and heterogenous category. The canonical definition describes IEs as expressions whose meaning cannot be derived from the meanings of their subparts [1, 2]. The typical example is `to kick the bucket`, whose meaning 'to die' cannot be inferred from 'kick', 'the', or 'bucket'. However, some cases do not fit this definition. For instance, the meaning of `to pull the strings` ('to use influence or connections') does bear a sort of (metaphorical) relation to its components. Another category of IEs can be identified, i.e. *potentially idiomatic expressions* or *PIEs* [3], which are expressions that can have a literal or an idiomatic meaning, depending on the context. That is the case of the first idiom presented as example, `to kick the bucket`, which can also take a literal meaning, as in `She got frustrated and kicked the bucket of paint across the garage`.

In light of this diversity, the traditional definition has been challenged in favour of a more complex, multifaceted view that emphasises the heterogeneous nature of idiomaticity, conceived of as a continuum where expressions can be placed depending on multiple factors [4].

Such complexity makes it challenging to deal with IEs in the field of natural language processing (NLP). Given the pervasive presence of IEs in language, effective idiom processing is needed to gain a deeper and more comprehensive understanding of language. Idiom-aware NLP can benefit downstream tasks, such as text summarisation, sentiment analysis, question answering, and machine translation [5, 6].

Most NLP applications focus on English, leaving multilingual idiom processing largely unexplored. Recent studies adopt encoder-based models [7, 8, 5], while studies on decoder-based ones remain relatively sparse. Another issue related to previous research is the models' lack of a robust generalisation and its poor performance on unseen idioms [5, 6].

To fill these gaps, we develop an instruction fine-tuned version of LLaMA 3.2 1B in three languages, English, Italian, and Portuguese, and on two tasks, sentence disambiguation and idiom identification:

Task 1: Sentence Disambiguation. Framed as a binary text classification task, it aims at discriminating idiomatic from literal sentences.

Task 2: Idiom Identification. Framed as a span labelling task, the model must identify the sequence of characters that correspond to an IE.

In Task 2, partial matches are considered partially valid: if the model identifies part of the IE, it still receives partial credit for a correct identification.

Both tasks are strictly interconnected: once a model recognises a sentence as idiomatic, it can identify the

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

*Corresponding author.

debora.ciminari@studio.unibo.it (D. Ciminari*);

a.barron@unibo.it (A. Barrón-Cedeño)

0000-0003-4719-34208 (A. Barrón-Cedeño)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2025-11-30

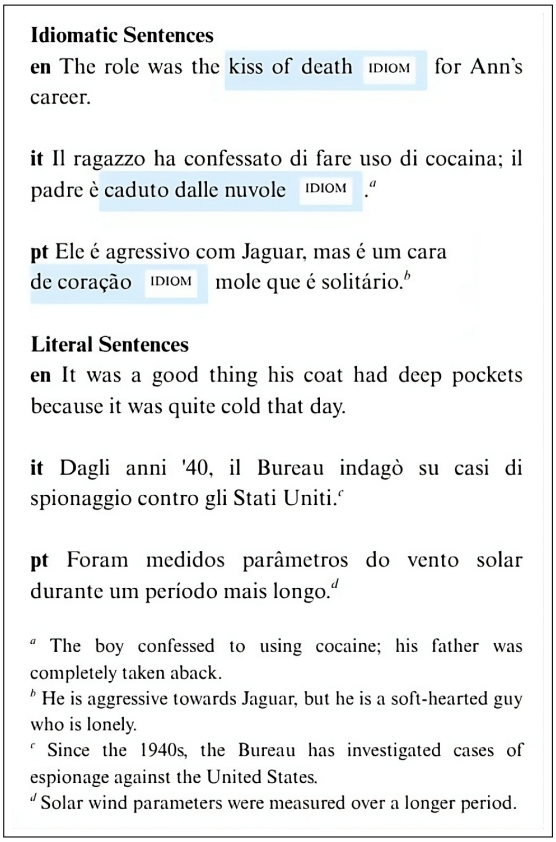

Figure 1: Examples of idiomatic and literal sentences in English, Italian, and Portuguese.

specific span constituting the IE. Figure 1 shows some examples of idiomatic and literal sentences.

Given this interdependence, our data is designed to address both tasks simultaneously. For instance, in the first example, the model's answer is expected to be `kiss of death`, showing that the model correctly identified the sentence as idiomatic and proceeded to detect the span where the idiom occurs.

Starting from annotated corpora, we design our instruction-formatted data, comprising an instruction (the task description), the input (the sentence), and the expected output [9]. Additionally, our dataset is multilingual in that it comprises inputs in all three languages. What differs is the instruction language, for which three subsets are created. We then fine-tune LLaMA 3.2 1B on a subset of our corpus and carry out evaluation based on the F_1-measure. We thus examine the effectiveness of instruction fine-tuning. Besides, we investigate the impact of the instruction language in scenarios where the instruction language and the input language are the same and scenarios where they differ. To date, such an impact has received scant attention in the research literature, with few studies providing contradictory results. Muennighoff et al. [10] explore English and multilingual Multitasked prompting finetuning (MTF), concluding that non-English prompts lead to improved performance, but English prompts still produce satisfactory results on data in other languages. On the other hand, Phelps et al. [11] concentrate specifically on automatic idiom processing and test different models on multiple prompting approaches. They find that, if the instruction language is the same as the input language, the models exhibit a better performance. However, the improvements are not consistent across all models and across Portuguese and Galician, the two languages included.

The contributions of this paper are the following: *(i)* We produce and release a multilingual, instruction-formatted dataset that includes both literal and idiomatic examples and that covers a wide range of idiom types. *(ii)* We show the improvements produced by an approach to idiom processing based on instruction fine-tuning (particularly for Task 1). *(iii)* We highlight the impact of the instruction language on the model's performance.[1]

The paper is structured as follows. Section 2 presents research on automatic idiom processing and sheds light on possible gaps and further developments. Section 3 illustrates the creation of the instruction-formatted dataset in English, Italian, and Portuguese. Section 4 presents the evaluation framework adopted and the specific settings of the experiments we carry out. Section 5 describes the results obtained and compares the model's performance before and after the fine-tuning. Finally, Section 6 draws conclusions and proposes future developments.

2. Related Work

The need to develop *ad hoc* techniques for the automatic processing of idioms is widely acknowledged to acquire a better understanding of language [12, 13, 14]. Multiple natural language understanding (NLU) tasks face challenges related to IEs, despite the use of state-of-the-art (SOTA) solutions. Among these tasks are sentiment analysis [15], paraphrase generation [16], natural language inference [17], dialog models [18], and machine translation [19, 20].

Recent approaches employ encoder-based models, like BERT [21], and leverage their contextual language embedding. Studies have found that this type of models struggles with non-compositionality and has difficulty in disambiguating between literal and idiomatic meanings [22, 7, 23]. Yu and Ettinger [7] explore the ability of encoder-based models to handle semantic compositionality. In particular, they use five models, such as BERT and

[1] The dataset and the implementation are both available at https://github.com/TinfFoil/MultIdiomLlama

some of its variants, to examine to what extent these can represent words in isolation and in phrases. They reach the conclusion that these models grasp the meaning of individual words but struggle to capture composed meaning. Zeng and Bhat [8], instead, propose the iDentifier of Idiomatic expressions via Semantic Compatibility (DISC) to perform extraction and identification of PIEs. Their framework leverages BERT to harness both the semantic and the syntactic properties of PIEs, and extract and identify all the expressions from a corpus. Results show that their model is able to outperform SOTA baselines, even in zero-shot settings, but it exhibits poor cross-domain performance. In addition, while including a notable array of idiom types, it focuses on English data only.

Some approaches take steps to include multilinguality. Tayyar Madabushi et al. [5] release *AStitchinLanguageModels*, a dataset in English and Portuguese, and expand it with Galician data for the SemEval-2022 Task 2 [24]. Working on the idiomaticity detection task, they employ models like BERT and XLNET [25] and conclude that models do not benefit from the inclusion of the context and that the zero-shot setting still produces poor results. This corpus represents the first significant attempt to include multilinguality for the automatic idiom processing and provides baselines for languages other than English. This dataset is, however, limited in that it only contains noun compounds, thus lacking diversity and failing to incorporate other types, such as verb and prepositional phrases. Another attempt at multilingual idiom processing is Tedeschi et al. [6]'s ID10M. They develop a framework of systems and training and validation data for the idiom identification task in 10 languages. Their findings confirm the distinction between zero-shot and few-shot performance.

Sentsova et al. [26] release the *Multilingual Corpus of Potentially Idiomatic Expressions* (MultiCoPIE) in Russian, Italian, and Catalan, which includes additional linguistic features, such as semantic compositionality, head part-of-speech, and English equivalents. By fine-tuning XML-RoBERTa, they explore cross-lingual transfer, which might benefit lower-resourced languages. Moreover, the inclusion of idioms having an English equivalent in the training set has proved helpful in disambiguating between literal and idiomatic usages.

Encoder-decoder models have also been used for the development of idiom-aware systems. Zeng and Bhat [27] opt for the BART [28] sequence-to-sequence (seq2seq) model. Their Generation of Idiom Embedding with Adapter (GIEA) model exhibits an improved ability at representing idiomaticity, but it is limited to English and does not show an enhanced generalisation capability.

Other studies have examined the performance of large language models (LLMs) [29, 11], finding that they fail to handle idiomaticity and that they tend to be outperformed by other transformer-based models.

Previous work falls short of capturing the complexity associated with IEs on multiple levels. On the one hand, studies have mostly focused on English, leaving other languages aside. On the other hand, they have failed to cover a wide enough variety of idiom types. Furthermore, studies agree on the limited ability of different models to handle and process unseen idioms.

3. Instruction Data Creation

Source Datasets. We start from three datasets to build our instruction-formatted data in English, Italian, and Portuguese: *AStitchInLanguageModels* [5], ID10M [6], and MultiCoPIE [26].

AStitchInLanguageModels is a dataset of idiomatic multi-word expression (MWE) usage in English and Portuguese. It comprises examples containing PIEs in the form of noun compounds, annotated according to two different schemes. In the first one, sentences are labelled as having an idiomatic or a literal meaning. The second one is more fine-grained in that it provides a paraphrase of the MWE's meaning and labels each example into one of five categories: literal, idiomatic, non-idiomatic, proper noun, or meta usage. We use data labelled with the first annotation scheme for the zero-shot scenario, with no overlap of PIEs between the training and the test sets.

ID10M is a framework that introduces a multilingual Transformer-based architecture for sentence disambiguation and idiom identification and provides annotated datasets in multiple languages. It includes gold-standard data in English, German, Italian, and Spanish, and silver-standard data automatically annotated in 10 languages: Chinese, Dutch, English, French, German, Italian, Japanese, Polish, Portuguese, and Spanish. A list of MWEs is compiled from the Wiktionary,[2] and sentences containing MWEs are collected from WikiMatrix [30],[3] a multilingual corpus in 83 languages with parallel sentences retrieved from Wikipedia. The gold-standard data are curated by native professional annotators, while the silver-standard data are annotated based on the Wiktionary entry of MWEs: when the MWE is marked as idiomatic, all occurrences of the MWEs are labelled as idiomatic, and vice versa. Since these annotations do not necessarily reflect the actual MWE usage in context, Tedeschi et al. develop a dual-encoder architecture to refine silver-standard data. They also incorporate a BIO tagging scheme [31] to identify the tokens belonging to the MWE, where *B* indicates the first token of a span, *I* signals the intermediate token(s), and *O* designates the tokens out of any span.

[2] https://pypi.org/project/wiktextract/
[3] https://github.com/facebookresearch/LASER/tree/main/tasks/WikiMatrix

MultiCoPIE is a dataset annotated for sentence disambiguation and idiom identification in Russian, Italian, and Catalan. To build this dataset, a list of PIEs is compiled for each language from online resources, such as the Dizionario italiano De Mauro[4], the Russian Wiktionary[5], the Diccionari català-anglès/anglès-català de locucions i frases fetes[6]. PIEs with varying characteristics are included, specifically, PIEs with different parts of speech as heads. For example, `appeso a un filo` ('hung by a thread') has the adjective `appeso` ('hung') as head, while `con l'acqua alla gola` (literally 'with water up to the throat', meaning 'to be in serious difficulty') is headed by the preposition `con` ('with'). The dataset also covers PIEs with diverse degrees of semantic compositionality. PIEs with a higher level of compositionality comprise at least one cue to the meaning of the expression. An example is `ammazzare il tempo` ('to kill time'), where the word `tempo` ('time') helps interpreting the expression as 'to spend time trying not to get bored'. On the other hand, `essere al settimo cielo` ('to be on cloud nine') is more opaque since it does not comprise any hints about the meaning 'to be at the peak of happiness'. After selecting the PIEs, sentences are automatically extracted from the Open Super-large Crawled Aggregated coRpus (OSCAR)[7] [32], a multilingual corpus generated from Common Crawl[8], and refined through manual selection. The two surrounding sentences are included to provide context. Opening and closing tags are also employed to locate the lexicalised components of PIEs. The tags are used to identify all PIEs present in the target sentence and the preceding and following sentences.

Creation of the Instruction Templates. To create a dataset of instruction-formatted instances, we design instructions in English, Italian, and Portuguese. We first translate a seed instruction written in English into Italian and Portuguese using LLaMA 3.2 3B[9] via ollama[10]. With the same model, we generate three paraphrased versions of the instructions. We design the prompts to produce different writing styles and perspectives, ensuring a varied dataset and a high linguistic diversity. These instructions are then organised in empty templates. The starting point to construct such templates is the work by Taori et al. [33], who fine-tune LLaMA 7B on instruction-formatted demonstrations. They design a template in English to create the instruction-formatted examples and carry out the fine-tuning. We translate their template into Italian and Portuguese. The 'prompt no input' option is discarded since all our samples include an input sentence. Finally, we change the structure of the template. While the Alpaca template[11] organises the instruction in "Instruction", "Input", and "Response", we modify the order so that the input is first presented, followed by the instruction and the response, since this order better fits language modeling underlying LLMs. This order meshes well with the left-to-right autoregressive nature of LLaMA: as shown in Table 1, the instruction leaves an empty slot at the end, where the model's response is expected. Finally, the 'input' and 'output' keys are left empty to be filled in the following step.

Table 1
Examples from the instruction dataset with the output produced given an instruction and an input in different language combinations.

Examples
Input (en): Although the encounter was bathed in sunshine, the match failed to reach boiling point. **Instr. (en)**: Can you spot the idiomatic expressions lurking within this sentence? They are: **Output**: boiling point
Input (pt): Nos últimos anos, muitas universidades têm mostrado quadricópteros realizando manobras aéreas.[a] **Instr. (it)**: Un'analisi della frase rivela la presenza delle seguenti costruzioni idiomatiche:[b] **Output**: Nessuna.[c]
Input (en): After the day I had today , I feel like I could walk on water. **Instr. (pt)**: A frase contém as seguintes expressões idiomáticas:[d] **Output**: walk on water

[a] In recent years, many universities have demonstrated quadcopters performing aerial manoeuvres.
[b] An analysis of the sentence reveals the presence of the following idiomatic constructions:
[c] None.
[d] The sentence contains the following idiomatic expressions:

Creation of the Final Dataset. We then proceed with the creation of the final dataset. We extract IEs and examples from the aforementioned datasets. For English and Portuguese, we use ID10M and *AStitchInLanguageModels*, while, for Italian, we employ ID10M and MultiCoPIE. The processing of the *AStitchInLanguageModels* mainly focuses on extracting the actual MWEs present in the sentences since it includes the dictionary form. For ID10M, we process the data by reconstructing full sentences and identifying idiomatic spans. We then create a training and test split combining data from both ID10M and *AStitchInLanguageModels*, while ensuring that no PIEs in the test set overlap with those in the training

[4] https://dizionario.internazionale.it/
[5] https://ru.wiktionary.org/wiki/
[6] https://visca.com/apac/dites/
[7] https://huggingface.co/oscar-corpus
[8] https://commoncrawl.org/
[9] https://huggingface.co/meta-llama/Llama-3.2-3B
[10] https://ollama.com/
[11] https://github.com/tloen/alpaca-lora/blob/main/templates/alpaca_short.json

Table 2
Statistics of an instruction subset for each of the three languages.

Language	Idiomatic	Literal	Instances
en	12,415	12,415	24,830
it	10,999	10,999	21,998
pt	6,448	6,448	12,896
Total	29,862	29,862	59,724

data. Finally, we apply text cleaning operations, such as fixing contractions and punctuation spacing and export two final processed splits per language.

For the MultiCoPIE data, we retrieve sentences and the relative PIEs enclosed in annotation tags to obtain the non-lemmatised versions. Next, we combine these data with ID10M's Italian data and balance the whole dataset by undersampling literal instances and splitting into training and test sets, while preventing PIEs overlap between them. Finally, text cleaning is applied to improve consistency.

Once extracted the IEs and the sentences for all three languages, we merge all sets into unified training and test datasets containing instances from English, Italian, and Portuguese. We then populate the 'input' and 'output' fields of the templates with such examples. The final dataset comprises three subsets containing the same set of examples, with English, Italian, and Portuguese as the input languages. The subsets thus have identical sizes and balanced distributions of idiomatic and literal sentences and only differ in the instruction language. Table 2 shows the statistics for one representative subset.

4. Experimental Settings

Evaluation Framework. To account for both tasks, we propose a two-fold evaluation methodology, which allows for a comprehensive understanding of the model's ability to handle both the classification and the identification challenges.

We design an evaluation framework to assess the model's performance on the sentence disambiguation and idiom identification tasks across various language combinations. For Task 1 we develop a labelling mechanism that considers multiple linguistic markers. Such markers are used for both ground truths and predictions to determine the label (0 or 1) to assign to each example. These keywords are language-specific and are:

- Portuguese: 'nenhuma', 'não', 'ausente';
- Italian: 'nessuna', 'non';
- English: 'none', 'no idiom', 'not contain', 'not'.

The label assignment can be represented as follows:

$$\text{label} = \begin{cases} 0 & \text{if keywords are present} \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

According to the assigned labels, precision, recall, and F_1-measure scores are computed.

For Task 2, the evaluation approach adopts partial text span matching. Following the approach proposed by Da San Martino et al. [34], we give credit to partial matches of the identified spans. This means that if the model identifies a portion of the IE but not the whole span, it will be credited for the overlap. This approach allows for a more flexible assessment of the model's performance. The span matching is character-based and is computed through the Longest Common Subsequence (LCS), which aims to find the longest subsequence two sequences have in common, keeping the order of characters unchanged. LCS is thus order-sensitive, but the characters in the subsequence do not necessarily have to be contiguous. This is important because IEs may contain lexicalised components that are spread across a span, while particles (such as auxiliary verbs or personal pronouns) may appear within the expression but do not need to be included in the span for the match to be considered correct. In other words, LCS is flexible in terms of character proximity, as it allows for gaps in the sequence, but requires that the identified IE maintains the same order found in the ground truth span. This is crucial since it enables to account for variations occurring within the IE.

Based on LCS, character overlap is determined and used to compute precision and recall:

$$P(S,T) = \frac{1}{|S|} \sum_{d \in D} \sum_{s \in S_l, t \in T_l} \frac{|s \cap t|}{|t|} \tag{2}$$

$$R(S,T) = \frac{1}{|T|} \sum_{d \in D} \sum_{s \in S_l, t \in T_l} \frac{|s \cap t|}{|s|} \tag{3}$$

where s is the predicted span, t is the ground truth span, S is the set of predicted spans, T is the set of gold standard spans, d is a sample, and D represents the whole dataset. The F_1-measure is then computed as the harmonic mean of precision and recall.

Settings. The instruction fine-tuning is implemented on a subset of our dataset. This subset comprises 18,397 samples and retains the balance of the instruction dataset. To optimise the fine-tuning, QLoRA [35] is also employed to reduce computational cost and memory usage.

As for the instruction fine-tuning, a set of default hyperparameters is configured to implement the fine-tuning of the LLaMA-3.2 1B model for the sentence disambiguation and idiom identification tasks. The model is trained with a batch size of 32 across 2 epochs, using a cutoff

Table 3
Results for Task 1 in terms of F_1-measure for the three instruction (left) and the three input languages (top). Best in bold.

Inst.	Input		
	en	**it**	**pt**
Baseline			
en	0.7134±0.0014	0.6278±0.0000	0.6101±0.0055
it	0.7075±0.0012	0.6508±0.0051	0.6172±0.0040
pt	0.7061±0.0028	0.6580±0.0.0013	0.5573±0.0115
Ours			
en	**0.9557±0.0076**	**0.9020±0.01380**	**0.8898±0.0376**
it	0.9360±0.0180	0.8585±0.0382	0.8450±0.0802
pt	0.9165±0.0543	0.8983±0.0405	0.8403±0.0347

Table 4
Results for Task 2 in terms of F_1-measure for the three instruction (left) and the three input languages (top). Best in bold.

Inst.	Input		
	en	**it**	**pt**
Baseline			
en	0.3064±0.0078	0.3029±0.0036	0.2394±0.0057
it	0.3190±0.0013	**0.3280±0.0050**	0.2804±0.0012
pt	0.3200±0.0015	0.3091±0.0090	0.2754±0.0125
Ours			
en	0.4494±0.0570	0.2033±0.0632	0.2634±0.0875
it	0.5324±0.0921	0.2335±0.0920	0.3139±0.0946
pt	**0.5465±0.0814**	0.3261±0.1220	**0.3234±0.1230**

length of 128 tokens for input sequences. For parameter-efficient fine-tuning, LoRA [36] is employed with a rank (r) of 8, alpha of 16, and dropout rate of 0.05, specifically targeting the query and key projection matrices. The implementation of LoRA enables to update only 851,968 out of more than 1 billion parameters. The optimisation process uses 4-bit quantization with NF4 format to reduce memory requirements. The learning process is managed with a learning rate of 3e-4, weight decay of 0.01, and a warmup ratio of 0.1, using the Paged AdamW 32-bit optimizer and cosine learning rate schedule with restarts. Gradient accumulation is set to 2 steps with a maximum gradient norm of 1.0, and gradient checkpointing is enabled to optimise memory usage. The training uses mixed-precision computation (FP16) and employs early stopping.

5. Results and Discussion

Sentence Disambiguation Task. Table 3 shows the F_1 scores for Task 1, averaged over 3 runs, for all combinations of instruction and input language, before and after the fine-tuning. When comparing our model against the baseline model without fine-tuning, we can see that the best results are achieved after the instruction fine-tuning: the performance gains more than 2 points across all combinations, with the Portuguese monolingual pair increasing by almost 3 points. These findings suggest that the approach we adopted consistently enhances the model's performance, regardless of the instruction-input language combination. Turning to the impact of the instruction language, the baseline results indicate that English inputs tend to prefer English instructions. On the other hand, there seem to exist some sort of interplay between Italian and Portuguese: a slight improvement is produced when Italian data are associated with Portuguese instructions and vice versa. Conversely, our results show that the model yields better F_1 scores when prompted with instructions in English across all language combinations. This suggests that the fine-tuning leads the model to prefer instructions written in English when disambiguating between literal and idiomatic sentences.

Idiom Identification Task. Table 4 shows F_1 scores, averaged over 3 runs, for Task 2, before and after instruction fine-tuning. We can see that, in general, the model exhibits poor performance and struggles to identify the idiom contained in the input sentence. In the idiom identification task, the improvements produced by the instruction fine-tuning are mostly lower or non-existent. The English inputs tend to benefit more from this approach, gaining 2 points almost with all languages. Conversely, the model seems to struggle on Italian data, and, when associated with Italian and English instructions, it suffers from the fine-tuning, losing 1 point. When dealing with Portuguese sentences, instead, the model produces slightly improved results. Instruction fine-tuning, therefore, does not significantly and consistently help the model in identifying idioms. However, we should consider that Task 2 is much more challenging in that it consists in the identification of the idiom contained in a given sentence, at the character level. As for the instruction language, unlike Task 1, the instruction fine-tuning does not lead the model to favour English. Instead, Portuguese instructions seem to better help the model in detecting the idiom.

Interactions between Instruction and Input Language. The results abovementioned provide insights into the interactions between instruction and input language. For Task 1, English instructions seem to aid the model in distinguishing between idiomatic and literal sentences. Sentence disambiguation represents a simpler task that requires a global understanding of the input sentence. English, on which the model is mostly pretrained [37], might better allow LLaMA 3.2 to comprehend the task to carry out. Idiom identification, instead, is a much more complicated task requiring the model to have a deeper and more precise comprehension, not only at the sentence level, but also at the phrase level. This

entails a finer knowledge of the input language as well. Besides, when the instruction and the input language differ, the model is prompted in one language and asked to answer in another, which creates an additional layer of complexity. Different types of interactions between instruction and input language thus emerge, and future research is needed to investigate such interactions based on the languages involved and the task under study.

6. Conclusions

In this paper, we developed a fine-tuned LLaMA 3.2 1B on two tasks: sentence disambiguation and idiom identification. We adopted a multilingual approach in that we considered three languages, English, Italian, and Portuguese, and we employed instruction fine-tuning. To carry out the fine-tuning, we first constructed a multilingual dataset consisting of instruction-formatted data designed for idiomatic expressions (IEs). We examined the two tasks in a multilingual setting involving the abovementioned languages, which were used as both instruction and input languages, covering all possible combinations. This fine-tuning provided some valuable insights.

For the sentence disambiguation task, our instruction-based approach yielded better F_1 scores, compared to the baseline results, which suggests that it aids the model in distinguishing between idiomatic and literal meanings. Nevertheless, after the fine-tuning, the models seemed to favour English instructions across all input languages. This might indicate that we can achieve satisfactory results prompting models with English instructions [10], and that we can limit instruction engineering to only one language [38]. On the other hand, this can be disadvantageous for other languages, potentially reducing model performance and usability in multilingual contexts.

For the idiom identification task, the model struggled to correctly identify the idiom included in the sentence, both before and after the fine-tuning. Our instruction-based approach did not necessarily lead to a significantly improved performance, and, in some cases, it produced lower F_1 scores. Unlike Task 1, Task 2 represents a far more challenging task consisting in detecting IE at the character level, which might explain such a poor performance. Besides, the model did not exhibit a consistent preference for one language and produced mixed results.

Instruction fine-tuning might be beneficial for Task 1 but not necessarily for Task 2, and the instruction language plays a crucial role in the model's performance.

However, further research is needed. From a methodological perspective, we used a relatively small model, and experiments with larger ones can be conducted. Other LLMs beyond LLaMA could be fine-tuned as well, not only to assess their performance but also to compare encoder-based and encoder-decoder models on the same IE-related tasks. We did not implement hyperparameter tuning and limited the fine-tuning to a small subset. Future research could explore optimised hyperparameters to improve performance, as well as use a larger dataset. Our study was also limited to three languages, and the scope could be expanded to others, even from different families, to gain a deeper understanding of cross-linguistic interactions. Finally, a promising direction would be the creation of datasets annotating idiomaticity on a continuum rather than as a binary distinction, aligning with more recent linguistic theories.

References

[1] B. Fraser, Idioms within a Transformational Grammar, Foundations of Language 6 (1970) 22–42.

[2] N. A. Chomsky, Rules and Representations, Behavioral and Brain Sciences 3 (1980) 1–15. doi:10.1017/s0140525x00001515.

[3] H. Haagsma, J. Bos, M. Nissim, MAGPIE: A large corpus of potentially idiomatic expressions, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 279–287. URL: https://aclanthology.org/2020.lrec-1.35/.

[4] S. Wulff, Rethinking Idiomaticity: A Usage-based Approach, Research in Corpus and Discourse, Continuum, London and New York, 2008.

[5] H. Tayyar Madabushi, E. Gow-Smith, C. Scarton, A. Villavicencio, AStitchInLanguageModels: Dataset and Methods for the Exploration of Idiomaticity in Pre-Trained Language Models, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 3464–3477. URL: https://aclanthology.org/2021.findings-emnlp.294/. doi:10.18653/v1/2021.findings-emnlp.294.

[6] S. Tedeschi, F. Martelli, R. Navigli, ID10M: Idiom Identification in 10 Languages, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 2715–2726. URL: https://aclanthology.org/2022.findings-naacl.208/. doi:10.18653/v1/2022.findings-naacl.208.

[7] L. Yu, A. Ettinger, Assessing Phrasal Representation and Composition in Transformers, in:

B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4896–4907. URL: https://aclanthology.org/2020.emnlp-main.397/. doi:10.18653/v1/2020.emnlp-main.397.

[8] Z. Zeng, S. Bhat, Idiomatic expression identification using semantic compatibility, Transactions of the Association for Computational Linguistics 9 (2021) 1546–1562. URL: https://aclanthology.org/2021.tacl-1.92/. doi:10.1162/tacl_a_00442.

[9] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, G. Wang, Instruction Tuning for Large Language Models: A Survey, 2024. URL: https://arxiv.org/abs/2308.10792. arXiv:2308.10792.

[10] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. Le Scao, M. S. Bari, S. Shen, Z. X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, C. Raffel, Crosslingual generalization through multitask finetuning, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 15991–16111. URL: https://aclanthology.org/2023.acl-long.891/. doi:10.18653/v1/2023.acl-long.891.

[11] D. Phelps, T. Pickard, M. Mi, E. Gow-Smith, A. Villavicencio, Sign of the times: Evaluating the use of large language models for idiomaticity detection, 2024. URL: https://arxiv.org/abs/2405.09279. arXiv:2405.09279.

[12] I. A. Sag, T. Baldwin, F. Bond, A. Copestake, D. Flickinger, Multiword expressions: A pain in the neck for NLP, in: A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 1–15.

[13] A. Villavicencio, F. Bond, A. Korhonen, D. McCarthy, Editorial: Introduction to the special issue on multiword expressions: Having a crack at a hard nut, Comput. Speech Lang. 19 (2005) 365–377. URL: https://doi.org/10.1016/j.csl.2005.05.001. doi:10.1016/j.csl.2005.05.001.

[14] T. Baldwin, S. N. Kim, Multiword Expressions, CRC Press LLC, 2010, pp. 267–292.

[15] R. Biddle, A. Joshi, S. Liu, C. Paris, G. Xu, Leveraging sentiment distributions to distinguish figurative from literal health reports on twitter, in: Proceedings of The Web Conference 2020, WWW '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1217–1227. URL: https://doi.org/10.1145/3366423.3380198. doi:10.1145/3366423.3380198.

[16] J. Zhou, Z. Zeng, H. Gong, S. Bhat, Idiomatic Expression Paraphrasing without Strong Supervision, 2021. URL: https://arxiv.org/abs/2112.08592. arXiv:2112.08592.

[17] T. Chakrabarty, D. Ghosh, A. Poliak, S. Muresan, Figurative language in recognizing textual entailment, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 3354–3361. URL: https://aclanthology.org/2021.findings-acl.297/. doi:10.18653/v1/2021.findings-acl.297.

[18] H. Jhamtani, V. Gangal, E. Hovy, T. Berg-Kirkpatrick, Investigating robustness of dialog models to popular figurative language constructs, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7476–7485. URL: https://aclanthology.org/2021.emnlp-main.592/. doi:10.18653/v1/2021.emnlp-main.592.

[19] M. Fadaee, A. Bisazza, C. Monz, Examining the tip of the iceberg: A data set for idiom translation, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1148/.

[20] E. Liu, A. Chaudhary, G. Neubig, Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 15095–15111. URL: https://aclanthology.org/2023.emnlp-main.933/. doi:10.18653/v1/2023.emnlp-main.933.

[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[22] N. Nandakumar, T. Baldwin, B. Salehi, How well do embedding models capture non-compositionality? A view from multiword expressions, in: A. Rogers, A. Drozd, A. Rumshisky, Y. Goldberg (Eds.), Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP, Association for Computational Linguistics, Minneapolis, USA, 2019, pp. 27–34. URL: https://aclanthology.org/W19-2004/. doi:10.18653/v1/W19-2004.

[23] M. Garcia, T. Kramer Vieira, C. Scarton, M. Idiart, A. Villavicencio, Probing for idiomaticity in vector space models, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 3551–3564. URL: https://aclanthology.org/2021.eacl-main.310/. doi:10.18653/v1/2021.eacl-main.310.

[24] H. Tayyar Madabushi, E. Gow-Smith, M. Garcia, C. Scarton, M. Idiart, A. Villavicencio, SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 107–121. URL: https://aclanthology.org/2022.semeval-1.13/. doi:10.18653/v1/2022.semeval-1.13.

[25] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, XLNet: generalized autoregressive pretraining for language understanding, Curran Associates Inc., Red Hook, NY, USA, 2019.

[26] U. Sentsova, D. Ciminari, J. V. Genabith, C. España-Bonet, MultiCoPIE: A multilingual corpus of potentially idiomatic expressions for cross-lingual PIE disambiguation, in: A. K. Ojha, V. Giouli, V. B. Mititelu, M. Constant, G. Korvel, A. S. Doğruöz, A. Rademaker (Eds.), Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025), Association for Computational Linguistics, Albuquerque, New Mexico, U.S.A., 2025, pp. 67–81. URL: https://aclanthology.org/2025.mwe-1.8/.

[27] Z. Zeng, S. Bhat, Getting BART to ride the idiomatic train: Learning to represent idiomatic expressions, Transactions of the Association for Computational Linguistics 10 (2022) 1120–1137. URL: https://aclanthology.org/2022.tacl-1.65/. doi:10.1162/tacl_a_00510.

[28] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703/. doi:10.18653/v1/2020.acl-main.703.

[29] F. De Luca Fornaciari, B. Altuna, I. Gonzalez-Dios, M. Melero, A hard nut to crack: Idiom detection with conversational large language models, in: D. Ghosh, S. Muresan, A. Feldman, T. Chakrabarty, E. Liu (Eds.), Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024), Association for Computational Linguistics, Mexico City, Mexico (Hybrid), 2024, pp. 35–44. URL: https://aclanthology.org/2024.figlang-1.5/. doi:10.18653/v1/2024.figlang-1.5.

[30] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, F. Guzmán, Wikimatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics., 2021, pp. 1351–1361. doi:10.18653/v1/2021.eacl-main.115.

[31] L. Ramshaw, M. Marcus, Text chunking using transformation-based learning, in: Third Workshop on Very Large Corpora, 1995. URL: https://aclanthology.org/W95-0107/.

[32] P. J. Ortiz Suárez, B. Sagot, L. Romary, Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures, Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, Leibniz-Institut für Deutsche Sprache, Mannheim, 2019, pp. 9 – 16. URL: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215. doi:10.14618/ids-pub-9021.

[33] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following LLaMA model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[34] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles, in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1377–1414. URL: https://aclanthology.org/2020.semeval-1.186/. doi:10.18653/v1/2020.semeval-1.186.

[35] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized

LLMs, 2023. URL: https://arxiv.org/abs/2305.14314. arXiv:2305.14314.
[36] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.
[37] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, et al., The Llama 3 herd of models, arXiv (Cornell University) (2024). doi:10.48550/arxiv.2407.21783.
[38] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, 2025. URL: https://arxiv.org/abs/2303.18223. arXiv:2303.18223.

Declaration on Generative AI

During the preparation of this work, the author(s) did not use any generative AI tools or services.