Alma Mater Studiorum · Università di Bologna

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE
Corso di Laurea magistrale in Specialized Translation (classe LM-94)

TESI DI LAUREA
in NATURAL LANGUAGE PROCESSING

A Tough Row to Hoe: Instruction Fine-Tuning LLaMA 3.2 for Multilingual Sentence Disambiguation and Idiom Identification

CANDIDATA:
DEBORA CIMINARI

Relatore:
Alberto Barrón-Cedeño

Correlatrice:
Maja Miličević Petrović

Anno Accademico 2023/2024
Terzo Appello

Acknowledgements

First and foremost, I would like to thank prof. Barrón and prof. Miličević, not only for their guidance throughout my thesis work, but also for their availability and understanding. This thesis has been a real challenge, and their help was essential to complete it.

A special thank you goes to my family: to my mom, who has always encouraged me to pursue what I am passionate about and to fulfil myself; to my brothers, Michael and Kevin, who have shown their support in their own, very different ways; and to my sister-in-law, Marta, who has been an inspiration in my university journey.

I am also deeply grateful to my best friend Sofia. In over ten years of friendship, you have always made feel like I belong. In the hardest times you have always been by my side in the way I needed the most, which shows how well you know me. Despite the distance of the past few years, you remain an irreplaceable point of reference, and I can only thank you for choosing to be my best friend.

I would also like to thank the people I met in Forlì: my friends and flatmates from Apartment 31, Adele, Martine, Megane, Giusy, and our honorary flatmates, Shirley and Francesca. The time spent in Forlì has been among the most exciting and memorable, thanks to you. The lightness and joy you have brought into my life are invaluable. I am so happy that I get to spend some more months with you. A special thank you goes to Adele and Martine: you opened up to me and made me feel free to do the same with you. I know that once we go our separate ways, I will deeply miss our midnight conversations where we talk about the most personal (and sometimes embarrassing) things. My growth of the past few years has been possible also thanks to you.

I would also like to thank the people at the Sassi Masini, whom I have got to know over the last few months. You have made this last period of my

MA much lighter and bearable.

Finally, I would like to dedicate this thesis to my little Mariasole. The love you have brought into my life is unbelievable. I hope that, through this and my future work, I can become an aunt who sets an ever-inspiring example for you.

Abstract

Idiomatic expressions (IEs) are a fundamental aspect of language, traditionally defined as expressions whose meanings cannot be inferred from their individual components. However, modern linguistic theories propose a more complex definition of idiomaticity, which is now understood as a continuum where IEs can be placed depending on multiple factors. This complexity poses challenges for natural language processing (NLP) applications, where effective handling of IEs can improve performance in various tasks, including sentiment analysis, question answering, text summarisation, and machine translation. This thesis contributes to the study of IEs in NLP by instruction fine-tuning LLaMA 3.2 1B on two tasks: sentence disambiguation and idiom identification. To this end, a multilingual instruction-formatted dataset was created, incorporating English, Italian, and Portuguese as both instruction and input languages. This enabled to investigate the interaction between the instruction and input language and examine the model's performance when they match and when they differ. The findings showed that aligning instruction and input languages does not always improve performance, highlighting complex cross-linguistic interactions. However, while fine-tuning enhanced idiom identification, it led to slight declines in sentence disambiguation, possibly due to dataset limitations and lack of hyperparameter tuning. Future work could expand language diversity, refine fine-tuning strategies, and explore other LLM architectures for better performance.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Idiomatic expressions (IEs) are a large and fundamental part of language. Due to their heterogenous nature, they resist clear and shared categorisation. Traditionally, IEs were defined as expressions whose meaning cannot be entirely derived from the meanings of their subparts (e.g. Chomsky, 1980; Fraser, 1970). Over time, the conceptualisation of IEs evolved from a static, monolithic definition to a multi-faceted scalar conception. Idiomaticity is no longer seen as an all-or-nothing characteristic, but as a continuum where various factors come into play (Wulff, 2008). These factors pertain to linguistic levels, such as semantics and syntax. For instance, consider *to kick the bucket*, which conveys the idiomatic meaning 'to die'; while this expression is relatively fixed, other idioms, such as *to spill the beans*, license greater syntactic variation (e.g., the passive *the beans were spilled*).

Given such complexity, natural language processing (NLP) applications struggle to deal with IEs. Yet, developing methods to grasp idiomaticity allows for the creation of systems that have a more nuanced understanding of language. This can benefit downstream tasks, such as sentiment analysis, question answering, text summarisation, and machine translation (Tayyar Madabushi et al., 2021; Tedeschi et al., 2022).

Recent approaches mainly rely on encoder-based models, like BERT (Devlin et al., 2019) and its variants. Despite the significant progress, studies are limited due to idiom diversity and variability. They also tend to focus on English, leaving multilingual idiom processing largely unexplored. Additionally, such models lack robust generalisation and do not perform well on unseen IEs.

Recently, studies have started exploring the use of large language models

Instruction	**Input**	**Output**
Determine if the sentence has an idiomatic or literal meaning.	*Unni, the stylist, is on cloud nine after having an opportunity to style the beard of his favourite star.*	Idiomatic

Table 1.1: An example of an instruction-formatted instance.

(LLMs) for idiom processing, particularly LLaMA (Touvron et al., 2023), an open-source collection of models pre-trained on publicly available data. These studies show that LLMs generally underperform in idiom-related tasks, compared to encoder-based and encoder-decoder models.

However, research on LLMs for idiomaticity remains sparse and fails to provide such models with a fine-tuning specific to IEs. For example, these large-scale models could benefit from instruction fine-tuning, which involves adapting LLMs using instruction-output pairs (Zhao et al., 2023). As shown in 1.1, the instruction provides a task description in natural language, while the output represents the desired result. An input may also be included, for instance, in the form of the target sentence to classify as idiomatic or literal.

This approach enhances LLMs' generalisation and controllability, enabling better performance on unseen tasks and more predictable behaviour. Instructions can be either LLM-oriented (more aligned with the pre-training objective of LLMs) or human-oriented (more descriptive). Additionally, they can be expressed in different languages, but results on the role of the instruction language are inconsistent. According to Zhang et al. (2023), instructions designed in English generally lead to satisfactory results. On the other hand, in their study focused on automatic idiom processing, Phelps et al. (2024) suggest that translating the instruction into the target language can boost the model's performance. Given that most instruction datasets are predominantly in English, there is a clear need for a more multilingual approach to instruction fine-tuning ((Lou et al., 2024; Peng et al., 2023, ;).

To address these gaps, this study aims to develop an instruction finetuned version of LLaMA 3.2, tailored specifically for the tasks of sentence disambiguation and idiom identification in three language, English, Italian, and Portuguese. To achieve our objective, we attempt to answer the following research questions:

Research Question 1: when the instruction language and the

input language are the same, does the model's performance improve, as opposed to when they differ?

Research Question 2: are there specific language combinations that facilitate cross-lingual transfer learning?

To answer the research questions, I construct two separate instruction datasets based on the instruction type: LLM-oriented and human-oriented. Each dataset is further split into three subsets according to the instruction language, which can be English, Italian, or Portuguese, while including inputs from all three languages. I extract IEs and the input sentences in which they occur from already annotated corpora. Once I have created the datasets, I instruction fine-tune LLaMA 3.2 1B for sentence disambiguation and idiom identification. Within the scope of this thesis, I focus on fine-tuning LLaMA on the LLM-oriented instruction data. I then carry out the evaluation phase. Specifically, I examine the F1 scores resulting from each combination of instruction language and input language and attempt to answer the research questions. Evaluation is conducted in a zero-shot setting to investigate the performance of the model on unseen IEs.

This thesis seeks to deepen our understanding of how decoder-based models handle IEs, particularly within a multilingual context and with different types of expressions. I present a targeted fine-tuning approach designed to enhance LLaMA 3.2's ability to interpret idioms. Additionally, this thesis makes a significant contribution by constructing a multilingual instruction fine-tuning dataset in English, Italian, and Portuguese, specific to IEs. Finally, one of the byproducts of this research is the publication of the following scientific article, to be presented at the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL):

Uliana Sentsova, Debora Ciminari, Cristina España-Bonet, and Josef van Genabith. MultiCoPie: A multilingual corpus of potentially idiomatic ex- pressions for cross-lingual pie disambiguation. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, June 2025. To appear.

The remaining part proceeds as follows.

Chapter 2 provides the theoretical framework on which this study hinges. First, I delve into IEs and shed light on the complexity of defining such expressions, presenting the various linguistic approaches that have contributed to our understanding of IEs. I then provide an overview of LLMs, addressing their main characteristics and types, and I present the LLaMA's model family. Finally, I describe LLMs' learning paradigm, focusing on instruction fine-tuning.

Chapter 3 describes the process of creating the instruction-formatted instances in English, Italian, and Portuguese. Specifically, I first present the extraction of the annotated sentences from pre-existing corpora, along with the IEs they include. Then, I outline the steps to construct the different subsets of the dataset, based on the instruction language and the instruction type (LLM-oriented and human-oriented).

Chapter 4 is dedicated to the experiment I conducted. First, I describe the experiment's details, including QLoRa (Dettmers et al., 2023), a parameter-efficient technique. I then report the results of the finetuning in terms of precision, recall, and F1 score, and attempt to answer the research questions.

Chapter 5 draws conclusions and proposes some future steps.

Chapter 2

Background

This chapter describes the theoretical foundations underpinning this study. In particular, Section 2.1 discusses idiomatic expressions (IEs) from a linguistic perspective, providing an overview of the different definitions and classifications developed over time. Section 2.2 is concerned with the description of LLMs and illustrates their characteristics, types, and fine-tuning methods, with particular emphasis on instruction tuning. Finally, Section 2.3 reviews previous studies on automatic idiom processing, highlighting their benefits and limitations.

2.1 Idiomatic Expressions

IEs are an important component of natural language, and multiple approaches have been adopted to analyse them. However, a precise definition has proved elusive, and there is little consensus about what an IE actually constitutes. The following sections illustrate the evolution of the conceptualisation of IEs and underscore the complex aspects that challenge their categorisation.

2.1.1 The Non-Compositionality View

One of the most influential contemporary linguistic theories, i.e. generative grammar, equates idiomaticity with non-compositionality and provides the canonical definition of an IE, i.e. an expression whose meaning cannot be deduced from the meanings of its component words (e.g. Chomsky, 1980; Fraser, 1970). The IE typically used to exemplify non-compositionality is

to kick the bucket, whose meaning ('to die') cannot be inferred from *kick*, *the*, or *bucket*. According to Cacciari (1993, p. 33), such a definition is supported by three arguments. First, she examines the verb *to break* in two different occurrences: *to break a cup* and *to break the ice*: while the semantic interpretation of the former only requires the knowledge of the meaning of each word, the interpretation of the latter (assumed a figurative meaning) is possible only by retrieving the expression from the lexicon memory. Second, since IEs are included in figurative language (which also comprises metaphor, irony, and metonymy, among others), non-compositionality is considered the only factor that makes it possible to draw a distinction between IEs and metaphors. According to such a view, while the meanings of the component words are exploited by metaphors, they do not contribute to the meaning of IEs. This opposition is also mirrored in the creativity associated with metaphors and the frozenness associated with IEs. This fixedness also serves as the third point on which the non-compositionality hypothesis hinges, since any internal modification would entail a shift in the meaning of the IE.

2.1.2 Beyond Non-Compositionality

Psycholinguistic studies have challenged the strict non-compositional view of IEs, arguing that variability is a property of idioms. These studies suggest that idioms can sometimes be understood compositionally and that non-compositionality does not always imply idiomaticity. IEs' frozenness has been contested on the grounds that they do exhibit some degree of syntactic flexibility, and some preserve their meaning when subject to adjectival modification, quantification, topicalisation, and ellipsis (e.g., Wasow et al., 1983; Cacciari, 1993; Titone and Connine, 1999; Nunberg et al., 1994). For example, the IE *to pull the strings* (meaning 'to use influence or connections') licenses quantification without changing its meanins, as in 'She pulled some strings to get the job'. The second argument against the non-compositional approach runs counter to the assumption that the relationship between idiom meanings and idiom parts is arbitrary. Some IEs are grounded in conceptual metaphors that connect the meaning of the expression to the meanings of its parts (e.g., Nunberg et al., 1994). This relationship is the basis of what has been termed 'motivation' (Gibbs and O'Brien, 1990; Gibbs, 1992). Gibbs and his collaborators have found that speakers have intuitions about the rationale of some IEs, which help them in idiom processing and compre-

hension. The notion of motivation points towards the notions of 'semantic compositionality' and 'analysability': the knowledge of components' meanings, coupled with the knowledge of the underlying metaphor, allows speakers to understand an IE and to map the literal-local level of meaning to the figurative-global one (Cacciari, 1993, p. 35).

Based on semantic compositionality, Gibbs and Nayak (1989) present a threefold classification of idioms (see also Nunberg, 1978): normally decomposable, abnormally decomposable, and non-decomposable. In normally decomposable IEs, all component words refer to their own idiomatic referents; an example is *to pop the question*, where the noun *question* refers to the marriage proposal, and the verb *pop* refers to the act of making it. On the other hand, abnormally decomposable IEs are identified when no such a relation can be established for all components, and idiom comprehension stems from the knowledge of the conceptual metaphor underlying the expression. For example, the understanding of *to carry a torch* ('to be in love') is based on *torch* being a conventional metaphor for warm feelings. Finally, IEs whose meaning does not bear any sort of relation to its parts fall into the third category of non-decomposable IEs.

According to Cacciari and Glucksberg (1991, p. 230), in both normally and abnormally decomposable IEs, the idiom meaning establishes a relation with the idiom parts (quasi-literal and metaphorical, respectively), and such a distinction has no bearing on idiom comprehension. In their functional typology of IEs, they propose four categories that are determined by the existence and nature of the functional relations between the IE's elements and the IE's meaning. Where there is no such a relation, idioms are 'non-analysable', while, if some relationships can be discerned, idioms are analysable. Within this category, idioms can be:

1. analysable-opaque (the relations might be opaque, but the meaning of the idiom parts can constrain interpretation and use, as in *to kick the bucket*);

2. analysable-transparent (there is a clear correspondence between the idiom's elements and components of the idiom's meaning that is usually metaphorical, for instance *to spill the beans*);

3. quasi-metaphorical (the whole idiom meaning constitutes a metaphor and can be associated with a literal referent, which serves as an ideal

> example of a certain concept, such as *to give up the ship* referring to the act of surrendering).

The distinction between non-analysable and analysable translates, in Nunberg et al. (1994), into the distinction between idiomatic phrases (IPs) and idiomatically combining expressions (ICEs). While these types differ for the compositional nature, they share what Nunberg et al. (1994, p. 498) describes as 'conventionality', which they define as

> the discrepancy between the idiomatic phrasal meaning and the meaning we would predict for the collocation if we were to consult only the rules that determine the meanings of the constituents in isolation, and the relevant operations of semantic composition.

Nunberg and his colleagues hold the view that, out of the many dimensions of idiomaticity (including compositionality), conventionality represents a necessary condition, since the meaning of an idiom cannot be predicted on the basis of the conventions governing the individual words in isolation.

The typologies described above are one-dimensional and emphasise one feature of IEs that is considered the prominent one. Other studies have made the case for a different conception of idiomaticity, conceived of more as a continuum.

2.1.3 Idiomaticity as a Continuum

Given the complexity associated with IEs, Barkema (1996) argues that approaches focusing on only one characteristic are limited, and that a multidimensional model fits such complexity in a more comprehensive manner. In his view, idiomaticity is conceived of as a scalar notion determined by three factors, i.e., compositionality, flexibility, and collocability. Collocability is defined as 'the degree to which it is possible to substitute a lexical item from an open class in a construction with alternatives from the same class: a verb by other verbs, etc.' (Barkema, 1996, p. 145). These factors are continua themselves, and the endpoints and the middle section of each continuum are identified through three categories. For instance, the flexibility continuum features 'fully flexible', 'semi-flexible' and 'flexible' constructions.

Wulff (2008), on the other hand, posits that two different continua exist: the idiomaticity continuum and the idiomatic variation continuum. The former represents what speakers construct by employing the information they

deem salient about an expression and more helpful in positioning it along such a continuum. The latter can be described as 'the range of values that constructions can actually take on with respect to their semantic and syntactic behaviour' (Wulff, 2008, p. 5). In her corpus-driven study, Wulff highlights that idiomaticity constitutes a scalar and multi-factorial notion where multiple factors come into play in shaping how idiomatic an expression is. In this way, she confirms that one-dimensional approaches are inadequate to grasp the multifaceted phenomenon of IEs.

To conclude, IEs demonstrate the complexities of natural language. To deal with such complexities, NLP has profited from the notable advances made possible by large language models (LLMs).

2.2 Large Language Models

Over the last few years, LLMs have driven remarkable advancements in NLP. At the core of this revolution lies *language modelling* (LM), which 'aims to model the generative likelihood of a word sequence, so as to predict the probabilities of future (or missing) tokens' (Zhao et al., 2023, p. 1).

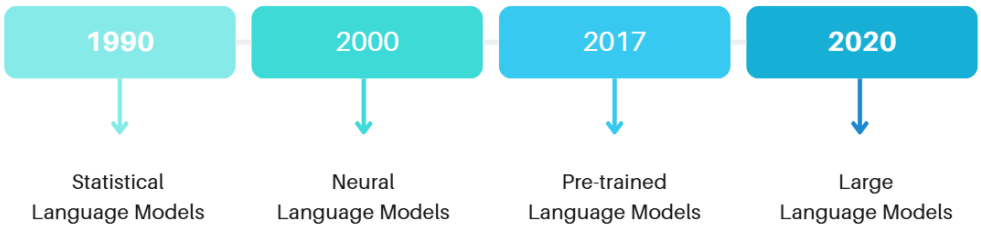

Figure 2.1: The development stages of LM.

As shown in Figure 2.1, the evolution of language modelling was inaugurated by statistical language models (SLMs), which adopt a statistical perspective. The rise of neural networks led to neural language models (NLMs), which present several advantages over SLMs, such as the ability of dealing with longer sequences. The next major development in LM was the advent of pre-trained language models (PTLMs), which are developed using self-supervised learning on large-scale text datasets. Pre-training allows

models to grasp fundamental patterns and structures of language and create a generic representation that can be applied to a wide range of NLP tasks (Naveed et al., 2023; Chu et al., 2024). The evolution then culminated in the introduction of large language models (LLMs), whose model size, data size, and training compute reach outstanding magnitude. This scaling has given rise to the emergence of new abilities, which make LLMs more powerful through reasoning, answering in zero-shot settings, instruction following, etc. (Zhao et al., 2023; Naveed et al., 2023).

2.2.1 Architecture

The majority of LLMs are based on the Transformer architecture introduced by Vaswani et al. (2017). This transduction model consists of two primary components: the encoder, which processes the input sequence to generate contextual representations, and the decoder, which uses the encoder's representations to produce the output sequence token by token. As illustrated in Figure 2.2, the encoder is built from N identical layers, each containing a multi-head self-attention mechanism and a dense feed-forward network. The decoder includes the same components, but it incorporates an additional sublayer for masked multi-head self-attention to facilitate autoregressive generation.

Self-attention enables the model to prioritise the relevant parts of the input sequence. Vaswani et al. (2017, p. 2) explain self-attention in terms of query (Q), key (K), and value (V) matrices through the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

where Q represents the word being attended to, K corresponds to all the words in a given sequence, and V holds the information associated with each word. The attention weights are obtained through the dot product between Q and K, with $\sqrt{d_k}$ serving as the scaling factor (d_k represents the dimension of Q and K). The weights are then normalised through the softmax function, and the resulting matrix is multiplied by V to obtain the output. Models like BERT draw on self-attention, whereas models like LLaMA rely on masked self-attention, which prevents them from accessing to subsequent words when predicting a given word. In order to gain a more nuanced representation, multi-head self-attention is also deployed to compute attention in parallel

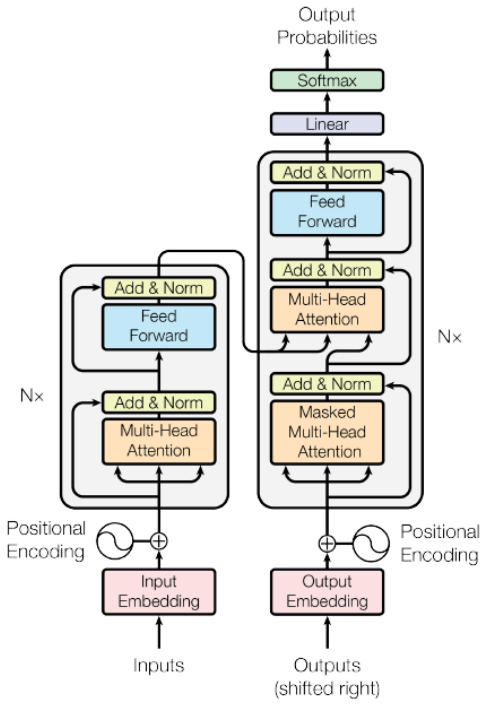

Figure 2.2: An example of the transformer model. From Vaswani et al. (2017).

across several independent heads, each attending to different patterns and features of the input.

According to the architecture, Liu et al. (2024), Zhao et al. (2023), and Naveed et al. (2023) categorise LLMs into the following types:

Encoder-decoder. It is based on the vanilla Transformer model and consists of two stacks of Transformer blocks as the encoder, which encodes the input and generates its representations, and the decoder, which autoregressively generates the target sequence.

Decoder-only. It only uses the Decoder module of the Transformer. Two sub-types can also be distinguished:

- *Causal decoder.* It performs unidirectional attention to ensure that each input token can only attend to the past tokens and itself.

- *Prefix decoder*. It performs bidirectional attention over the prefix tokens and unidirectional attention on generated tokens only.

Naveed et al. (2023) and Zhao et al. (2023) include another category: mixture-of-experts is an architecture where a subset of neural network weights are sparsely active, allowing scaling of the previous models, while maintaining the computational cost.

It is noteworthy that there is a lack of agreement on what should be encompassed under the definition of LLMs. While some definitions exclude encoder-only architectures, others, such as those proposed by Alammar and Grootendorst (2024) and Chu et al. (2024), explicitly include them. In particular, Alammar and Grootendorst (2024) draw a clear distinction between encoder-only models, which they also call 'representation models', and decoder-only models, also termed 'generative models'.

2.2.1.1 The LLaMA Model Family

An example of a causal decoder model is LLaMA (Touvron et al., 2023a), a series of models released by Meta AI in 2023. They are available to the research community upon request and are pre-trained on a large amount of public data, including English CommonCrawl[1], Project Gutenberg[2], Github[3] and ArXiv[4].

As illustrated in Figure 2.3, LLaMA retains the decoder component of the Transformer and introduces innovative aspects. For example, it employs RMSNorm (Zhang and Sennrich, 2019) for pre-normalisation. RMSNorm is a computationally efficient alternative to layer normalisation, since it does not depend on the mean and the variance but is computed from the root mean square of the input vector x. This allows for a higher training stability and more computational efficiency. As opposed to the Transformer's absolute positional embeddings, LLaMA uses rotary positional embeddings (RoPE) (Su et al., 2021), which incorporate rotation operations into the process of positional encoding to enable the model to generate dynamic positional representation during training. Another major component is grouped-query attention (Ainslie et al., 2023), where the query is divided into different group, each

[1] `https://commoncrawl.org/`
[2] `https://www.gutenberg.org/`
[3] `https://github.com/`
[4] `https://arxiv.org/`

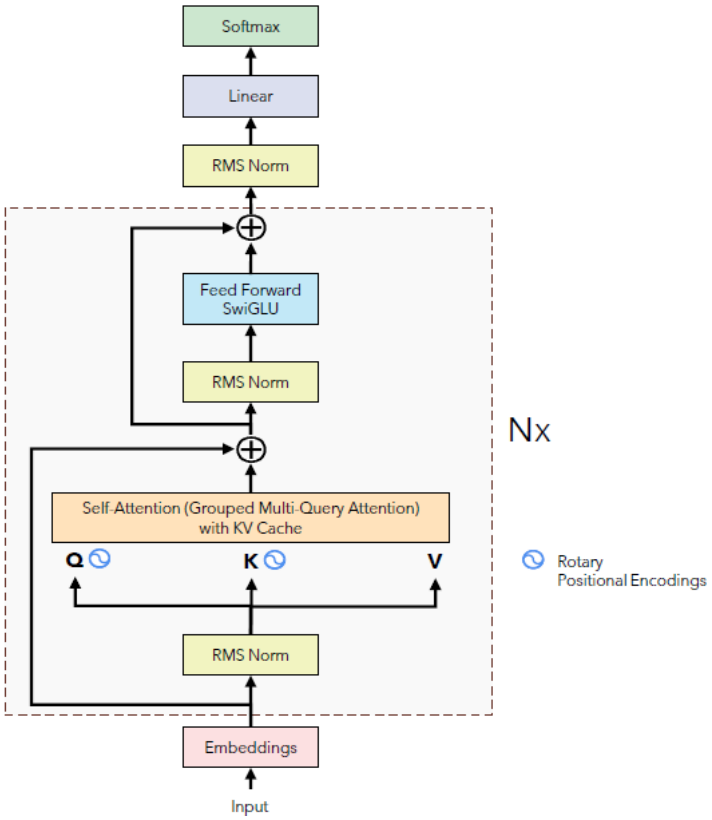

Figure 2.3: LLaMA's architecture. From Jamil (2023).

associated with one set of keys and values. This allows for fewer computations and faster inference. Finally, LLaMA uses SwiGLU activation function (Shazeer, 2020), which has proved to perform well on various benchmarks.

Given its success, two follow-up versions were released, with the latest being LLaMA 3 (Dubey et al., 2024), a multilingual and multimodal variant. Compared to the first version, small modifications have been applied to the architecture. The key enhancements of the latest release lies in the higher quality and diversity of pre-training data and the increase in the training scale. Additionally, LLaMA 3 officially supports eight languages, including Italian and Portuguese.

2.2.2 A Two-Step Paradigm: Pre-Train and Post-Train

As described by Alammar and Grootendorst (2024), the training paradigm of LLMs is twofold. The first step is pre-training, where the model is trained on a large amount of natural language data retrieved from a more or less diverse mixture of datasets. These datasets can be more general, including webpages and conversational text, or specialised, containing multilingual data[5] or texts belonging to specialised domains (Zhao et al., 2023). The pre-training data undergo a multi-stage preprocessing that aims at obtaining high quality and safety. Both the diversity and the quality of data constitute key factors in shaping the model's performance (Zhao et al., 2023; Liu et al., 2024).

Since the pre-training objective is mainly language modelling, pre-trained LLMs can undergo a post-training through two main strategies, alignment with human preferences and instruction fine-tuning (Zhao et al., 2023; Naveed et al., 2023). The former aims to avert unintended behaviours that LLMs might exhibit and that do not align to human values, such as helpfulness, honesty, and harmlessness (Askell et al., 2021). On the other hand, remarkable performance improvements are brought about by instruction fine-tuning. In the following section, I provide an overview of instruction fine-tuning, with a particular focus on instructions, which serve as the foundation for the creation of my dataset.

2.2.2.1 Instruction Tuning

Instruction tuning (IT), also referred to as supervised fine-tuning (SFT), consists in fine-tuning LLMs on instruction-output pairs, where the instruction is in the form of natural language and the output represents the desired output obtained from following the instruction (Zhang et al., 2023, p. 1). IT is a new supervision seeking paradigm that aims to boost LLMs' generalisation ability and controllability, allowing for a better performance on unseen tasks and a more predictable behaviour. In doing so, IT qualifies as a more user-oriented approach that attempts at bridging the gap between users' needs and the pre-training objective of next token prediction.

Studies diverge in defining instruction: Zhang et al. (2023) and Zhao et al.

[5]I here abide by Zhao et al. (2023) in their classification; however, I do not agree with the inclusion of multilingual data as specialised data.

(2023) make a distinction between instruction, input, and output, while Lou et al. (2024) understands instructions as comprising:

Input X : the input, such as a piece of text. It is optional.

Output Y : the desired output.

Template T : a textual template explaining the task intent or the relationship between X and Y.

The instruction data construction plays a crucial role and aims at creating an optimal dataset of instruction-formatted instances. Datasets can be created either manually or synthetically. Human-crafted data are retrieved from online sources or already annotated data and usually result in smaller datasets. A less time-consuming method is the synthetic creation, which leverages LLMs to generate datasets. While being more economic, they can produce less diverse and heterogenous instruction-formatted instances, which might hinder the models' performance. Another categorisation is proposed by Lou et al. (2024), where different instruction types are distinguished depending on the combinations of X, Y, and T:

NLI-oriented instructions (I=T+Y) : they combine a template T with a label Y to explain the task semantics. As exemplified in Table 2.1, tasks are converted into natural language inference (NLI), where labels are turned into hypotheses, whose truth has to be determined. This approach preserves the task semantics and encode the relationship between input and output.

Task	**NLI premise (the input text)**	**NLI hypothesis**
Entity Typing	[Donald Trump] $_{\text{ent}}$ served as the 45th president of the United States from 2017 to 2021.	Donald Trump is a **politician**. Donald Trump is a **journalist**.

Table 2.1: Example of NLI-oriented instruction, where the hypothesis is used to explain the labels (in bold). The label highlighted in green is correct. From Lou et al. (2024).

LLM-oriented Instructions (I=T+X) : this is the combination of template T and input X and is usually employed in the form of prompts. Table 2.2 illustrates two different: the prefix prompt, where the input is prepended within the instruction, or cloze prompt, which takes the form of a cloze-question template. These formats are designed to adhere to two distinct pre-training objectives: prefix prompt fits the autoregressive nature of decoder-based models, such as LLaMA, while cloze prompt mirrors the masked language modelling of encoder-based models, such as BERT (Liu et al., 2022). Being more LLM-orientend, such a type lacks user-friendliness and requires knowledge that users might not own; its structure is also short and simplistic and does not lend itself to more elaborate tasks.

Task	**Input X**	**Template**	**Answer**	**Output Y**
Sentiment Analysis	I would like to buy it again.	[X] The product is _.	Great Wonderful ...	Positive
Entity Tagging	[Donald Trump] $_{\text{ent}}$ served as the 45th president of the United States from 2017 to 2021.	The entity in [X] is a _ class?	Politician President ...	People

Table 2.2: Examples of LLM-oriented instructions. Adapted from Lou et al. (2024).

Human-oriented Instructions (I=T+ optional $\{X_i, Y_i\}_{i=1}^{k}$) : as shown in Table 2.3, this type is more descriptive and human-readable and is able to handle more complex tasks. Nevertheless, their encoding might be challenging, due to their complex nature.

Data quality and diversity are the cornerstone of an effective IT. As Lou et al. (2024) and Zhao et al. (2023) point out, a higher diversity in terms of writing style and perspective can enhance the generalisation ability, even in smaller models. Consistency in instruction type across training and test can also benefit the performance.

Task	Input X	Template	Output Y
Sentiment Analysis	I am extremely impressed with its good performance.	*Task Definition:* In this task, you are given a product review, and you need to identify ... *Test Instance:* Input: [X] Output: _	Positive

Table 2.3: Example of human-oriented instructions. Adapted from Lou et al. (2024).

Zhao et al. (2023) underscores the beneficial effects of IT on performance: instruction fine-tuned models, even smaller ones, display improved abilities in seen and unseen tasks and in zero-shot scenarios, since IT endows the models with the ability of following human instructions, regardless of the use of demonstrations. Multilingual scenarios can also benefit from IT, even though further research is needed. In particular, the lack of parallel multilingual instruction datasets prevents us from exploring the effects of IT in settings with multiple languages. Besides, few studies investigate the impact of the instruction language on the models' performance and provide contradictory results. Muennighoff et al. (2023) argues that English-only instructions can produce satisfactory results on multilingual tasks. On the other hand, Phelps et al. (2024) concentrate specifically on idiom processing and find that, if the instruction language is the same as the input language, the models exhibit a better performance.

In conclusion, IT can improve LLMs' performance on various tasks and in a multilingual setting. To fully unlock IT's potential, it is essential to design the instruction data creation process carefully. Instruction type and language also need to be taken into account, even though the impact of the instruction language remains unexplored.

2.3 Related Work on Idiomaticity in NLP

In the field of NLP, IEs are conceived of as MWEs, i.e. expressions that cross word boundaries and have idiosyncratic interpretations (Sag et al.,

2002; Villavicencio et al., 2005; Baldwin and Kim, 2010). MWEs pervade language, and their presence in a speaker's lexicon is comparable in magnitude to that of single words (Jackendoff, 1997, p. 156). Such a presence entails a high degree of complexity, which poses considerable challenges to NLP, including what Sag et al. (2002, p. 2) terms 'idiomaticity problem'. The idiomaticity problem refers to the issues related to those MWEs whose meaning is unrelated to the meanings of their component words. Baldwin and Kim (2010, p. 269) extend idiomaticity to multiple levels of language and define it as 'markedness or deviation from the basic properties of the component lexemes, [which] applies at the lexical, syntactic, semantic, pragmatic, and/or statistical levels'. They also view idiomaticity as a necessary feature of MWEs, which must exhibit some sort of deviation on at least one level.

The above-mentioned studies agree on the fact that the size and heterogeneity of the MWEs class require *ad hoc* NLP techniques for a better understanding of natural language and a more linguistically precise NLP. Challenges posed by the presence of IEs have been identified across multiple natural language understanding (NLU) tasks even with state-of-the-art (SOTA) solutions, including sentiment analysis (Liu et al., 2017; Biddle et al., 2020), paraphrase generation (Zhou et al., 2022), natural language inference (Chakrabarty et al., 2021), dialog models (Jhamtani et al., 2021), and machine translation (Fadaee et al., 2018; Dankers et al., 2022; Liu et al., 2023).

One line of work employs encoder-based models, particularly BERT (Devlin et al., 2019) and its variants, to harness their capability for contextual language representation. Previous work has found that contextual embedding models fail to capture non-compositionality and struggle to distinguish between literal and idiomatic usages of MWEs (García et al., 2021; Hashempour and Villavicencio, 2020; Nandakumar et al., 2019; Yu and Ettinger, 2020). For instance, Yu and Ettinger (2020) explore the notion of semantic compositionality and take steps to examine how faithfully contextual embedding models represent phrases and the individual words used in isolation. The authors test five encoder-based models, including BERT, RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2020) and conclude that such models exhibit weak sensitivity to composed meaning, heavily relying on word senses.

To capture both the semantic and syntactic properties of idioms, Zeng and Bhat (2021) propose the iDentifier of Idiomatic expressions via Semantic Compatibility (DISC). DISC leverages BERT to obtain the contextualised representation of idioms and extract semantic information. Zeng and Bhat

test their model on multiple state-of-the-art baselines and find that DISC tends to outperform them, especially in zero-shot settings, where training set and test set contain different idioms. The combination between syntactic and semantic properties allows the model to get a sense of what an idiom is and generalise well on unseen expressions. However, when tested on a dataset different from the training one, DISC exhibits a performance drop due to the change of the sentence source, showing poor cross-domain performance. Moreover, DISC performs well on various idiom types with varying degrees of variability, but it is trained on English data only.

Other approaches also address multilinguality. Tayyar Madabushi et al. (2021) aim to investigate the performance of encoder-based PTLMs on idiomaticity detection and representation in a multilingual setting. They release AStitchinLanguageModels, a dataset in English and Portuguese, and provide baselines by implementing BERT, XML-RoBERTA (Conneau et al., 2020), and XLNET (Yang et al., 2020) models. Focusing on the idiomaticity detection task, they conduct the experiments and explore how their performance varies with:

- Different input features: the input might include context (the previous and the next sentence), and the MWE can be included and marked by the [SEP] token.

- Three setups: zero-shot, one-shot, and few-shot scenarios.

They find that context does not lead to significantly improved performance, while the inclusion of the MWE is beneficial to the model performance. Finally, the experiments produce poor results in zero-shot setting, especially in Portuguese, suggesting that the models might fail to generalise to unseen expressions. As far as the languages are concerned, the results in English outperform those in Portuguese. The authors suggest that both the pre-training data and their training set contain significantly less Portuguese data, and the higher degree of inflection might also play a role. These findings highlight the need, as Tayyar Madabushi and colleagues argue, to examine idiomaticity from a multilingual perspective.

AStitchInLanguageModels is further extended with Galician data in the SemEval-2022 Task 2 on multilingual idiomaticity detection and sentence embedding (Tayyar Madabushi et al., 2022). The top teams succeeded in narrowing the gap between zero-shot results and one-shot results, even though the former are still outperformed. Among the top three best performing

teams, Chu et al. (2022) use large-scale cross-lingual pre-trained language models, i.e., multilingual BERT and XLM-RoBERTa. They attempt to improve the results by incorporating data augmentation and contrastive learning. While data augmentation proves useful under zero-shot setting, contrastive learning leads to a performance drop in both scenarios. Yet this study relies on the SemEval-2022 Task 2 dataset, which only comprises noun compounds and lacks diversity in terms of idiom types.

A notable effort to explore idiom identification in multiple languages is made by Tedeschi et al. (2022). They release ID10M, a framework consisting of systems, training and validation data, and benchmarks for the identification of idioms in 10 languages. Their findings confirm the discrepancy between zero-shot and few-shot performance.

Other approaches implement encoder-decoder models to develop idiom-aware systems. For instance, Zeng and Bhat (2022) point out that contextual embedding models hinge on a compositional paradigm of representation and struggle to capture non-compositionality. Consequently, they attempt to build idiomaticity into the BART (Lewis et al., 2020) sequence-to-sequence (seq2seq) model, which combines the encoder and the decoder and can capture contextual cues while also being able to generate text. The resulting model, called Generation of Idiom Embedding with Adapter (GIEA), shows an improved understanding of idiomaticity compared to BART base. Even though they contribute to improving idiom representations, their work is restricted to English data, and there is no attempt at improving GIEA's generalisation ability to idioms unseen during training.

In the FigLang2022 Shared Task[6], Bigoulaeva et al. (2022) implement another seq2seq model, specifically T5 (Raffel et al., 2020). FigLang2022 proposes a NLI task that includes figurative-language hypothesis and requires participants to generate a textual explanation. To tackle both, Bigoulaeva and her collaborators leverage transfer learning and train the models on eSNLI (Camburu et al., 2018), a dataset comprising NLI labels and their relative explanations, and IMPLI (Stowe et al., 2022), a NLI dataset of figurative language. While demonstrating the usefulness of transfer learning, this work focuses exclusively on English and limits the examination of figurative language to the context of the NLI task.

Recent studies have also investigated the capability of LLMs to handle idiomatic expressions and conclude that LLMs are outperformed by other

[6]`https://figlang2022sharedtask.github.io`

methods, such as encoder-based models. Starting from data in English, Portuguese, and Galician, He et al. (2024) compare their encoder-based model to other methods, including LLaMA 2 (Touvron et al., 2023b), showing that LLaMA 2 tends to perform worse. On the same languages, Phelps et al. (2024) compare the performance of decoder-based models with that of encoder-based ones and find that the latter outperform LLMs in idiom-related tasks, with LLaMA 2 ranking last among LLMs. Finally, De Luca Fornaciari et al. (2024) employ fine-tuned LLMs, such as Llama-2-7b-chat, and highlight how such models still struggle on idiom-related tasks.

We can conclude that while several studies have adopted encoder-based models or encoder-decoder architectures to investigate idiomaticity, they often fall short in addressing the full spectrum of idiom types, their variability, or the range of languages considered. Conversely, research on the performance of LLMs in handling idiomatic expressions remains relatively sparse. In cases where LLMs are examined, these studies lack a targeted fine-tuning specifically designed to improve their idiom comprehension.

Consequently, this thesis contributes to bridging these gaps by improving our understanding of how decoder-based models perform when dealing with idiomatic expressions, particularly in a multilingual context where language variability adds another layer of complexity. To achieve this, some of the aforementioned datasets are used to construct the multilingual instruction dataset for IEs. Specifically, AStitchInLanguageModels and ID10M are employed to extract IEs and sentences in English and Portuguese. Additionally, a third annotated corpus, MultiCoPIE (Sentsova et al., 2025), is combined with ID10M to retrieve IEs and sentence samples in Italian. Based on the final instruction dataset, this thesis introduces a focused fine-tuning approach aimed at equipping these models with enhanced capabilities to interpret idioms effectively, providing a more comprehensive and nuanced evaluation of their potential in this domain.

Chapter 3

Instruction Dataset Construction

Chapter 3 is concerned with the creation of the instruction dataset[1]., i.e. a dataset comprising examples in the form of instructions. Section 3.1 describes the source annotated datasets from which multi-word expressions (MWEs) and sentences containing them were extracted. Section 3.2 details the creation of instruction-formatted templates and the subsequent compilation of the final dataset.

3.1 Source Datasets

This section describes AStitchInLanguageModels (Tayyar Madabushi et al., 2021), ID10M (Tedeschi et al., 2022), and MultiCoPIE (Sentsova et al., 2025), the three source annotated datasets used for the extraction of input sentences and the corresponding MWEs. The description of each dataset focuses on three main aspects relevant to this study: the covered languages, the types of MWEs included, and the annotation scheme employed.

[1]For clarification, 'examples' and 'samples' refer to the individual data points within the dataset that are in the form of instructions, while 'template' denotes the structured format of instruction-based examples, as described in Section 2.2.2.1

3.1.1 AStitchInLanguageModels

AStitchInLanguageModels (Tayyar Madabushi et al., 2021) is an annotated dataset of idiomatic MWE usage in English and Portuguese. It consists of examples that contain potentially idiomatic expressions (PIEs) in the form of noun compounds and that are annotated according to two different schemes. In the first one, the examples featuring the MWEs are labelled with either 0 (for idiomatic usage) or 1 (for literal usage). In contrast, the second, more fine-grained annotation framework includes a paraphrase of the MWE's meaning and classifies each example into one of five categories: 'literal,' 'idiomatic,' 'non-idiomatic,' 'proper noun,' or 'meta usage'.

This thesis employs data labelled with the first annotation scheme and designed for a zero-shot scenario, where the training set does not share any PIEs with the development and test sets. Table 3.1 illustrates four English examples, where 'Sentence1' refers to the sentence containing the IE, 'Sentence2' represents the MWE, and 'Label' indicates whether the meaning of the sentence is idiomatic (0) or literal (1).

Label	**Sentence1**	**Sentence2**
0	Turns out that these people were speaking double Dutch.	double dutch
0	She casts herself as an honorable champion of conservative thinking against the 'intolerant liberal views' of the Ivory Tower.	ivory tower
1	Michigan has a lot of fresh water lakes.	fresh water
1	In reality, he would probably use a silver spoon to pot a plant.	silver spoon

Table 3.1: English examples from the AStitchInLanguageModels dataset.

As shown in Table 3.2, the English subset is well-balanced, with 51% of the examples classified as literal and 49% as idiomatic. In contrast, Table 3.3 indicates that the Portuguese dataset is skewed towards idiomatic usage, which accounts for 64% of the total examples. Additionally, AStitchInLanguageModels presents certain limitations, particularly in terms of idiom type, restricted to noun compounds, and dataset size, both in terms of the number of MWEs (223 for English and 113 for Portuguese) and the total number of examples (4,276 for English and 1,716 for Portuguese). Despite these con-

Set	MWEs	Literal Examples	Idiomatic Examples	Total Examples
train	163	1,565	1,762	3,327
dev	30	284	182	466
test	30	334	149	483
total	223	2,183	2,093	**4,276**

Table 3.2: Statistics of the English subset of AStitchInLanguageModels.

Set	MWEs	Literal Examples	Idiomatic Examples	Total Examples
train	73	391	773	1,164
dev	20	119	154	273
test	20	114	165	279
total	113	624	1,092	1,716

Table 3.3: Statistics of the Portuguese subset of AStitchInLanguageModels.

straints, this dataset represents the first significant attempt to address idiom processing from a multilingual perspective, and it provides a benchmark for a language other from English, thereby enabling further examination of idiom processing and multilinguality.

3.1.2 ID10M

ID10M (Tedeschi et al., 2022) is a framework that proposes a multilingual Transformer-based architecture for sentence disambiguation and idiom identification and introduces annotated datasets in multiple languages. In particular, it presents gold-standard data, i.e. a manually annotated dataset, in English, German, Italian, and Spanish, and silver-standard data, which are automatically annotated in 10 languages (Chinese, Dutch, English, French, German, Italian, Japanese, Polish, Portuguese, and Spanish). MWEs were extracted using Wiktionary[2], and instances containing MWEs were collected from WikiMatrix (Schwenk et al., 2021), a multilingual corpus in 83 languages consisting of parallel sentences retrieved from Wikipedia[3]. The gold-standard data were curated by mother-tongue professional annotators who

[2] `https://pypi.org/project/wiktextract/`

[3] `https://github.com/facebookresearch/LASER/tree/main/tasks/WikiMatrix`

were tasked with tagging MWE-sentence pairs as 'idiomatic' or 'literal'. The silver-standard data, on the other hand, were automatically annotated on the basis of the Wiktionary entry of MWEs: when marked as idiomatic, all occurrences of the MWEs were tagged as idiomatic, and vice versa. Since these annotations do not necessarily reflect the idiomatic or literal meaning of the MWE in context, Tedeschi and his collaborators introduce a dual-encoder architecture to further refine silver-standard data.

Building on this binary annotation system, they incorporate a BIO tagging scheme, a token-level tagging scheme identifying the tokens belonging to the MWE. As exemplified in Table 3.4, *B* designates the first token of a span, *I* is associated with the intermediate token(s), and *O* signals the tokens out of the span.

Token	**Label**
The	O
test	O
was	O
a	O
piece	B-IDIOM
of	I
cake	I
!	O

Table 3.4: Example of the BIO scheme.

In constructing the instruction dataset, this study uses silver-standard data for English, Italian, and Portuguese, alongside gold-standard data for English and Italian. As shown in Table 3.5, there is an imbalance in the distribution of MWEs across the languages included in the dataset, and the datasets contain significantly more literal examples than idiomatic ones, despite the primary focus being on idiomaticity. Nevertheless, ID10M provides a considerable amount of annotated data across a diverse set of languages, is not restricted to specific idiom types, and incorporates linguistic variations.

To compensate for this imbalance, MultiCoPIE is incorporated, which, in contrast, exhibits an imbalance towards the idiomatic class and enables the inclusion of additional idiomatic examples.

3.1.3 MultiCoPIE

MultiCoPIE (Sentsova et al., 2025) is a dataset annotated for sentence disambiguation and idiom identification in Russian, Italian, and Catalan. Since I contributed to the construction of the Italian dataset (which is also used in

Language	Set	MWEs	Literal	Idiomatic	Total
EN	train + dev	4,568	27,408	10,501	37,919
	test	142	41	159	200
	total	4,609	27,449	10,670	38,119
PT	train + dev	559	24,816	10,670	30,492
	test	/	/	/	/
	total	559	24,816	10,670	30,942
IT	train + dev	452	20,506	9,107	29,523
	test	78	48	152	200
	total	530	20,554	9,259	29,813

Table 3.5: Statistics of the English, Portuguese, and Italian subsets of ID10M.

this thesis), the following description focuses on this particular dataset.

To build this subset, I compilted a list of PIEs from online resources, including the Dizionario italiano De Mauro[4] and the Dizionario dei Modi di Dire[5]. Efforts were made at including PIEs with varying characteristics. Specifically, PIEs with different parts of speech as heads were incorporated. For instance, *appeso a un filo* ('hung by a thread') is headed by the adjective *appeso*('hung'), 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 varying degrees of semantic compositionality. PIEs with a higher level of compositionality comprise at least one component serving as a cue to the meaning of the expression. An example is *ammazzare il tempo*('to kill time'), where the word *tempo*('time') aids in 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 to the meaning 'to be at the peak of happiness'. A distinction is also made between PIEs and IEs: the former refer to expressions which can acquire both a literal and an idiomatic meaning, while the latter exclusively occur in an idiomatic sense.

Finally, two further categories are included for each expression: whether there is a partially or entirely equivalent English expression, and the possible variations in which the PIE can occur. This categorisation is illustrated in Table 3.6. The feature 'Ambiguity' indicates whether a MWE has the po-

[4]https://dizionario.internazionale.it/

[5]https://dizionari.corriere.it/dizionario-modi-di-dire/

Idiom	**Head Pos**	**Sem. Comp.**	**Variation**	**Ambiguity**	**Equ.**
rompere il ghiacchio ('to break the ice')	Verb	True Idiom	/	True	Identical
chiedere la luna ('to ask for the moon', meaning 'to want or ask for something impossible')	Verb	Weak Idiom	*volere la luna*	False	/
gallina dalle uova d'oro ('golden goose')	Noun	True Idiom	/	False	Similar

Table 3.6: Examples of the annotations related to PIEs in MUltiCoPIE.

tential to be interpreted literally ('True') or not ('False'), while the labels 'Identical' and 'Similar' denote the presence of an equivalent English expression, with the former indicating a direct match and the latter suggesting minor formal variations. Table 3.7 reports some statistics related to such categories. The majority of MWEs extracted have a verb as head. Besides, most MWEs are true idiom, and their meaning is opaque in terms of semantic compositionality. Regarding variation, 68% are also used with variations, such as the example *chiedere la luna* in Table 3.6. As far as ambiguity is concerned, 68% of the extracted MWEs can be categorised as PIEs (the idiomatic or literal usage depends on the context, while the remaining 32% qualify as IEs, used only in the idiomatic sense. Finally, there is a balance between MWEs having an English equivalent (either identical or similari) and MWEs without an English equivalent.

After compiling a list of PIEs, examples were automatically extracted from the OSCAR corpus[6] (Ortiz Suárez et al., 2019), a multilingual corpus generated from Common Crawl[7], and subsequently refined through manual

[6] `https://huggingface.co/oscar-corpus`
[7] `https://commoncrawl.org/`

Head POS		Sem. Transp.		Variation		Ambiguity		Eng. equiv.	
Adj	4	True Idiom	91	Yes	36	Yes	36	No	57
Adv	2	Weak Idiom	20	No	75	No	75	Identical	35
Conj	5							Similar	19
Noun	19								
Prep	27								
Verb	56								

Table 3.7: Properties of idioms according to head part-of-speech, semantic transparency, head part of speech.

selection. To maximise coverage, regular expressions were employed to capture a wide range of linguistic variations. Where available, the two surrounding sentences were included to provide context. The opening and closing tags were also employed to precisely localise the lexicalised components of PIEs, as illustrated in Table 3.8. The tags were used to identify not only the PIEs under study, but all PIEs present in the target sentence and the preceding and following sentences.

The final Italian subset comprises 2,245 total examples:1,887 (84%) instances are labelled as idiomatic, while 358 (16%) are marked as literal.

Idiom	Previous sent.	Target Sent.	Next sent.	Label
tirare la cinghia	Sempre per i più esperti è possibile aggiunge più lavoro ai muscoli flessori...	<idiom2>**Mano a mano**</idiom2>che la posizione si scioglie <idiom>**stringere la cinghia**</idiom>.	È possibile prendersi le mani in Full Swan...	0

Table 3.8: Examples from the MultiCoPIE Italian data.

3.2 Creation of Instruction-Formatted Instances

This section outlines the process of constructing the instruction dataset for the fine-tuning of LLaMA 3.2. Specifically, I first describe the creation of the templates structured into 'instruction', 'input' and 'output'. Then, I

illustrate the pipeline used to extract MWEs and sentences from the source datasets and how these were integrated with the templates to form the final dataset.

3.2.1 Creation of Templates

To create a dataset of instruction-formatted instances, I first designed instructions in English, Italian, and Portuguese. As shown in Figure 3.1, the process began with the translation of a seed instruction written in English into Italian and Portuguese by using LLaMA 3.2 3B[8] via ollama[9]. Secondly, by prompting the same model, I generated three paraphrased versions of each instruction. The prompts used were tailored to produce varying writing styles and perspectives, ensuring a varied dataset and a higher linguistic diversity.

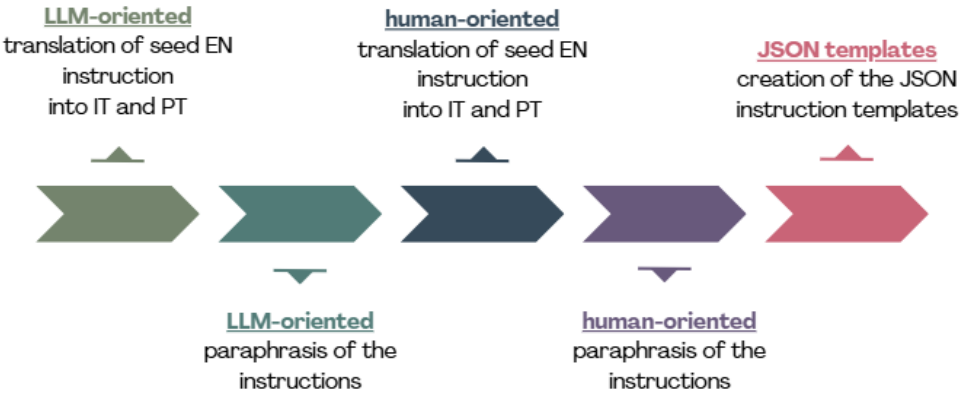

Figure 3.1: Stages of template creation.

I implemented this pipeline for both human-oriented (H) instructions and LLM-oriented (L) instructions (see Section 2.2.2.1). Within the human-oriented category, I further classified the instructions into three subtypes: I (the proper instruction), Y (positive response), and N (negative response). For example, an instruction I would ask whether a sentence contains an idiomatic expression, while the corresponding response Y would confirm its presence and specify the idiom, and the response N would indicate that no

[8] https://huggingface.co/meta-llama/Llama-3.2-3B
[9] https://ollama.dipintra.it/

idiom is present. Such categorisation was not provided for the LLM-oriented type since the response is the IE, when positive, or 'None', 'Nessuna', and 'Nenhuma', when negative. All the instructions were collected into a CSV file, where each row represents a predefined instruction paired with its language and category. Finally, I generated JSON templates based on this CSV file. The starting point to construct such templates is the work by (Taori et al., 2023), who a model fine-tuned from LLaMA 7B on instruction-formatted demonstrations. They design a template in English used to create the instruction-formatted examples to carry out the fine-tuning. Their template serves as the starting point to craft multilingual templates that enable the construction of LLM-oriented instructions. Table 3.9 illustrates the Alpaca template and the multilingual templates designed in this thesis. First, the template was translated into Italian and Portuguese, so that instruction-formatted samples in these two languages could have a corresponding template. Second, the 'prompt_no_input' option was discared since all my samples include an input (i.e. the input sentence). Finally, the structure of the template was changed to match the LLM-oriented instruction type. While the Alpaca template organises the instruction in "Instruction", "Input", and "Response", I changed the order so that the input is first presented, followed by the instruction and the response. This order better fits the LLM-oriented type, which features an empty slot at the end of the instruction that needs to be filled by the LLM. This structure is matched by having the instruction immediately followed by the response. The 'input' and 'output' keys are left empty to be filled in the following step.

	Alpaca template
	prompt_input: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ###Instruction: {instruction} ###Input: {input} ###Response: **prompt_no_input**: Below is an instruction that describes a task. Write a response that appropriately completes the request. ###Instruction: {instruction} ###Response: **response_split**: ###Response:
	Multilingual templates
EN template	**lang**: en **prompt_input**: Below is an input sentence, paired with an instruction that describes a task. Write a response that appropriately completes the request. ### Input: {input} ### Instruction: {instruction} ### Response: **response_split**: ### Response:
IT template	**lang**: it **prompt_input**: Di seguito si trova una frase di input, associata a un'istruzione che descrive una task. Scrivi una risposta che completi in modo appropriato la richiesta. ### Istruzione: {instruction} ### Input: {input} ### Risposta: **response_split**: ### Risposta:
PT template	**lang**: pt **prompt_input**: Abaixo está uma frase de entrada, acompanhada de instruções que descrevem uma tarefa. Escreva uma resposta que complete adequadamente a solicitação.### Instrução: {instruction} ### Input: {input} ### Resposta: **response_split**:### Resposta:

Table 3.9: Alpaca's template and the multilingual templates used in this thesis.

3.2.2 Creation of the Final Instruction Dataset

Once the templates with instructions were available, I proceeded with the creation of the final dataset, whose stages are summarised in Figure 3.2.

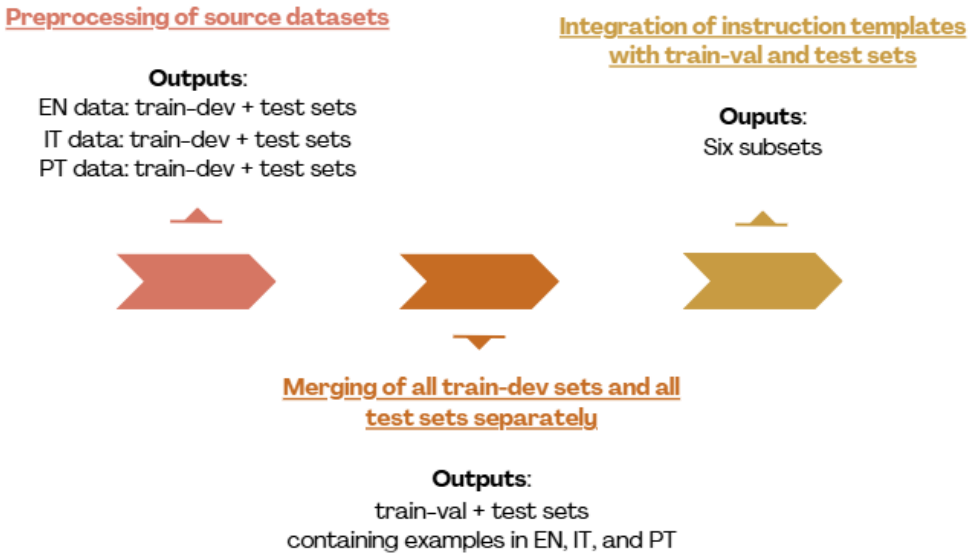

Figure 3.2: Stages of instruction dataset creation.

First, I extracted IEs and examples from the aforementioned datasets. For English and Portuguese, I used ID10M and AStitchInLanguageModels, while, for Italian, I employed ID10M and MultiCoPIE. I processed ID10M's files by reconstructing full sentences and identifying idiomatic spans. On the other hand, the processing of the AStitchInLanguageModel mainly focused on extracting the actual MWEs present in the sentences since the 'Sentence2' include the lemmatised version. Then, I created a training-development[10] and test split combining data from both ID10M and AStitchInLanguage-Models, while ensuring that no PIEs in the test set overlapped with those in the training-development data. Finally, I applied text cleaning operations, such as fixing contractions and punctuation spacing and exported two final processed splits per language, train-dev and test.

[10]Here, 'training-development' is used to indicate the data extracted from each source dataset. On the contrary, I use the term 'training-validation' set to specifically designate the training and validation sets of the final dataset I constructed for this thesis.

For the MultiCoPIE data, I extracted sentences and the relative PIEs enclosed in annotation tags to obtain the non-lemmatised versions. Next, I combined these data with ID10M's Italian data and balanced the whole dataset by undersampling literal instances and splitting into training-development and test sets, while preventing PIEs overlap between them. Finally, text cleaning is applied to improve consistency, and the train-dev and test sets are saved. Once I extracted the IEs and the sentences for all three languages, I merged all sets into unified training-development and test datasets containing instances from English, Italian, and Portuguese.

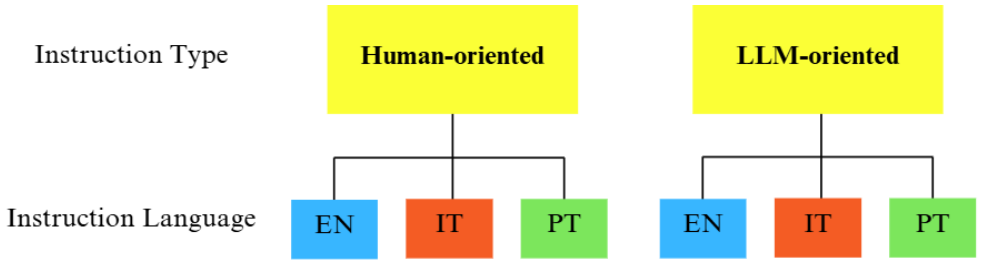

Figure 3.3: An overview of the dataset structure.

I then populated the JSON instruction templates with such examples. As illustrated in Figure 3.3, the final dataset comprises six subsets differing in instruction type and instruction language, while each containing English, Italian, and Portuguese as input languages.

The statistics of each instruction subset are consistent across all subsets, ensuring uniformity in the dataset. Table 3.10 provides an overview of the distribution and characteristics of each instruction subset, showing that this is well-balanced in terms of the number of the examples per class. An extract of the final dataset is shown in Appendix A, which illustrates instances from the subset with English as instruction language and LLM-oriented as instruction type.

Language of Examples	Set	Idioms	Idiomatic Examples	Literal Examples	Total Examples
EN	train+val	4,294	10,523	10,553	21,076
	test	840	1,892	1,862	3,754
IT	train+val	1,498	9,661	9,350	19,011
	test	275	1,338	1,649	2,987
PT	train+val	577	5,848	5,481	11,329
	test	120	600	967	1,567
Total	train+val	6,369	26,032	25384	51,386
	test	1,235	3,830	4,478	8,308

Table 3.10: Statistics of an instruction subset.

Chapter 4

Experiments

In this chapter, I first provide definitions for the two tasks addressed in this thesis: sentence disambiguation and idiom identification. Next, the evaluation framework is outlined, detailing the metrics used to assess the model's performance. The experiment[1] setting is then detailed by introducing QLoRA (Dettmers et al., 2023), a parameter-efficient technique used to reduce computational cost and memory usage during the fine-tuning. Additionally, the specific hyperparameters employed are reported. Finally, the results are analysed and discussed to address the research questions outlined in Chapter 1.

4.1 Task Definition

This thesis addresses two tasks, sentence disambiguation and idiom identification, across three languages: English, Italian, and Portuguese.

Task 1: Sentence disambiguation: this task is framed as a binary text classification problem aiming to classify a given sentence as either literal (labelled as '0') or idiomatic (labelled as '1').

Task 2: Idiom Identification: in this task, the objective is to identify the IE contained in sentences labelled as 'idiomatic'. This is framed as a span identification problem, where the model is tasked with identifying

[1]The code is available at `https://github.com/TinfFoil/MultIdiomLlama`

the sequence of characters that correspond to the IE. The approach relies on character-level overlap between the predicted span and the gold span corresponding to the IE. Partial matches are considered valid, meaning that even if the model identifies part of the idiomatic expression, it is still credited with identifying the idiom correctly.

These tasks are distinct but strictly interconnected: once the model recognises a sentence as idiomatic, it can identify the span constituting the IE. Given their interdependence, the instruction-formatted data are designed to address both tasks simultaneously. For example, in the sentence 'She broke the ice with a funny joke', the model's answer is expected to be 'broke the ice', demonstrating that the model correctly identified the sentence as idiomatic and proceeded to detect the span where the potentially idiomatic expression (PIE) occurs. To account for both tasks, I propose a two-fold evaluation methodology that includes metrics to assess the model's performance on Task 1 and Task 2. This approach allows for a comprehensive understanding of the model's ability to handle both the classification and identification challenges.

4.2 Evaluation Framework

This section outlines an evaluation framework designed to assess the model's performance on the sentence disambiguation and idiom identification tasks across various language combinations.

4.2.1 Task 1: Sentence Disambiguation

For Task 1, I developed 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:

$$\text{label} = \begin{cases} 0 & \text{if keywords like 'nenhuma', 'none', 'no' are present} \\ 1 & \text{otherwise} \end{cases} \tag{4.1}$$

According to the assigned labels, the binary classification metrics are computed as follows.

Precision. It measures the proportion of correctly predicted positive instances out of all instances predicted as positive. It indicates how reliable the model is when it predicts a positive class.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4.2}$$

Recall. It evaluates the proportion of correctly predicted positive instances out of all actual positive instances in the dataset. This metric focuses on how well the model identifies all the true positives, emphasising its ability to avoid false negatives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4.3}$$

F1 score. It is the harmonic mean of precision and recall and provides a balance between these two metrics. It is particularly useful when there is a class imbalance, as it considers both false positives and false negatives. The F1 score ranges from 0 to 1, where a higher value indicates better performance.

$$F1 \text{ Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

4.2.2 Task 2: Idiom Identification

For Task 2, the evaluation approach uses partial text span matching. The methodology follows the approach proposed by Da San Martino et al. (2020), who gives credit to partial matches of the identified spans, rather than requiring an exact match between the predicted span and the ground truth.

For instance, if the model correctly identifies a portion of the idiomatic expression but not the entire span, it will still be credited for the overlap. This approach allows for a more flexible assessment of the model's performance.

The span match is character-based and is computed through the Longest Common Subsequence (LCS). The goal of LCS is to find the longest subsequence that two sequences have in common, without changing the order of characters. LCS is thus order-sensitive and finds subsequences of characters that retain the same order between two sequences. However, the characters in the subsequence are not necessarily contiguous. This is significant 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, such as quantification (an example is that provided in Section 2.1.2).

Based on LCS, character overlap is determined and used to calculate precision and recall. These metrics are computed following the method proposed by Da San Martino et al. (2020) for the SemEval-2020 Task on the detection of propaganda techniques in news articles. In the context of the span identification subtask, they introduce formulas for precision and recall to evaluate the character overlap between the predicted span and the ground truth span:

$$P(S,T) = \frac{1}{|S|} \sum_{d \in D} \sum_{s \in S_l, t \in T_l} \frac{|s \cap t|}{|t|} \tag{4.5}$$

$$R(S,T) = \frac{1}{|T|} \sum_{d \in D} \sum_{s \in S_l, t \in T_l} \frac{|s \cap t|}{|s|} \tag{4.6}$$

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;
- D represents the dataset.

Based on the computed precision and recall, the F1 score is calculated as their harmonic mean, as shown in Equation 4.4.

4.3 Instruction Fine-Tuning

The instruction fine-tuning was implemented on a subset of the dataset of instructions oriented to large language models (LLMs), as described in Section 4.3.2. This subset comprises 18,397 samples and retains the balance of the instruction dataset. To optimise the fine-tuning, QLoRA (Dettmers et al., 2023) was also employed: QLoRA is a technique combining 4-bit quantization and Low-Rank Adaptation (LoRA) (Hu et al., 2021) and aiming to reduce computational cost and memory usage.

4.3.1 QLoRA

In implementing the instruction fine-tuning, I employed QLoRA (Dettmers et al., 2023), a technique combining quantization and LoRA (Hu et al., 2021) designed to reduce the computational cost of fine-tuning a LLM. Quantization is a technique used to reduce the computational and memory costs of running inference by converting the model's weights and activations from floating-point numbers (typically 32-bit floating-point) to lower-precision data types like 16-bit floating-point. Reducing the number of bits means the resulting model requires less memory storage. I opted for 4-bit quantization, which is typically used for QLoRa and which represents weights and activations using 4 bits as opposed to 32 bits.

Besides 4-bit quantization, Low-Rank Adaptation (LoRA) was used as well. LoRA is an efficient fine-tuning technique designed to adapt large language models (LLMs) with reduced computational cost. Instead of modifying the entire set of model parameters, which can be expensive in terms of memory and processing power, LoRA introduces small, trainable matrices that are injected into specific layers of the model, typically in the attention mechanism of transformers. By freezing the original model weights and only training these additional low-rank matrices, LoRA significantly reduces the

number of parameters that need updating, making fine-tuning much more efficient. This makes it particularly useful for adapting massive pre-trained models to specific tasks without requiring high-end hardware. Additionally, LoRA adapters can be easily stored, shared, and swapped, enabling quick adaptation to multiple tasks without retraining the full model.

4.3.2 Experimental Settings

The fine-tuning was implemented on a subset of the multilingual instruction dataset due to computational constraints. This subset consists of 18,397 instruction-formatted examples, split into 12,407 examples for training, 3,000 for validation, and 2990 for the test. To ensure a representative distribution, examples were randomly selected from the dataset, while maintaining balance in terms of samples per class, per input language and per instruction language.

Moreover, a set of default hyperparameters was configured to implement the fine-tuning of the LLaMA-3.2 1B model for the sentence disambiguation and idiom identification tasks. The model was trained with a batch size of 32 across 2 epochs, using a cutoff length of 128 tokens for input sequences. For parameter-efficient fine-tuning, LoRA was employed with a rank (r) of 8, alpha value of 16, and dropout rate of 0.05, specifically targeting the query and key projection matrices. The implementation of LoRA enabled to update only 851,968 out of more than 1 billion parameters. The optimisation process used 4-bit quantization with NF4 format to reduce memory requirements. The learning process was 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 was set to 2 steps with a maximum gradient norm of 1.0, and gradient checkpointing was enabled to optimise memory usage. The training uses mixed-precision computation (FP16) and employed early stopping.

4.4 Results and Discussion

This section reports the results of the instruction fine-tuned LLaMA 3.2. For each input language, precision, recall and F1 scores are examined according to the instruction language in the sentence disambiguation and idiom iden-

tification tasks. These results are then analysed in the light of the research questions outlined in Chapter 1.

4.4.1 Evaluation for Task 1

In this section, I report and discuss the results related to the sentence disambiguation task in terms of precision, recall and F1 scores.

Input Lang.	Instruction Lang.	Precision	Recall
en	en	0.4979	0.9307
	it	0.4927	0.9564
	pt	0.4898	0.8828
it	en	0.4248	0.9761
	it	0.4555	0.9557
	pt	0.4624	0.9077
pt	en	0.3447	0.9481
	it	0.4075	0.8950
	pt	0.4007	0.8467

Table 4.1: Precision and recall for Task 1.

Table 4.1 presents the precision and recall values for the different combinations of the instruction language and the input language.

When the input language is English, precision remains consistent across the different instruction languages, hinting at the model's moderate preciseness in identifying the positive class (idiomatic usage). On the other hand, recall stays high across the different instruction languages.

When the input is in Italian, precision tends to remain consistent across the different instruction languages, with the highest value (0.4624) achieved when in combination with Portuguese instructions. Recall stays high, especially when the instruction language is English (0.9564).

Shifting to the Portuguese input, with English instructions, the precision drops to 0.3447, while it is slightly higher with Italian and Portuguese instructions. This suggests that the model struggles the most when the instructions are written in English, making more false positive predictions. Recall, however, remains relatively high, with values ranging from 0.8467 (Portuguese instruction and input) to 0.9481 (English instruction, Portuguese input).

Overall, these results indicate a trend where recall remains consistently high across all instruction and input language combinations, while precision stays moderate. This suggests that the model is effective at capturing the majority of true positives, as evidenced by the high recall values. However, the moderate precision indicates that the model may also generate a significant number of false positives.

Instruction Lang	**Input Lang**		
	en	**it**	**pt**
en	0.6487	0.5919	0.5056
it	0.6503	0.6169	0.5601
pt	0.6300	0.6127	0.5440

Table 4.2: F1 scores for Task 1.

As for the F1 scores shown in Table 4.2, some notable trends emerge when different instruction languages are used.

For English as the input language, across different instruction languages, the F1 score remains consistent, with a slight improvement when instructions are written in Italian.

Even with the input in Italian, the F1 scores stay consistent across instructions written in English, Italian, and Portuguese, with the highest F1 score (0.6169) being achieved when the instructions are in Italian.

Finally, for Portuguese as the input language, the best performance is achieved with Italian instructions (0.5601), followed closely by Portuguese, and the lowest score occurs with English instructions (0.5056).

Regarding RQ1, these results suggest that, in general, instruction language matching with input language does not necessarily lead to higher F1 scores. Turning to RQ2, it could be argued that the exact effect varies slightly depending on the language pair. In this case, the Italian instructions appear to benefit the model's performance across all language combinations.

4.4.2 Evaluation for Task 2

For Task 2, Table 4.3 presents precision and recall scores across different instruction and input language combinations.

Input Lang.	Instruction Lang.	Precision	Recall
en	en	0.6797	0.2391
	it	0.6325	0.2112
	pt	0.5766	0.2495
it	en	0.6949	0.2234
	it	0.7202	0.2203
	pt	0.6094	0.2641
pt	en	0.7034	0.1751
	it	0.7133	0.1834
	pt	0.6002	0.2353

Table 4.3: Precision and recall for Task 2.

With the input in English, the highest precision is achieved with English instructions (0.6797), followed by Italian (0.6325) and Portuguese (0.5766). Yet, recall is relatively low across all instruction languages, with the highest recall seen when instructions are in Portuguese (0.2495).

When the input is written in Italian, the highest precision score is observed when instructions are in Italian (0.7202), followed by English (0.6949), while with Portuguese precision drops to 0.6094. Recall is highest with Portuguese instructions (0.2641), slightly outperforming Italian (0.2203) and English (0.2234).

Finally, for inputs in Portuguese, precision is highest when instructions are in Italian (0.7133), followed closely by English (0.7034), whereas it diminishes with Portuguese (0.6002). Recall, however, is best when instructions are in Portuguese (0.2353).

For the idiom identification task, an opposite trend can be observed, compared to Task 1: the model exhibits, in general, a higher precision and a lower recall, suggesting that it is accurate when predicting a character belonging to an idiomatic expression, but it tends to miss many characters as well.

Instruction Lang	Input Lang		
	en	**it**	**pt**
en	0.3538	0.3381	0.2804
it	0.3166	0.3374	0.2917
pt	0.3483	0.3685	0.3381

Table 4.4: F1 scores for Task 2.

Shifting to Table 4.4, the analysis of the F1 scores reveal that in the idiom identification task the model still struggles to identify the characters belonging to an idiomatic expression, even after the instruction fine-tuning.

Examining the performance across the different language combinations, with the input in English, the highest F1 score is achieved when instructions are in English (0.3538), closely followed by Portuguese (0.3483), while Italian instructions result in the lowest performance (0.3166).

For the Italian input, the best F1 score is obtained when the instructions are in Portuguese (0.3685), while English and Italian instructions reach similar scores, respectively 0.3381 and 0.3374.

In the case of the Portuguese input, the highest F1 score is achieved when instructions are in Portuguese (0.3381), followed by Italian (0.2917) and English (0.2804).

Considering RQ1, these results indicate that for the idiom identification task, the alignment between the instruction and the input language might lead to an improved performance in some cases (e.g., Portuguese). Consequently, as for RQ2, monolingual combinations (in this case, Portuguese instruction and Portuguese input, and English instruction and English input) might benefit the span identification task related to idioms.

4.4.3 Evaluation of Baseline vs. Fine-Tuned Results

This section presents a comparison between baseline and fine-tuned model performance. Table 4.5 shows differing effects of the fine-tuning across the

Task	Language	Baseline	Fine-tuned	Improvement
Task 1	en	0.6141	0.6041	-0.01
	it	0.6320	0.6233	-0.0086
	pt	0.6362	0.6084	-0.0278
Task 2	en	0.2995	0.3387	+0.0392
	it	0.3199	0.3205	+0.0006
	pt	0.3125	0.3536	+0.0412

Table 4.5: Comparison between baseline and fine-tuned in F1 scores.

two tasks. For Task 1, fine-tuning does not lead to an improvement in F1

scores, and slight performance drops are observed across all three languages. This suggests that the fine-tuning did not enhance the model's ability to distinguish between idiomatic and literal meanings.

In contrast, for Task 2, the instruction fine-tuning results in an overall improvement, with F1 score increases for English (+0.0392), Italian (+0.0006), and Portuguese (+0.0412). The largest gains are observed in Portuguese and English, while Italian shows a minimal but positive effect. These results suggest that the fine-tuning had a more noticeable impact on the model's ability to identify IEs within sentences and that instruction fine-tuning helped more with character-level span detection.

These findings suggest that, while idiom identification benefits from fine-tuning, sentence disambiguation shows slight declines. The fine-tuning could possibly lead to a more consistent improvement with a larger dataset or with hyperparameter tuning, which could help mitigate these declines and optimise performance.

Chapter 5

Conclusions and Future Work

This thesis introduced a multilingual dataset consisting of instruction-formatted examples, specifically designed for idiomatic expressions (IEs). Each example comprises an input sentence, an instruction defining the task that the model has to perform, and an expected output. This dataset allows for the exploration of two tasks related to IEs: sentence disambiguation, a binary classification task consisting in determining if a sentence contains an IE or not; and idiom identification, a span identification task aimed at detecting the span within the input sentence that corresponds to the IE. These tasks were examined in a multilingual setting involving three languages, English, Italian, and Portuguese, which were used as both instruction and input languages, covering all possible language combinations.

This thesis aimed to develop an instruction fine-tuned version of LLaMA 3.2 1B for these tasks and across the three languages. This fine-tuning provided insights into two research questions. Research question 1 investigates if having the instruction and the input in the same language could lead to an improved model's performance, while research question 2 focuses on the combinations of the instruction and input language yielding better results.

The findings suggest that when the instruction and the input language coincide, the model does not necessarily perform better, compared to when they differ. Instead, the interactions between languages appear to be more complex, with certain language combinations proving beneficial in ways that go beyond simple alignment. This suggests that cross-linguistic influences might play a role, where one language may enhance performance on another, rather than strict language matching being the primary driver of success.

However, this thesis was limited to three languages. Future work could

expand the scope to other languages, even from different families, to gain a deeper understanding of cross-linguistic interactions.

Additionally, a promising direction would be the creation of datasets that annotate idiomaticity on a continuum rather than as a binary distinction, aligning with recent linguistic theories that view idiomaticity as a scalar phenomenon. From a methodological perspective, this thesis did not implement hyperparameter tuning, instead relying on default values, and limited the fine-tuning to a small subset. This could explain the declines in the model's performance for Task 1, compared to the baseline.

Future research could explore optimised hyperparameters to improve performance, as well as use a larger dataset for the instruction fine-tuning. Moreover, other large language models (LLMs) beyond LLaMA could be fine-tuned, not only to assess their performance but also to compare encoder-based and encoder-decoder models on the same IE-related tasks.

Bibliography

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints, 10 2023. URL https://arxiv.org/abs/2305.13245.

Jay Alammar and Maarten Grootendorst. *Hands-On Large Language Models*. O'Reilly Media, Inc., 12 2024.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Thomas Henighan, Andrew M Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack A Clark, Samuel McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861, 12 2021. doi: 10.48550/arxiv.2112.00861.

Timothy Baldwin and Su Nam Kim. *Multiword Expressions*, pages 267–292. Handbook of Natural Language Processing. CRC Press LLC, 2010.

Henk Barkema. Idiomaticity and terminology: A multi-dimensional descriptive model. *Studia Linguistica*, 50:125–160, 1996. doi: https://doi.org/10.1111/j.1467-9582.1996.tb00347.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9582.1996.tb00347.x.

Rhys Biddle, Aditya Joshi, Shaowu Liu, Cecile Paris, and Guandong Xu. Leveraging sentiment distributions to distinguish figurative from literal health reports on twitter. *Proceedings of The Web Conference 2020*, 04 2020. doi: 10.1145/3366423.3380198.

Irina Bigoulaeva, Rachneet Singh Sachdeva, Harish Tayyar Madabushi, Aline Villavicencio, and Iryna Gurevych. Effective cross-task transfer learning for

explainable natural language inference with t5. *arXiv (Cornell University)*, 01 2022. doi: 10.18653/v1/2022.flp-1.8.

Cristina Cacciari. *The Place of Idioms in a Literal and Metaphorical World*, pages 27–55. Processing, structure and interpretation. Erlbaum, 1993. doi: 10.4324/9781315807133-10.

Cristina Cacciari and Sam Glucksberg. Chapter 9 understanding idiomatic expressions: The contribution of word meanings. *Advances in Psychology*, 77:217–240, 1991. doi: 10.1016/s0166-4115(08)61535-6.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations, 12 2018. URL `https://arxiv.org/abs/1812.01193`.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.297.

Noam Chomsky. Rules and representations. *Behavioral and Brain Sciences*, 3:1–15, 1980.

Zheng Chu, Ziqing Yang, Yiming Cui, Zhigang Chen, and Ming Liu. Hit at semeval-2022 task 2: Pre-trained language model for idioms detection. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 221–227, 01 2022. doi: 10.18653/v1/2022.semeval-1.28.

Zhibo Chu, Shiwen Ni, Zichong Wang, Xi Feng, Min Yang, and Wenbin Zhang. History, development, and principles of large language models-an introductory survey, 06 2024. URL `https://arxiv.org/abs/2402.06853`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 8440–8451. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.747. URL `https://aclanthology.org/2020.acl-main.747/`.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.186. URL https://aclanthology.org/2020.semeval-1.186/.

Verna Dankers, Christopher Lucas, and Ivan Titov. Can transformer be too compositional? Analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 01 2022. doi: 10.18653/v1/2022.acl-long.252.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. A hard nut to crack: Idiom detection with conversational large language models, 2024. URL https://aclanthology.org/2024.figlang-1.5.pdf.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized LLMs, 2023. URL https://arxiv.org/abs/2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North*, 1, 2019. doi: 10.18653/v1/n19-1423. URL https://aclanthology.org/N19-1423.pdf.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy,

Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Alwala Kalyan Vasuden, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Koura Punit Singh, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudet, Zheng Yan, Zhengxing Chen, Zoe Papakipos,

Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Re-

strepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The Llama 3 herd of models. *arXiv (Cornell University)*, 07 2024. doi: 10.48550/arxiv.2407.21783.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), 05 2018. URL `https://aclanthology.org/L18-1148/`.

Bruce Fraser. Idioms within a transformational grammar. *Foundations of Language*, 6:22–42, 1970.

Marcos García, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. Probing for idiomaticity in vector space models. *White Rose Research Online (University of Leeds)*, 01 2021. doi: 10.18653/v1/2021.eacl-main.310.

Raymond W. Gibbs. What do idioms really mean? *Journal of Memory and Language*, 31:485–506, 08 1992. doi: 10.1016/0749-596x(92)90025-s. URL `https://www.sciencedirect.com/science/article/pii/0749596X9290025S`.

Raymond W. Gibbs and Nandini P. Nayak. Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology*, 21:100–138, 01 1989. doi: 10.1016/0010-0285(89)90004-2.

Raymond W. Gibbs and Jennifer E. O'Brien. Idioms and mental imagery: The metaphorical motivation for idiomatic meaning. *Cognition*, 36:35–68, 07 1990. doi: 10.1016/0010-0277(90)90053-m.

Reyhaneh Hashempour and Aline Villavicencio. Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80, 12 2020. URL `https://aclanthology.org/2020.cogalex-1.9/`.

Wei He, Marco Idiart, Carolina Scarton, and Aline Villavicencio. Enhancing idiomatic representation in multiple languages via an adaptive contrastive triplet loss. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12473–12485, 01 2024. doi: 10.18653/v1/2024.findings-acl.741.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021. URL `https://arxiv.org/abs/2106.09685`.

Ray Jackendoff. *The architecture of the language faculty*. MIT Press, 1997.

Umar Jamil. Github - hkproj/pytorch-llama-notes: Notes about LLaMA 2 model, 2023. URL https://github.com/hkproj/pytorch-llama-notes/tree/main.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 01 2021. doi: 10.18653/v1/2021.emnlp-main.592.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.703.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 15095–15111. Association for Computational Linguistics, 01 2023. doi: 10.18653/v1/2023.emnlp-main.933. URL https://aclanthology.org/2023.emnlp-main.933/.

Pengfei Liu, Kaiyu Qian, Xipeng Qiu, and Xuanjing Huang. Idiom-aware compositional distributed semantics. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1213, 2017. doi: 10.18653/v1/d17-1124. URL https://aclanthology.org/D17-1124/.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55, 09 2022. doi: 10.1145/3560815. URL https://arxiv.org/pdf/2107.13586.pdf.

Yiheng Liu, Hong-di He, Tianle Han, Zhiwei Xu, Mengyuan Liu, Jiaming Tian, Yutong Zhang, Jiaqi Wang, Xiaohui Gao, Tianyang Zhong, Yi Peng, Shu Xu, Zihao Wu, Zhengliang Liu, Xin Zhang, Shu Zhang, Xintao Hu, Tuo Zhang, Qiang Niu, Tianming Liu, and Bao Ge. Understanding LLMs:

A comprehensive overview from training to inference. *arXiv (Cornell University)*, 01 2024. doi: 10.48550/arxiv.2401.02038.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar S Joshi, Danqi Chen, Omer Levy, Michael Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv (Cornell University)*, 07 2019. doi: 10.48550/arxiv.1907.11692.

Renze Lou, Kai Zhang, and Wenpeng Yin. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, 50:1053–1095, 06 2024. doi: 10.1162/coli_a_00523. URL `https://aclanthology.org/2024.cl-3.7/`.

Niklas Muennighoff, Thomas J Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 15991–16111. Association for Computational Linguistics, 01 2023. doi: 10.18653/v1/2023.acl-long.891.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. How well do embedding models capture non-compositionality? a view from multiword expressions. In Anna Rogers, Aleksandr Drozd, Anna Rumshisky, and Yoav Goldberg, editors, *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34, 01 2019. doi: 10.18653/v1/w19-2004. URL `https://aclanthology.org/W19-2004/`.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 08 2023. URL `https://arxiv.org/abs/2307.06435`.

Geoffrey Nunberg. *The pragmatics of reference*. PhD thesis, 1978.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. Idioms. *Language*, 70: 491, 09 1994. doi: 10.2307/416483.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent 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, pages 9 – 16, Mannheim, 2019. Leibniz-Institut für Deutsche Sprache. doi: 10.14618/ids-pub-9021. URL http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with GPT-4. 04 2023. doi: 10.48550/arxiv.2304.03277.

Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. Sign of the times: Evaluating the use of large language models for idiomaticity detection. *arXiv (Cornell University)*, 05 2024. doi: 10.48550/arxiv.2405.09279.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-45715-2.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 02 2020. URL https://arxiv.org/abs/1910.01108v4.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Honghan Gong, and Francisco Guzmán. Wikimatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. pages 1351–1361. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics., 01 2021. doi: 10.18653/v1/2021.eacl-main.115.

Uliana Sentsova, Debora Ciminari, Cristina España-Bonet, and Josef van Genabith. MultiCoPIE: A multilingual corpus of potentially idiomatic expressions for cross-lingual pie disambiguation. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, June 2025. To appear.

Noam Shazeer. GLU variants improve transformer, 02 2020. URL `https://arxiv.org/abs/2002.05202v1`.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. Impli: Investigating nli models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 5375–5388. Association for Computational Linguistics, 01 2022. doi: 10.18653/v1/2022.acl-long.369.

Jianlin Su, Yu Lu, Sheng-Feng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. 04 2021. doi: 10.48550/arxiv.2104.09864.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. Astitchinlanguagemodels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. *Empirical Methods in Natural Language Processing*, pages 3464–3477, 11 2021.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 01 2022. doi: 10.18653/v1/2022.semeval-1.13.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. Id10m: Idiom identification in 10 languages. *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, 01 2022. doi: 10.18653/v1/2022.findings-naacl.208. URL `https://aclanthology.org/2022.findings-naacl.208/`.

Debra A. Titone and Cynthia M. Connine. On the compositional and non-compositional nature of idiomatic expressions. *Journal of Pragmatics*, 31: 1655–1674, 11 1999. doi: 10.1016/s0378-2166(99)00008-9.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv (Cornell University)*, 02 2023a. doi: 10.48550/arxiv.2302.13971.

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 07 2023b. URL `https://arxiv.org/abs/2307.09288`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 12 2017. URL `https://arxiv.org/abs/1706.03762`.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech Language*, 19:365–377, 10 2005. doi: 10.1016/j.csl.2005.05.001. URL `doi:10.1016/j.csl.2005.05.001`.

Thomas Wasow, Ivan Sag, and Geoffrey Nunberg. Idioms: An interim report.

In S. Hattori and K. Inoue, editors, *Proceedings of the XIIIth International Congress of Linguistics*, 1983.

Stefanie Wulff. *Rethinking idiomaticity : a usage-based approach*. Continuum, 1st edition, 2008.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 01 2020. URL `https://arxiv.org/abs/1906.08237v2`.

Lang Yu and Allyson Ettinger. Assessing phrasal representation and composition in transformers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907. Association for Computational Linguistics, 01 2020. doi: 10.18653/v1/2020.emnlp-main.397. URL `https://aclanthology.org/2020.emnlp-main.397/`.

Ziheng Zeng and Suma Bhat. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562, 2021. doi: 10.1162/tacl_a_00442.

Ziheng Zeng and Suma Bhat. Getting BART to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137, 2022. doi: 10.1162/tacl_a_00510.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/1e8a19426224ca89e83cef47f1e7f53b-Paper.pdf`.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey, 10 2023. URL `https://arxiv.org/abs/2308.10792`.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *arXiv (Cornell University)*, 03 2023. doi: 10.48550/arxiv.2303.18223.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. Idiomatic expression paraphrasing without strong supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:11774–11782, 06 2022. doi: 10.1609/aaai.v36i10.21433.

Appendix A

Instruction Dataset Examples

Type	**Inst. Lang.**	**Input Lang.**	**Example**
L	EN	EN	**I**: Can you spot the idiomatic expressions lurking within this sentence? They are: **Input**: Although the encounter was bathed in sunshine, the match failed to reach boiling point but that will be of little concern to Gerard Houllier's team. **Output**: boiling point
H	IT	PT	**I**: La frase è caratterizzata dalla presenza di una o più espressioni idiomatiche? **Input**: Nos últimos anos, muitas universidades têm mostrado quadricópteros realizando manobras aéreas. **Output**: No, la frase non contiene alcuna espressione idiomatica.
H	IT	IT	**I**: La frase è caratterizzata dalla presenza di una o più espressioni idiomatiche? **Input**: Diplo ha definito la lunga attesa della produzione vale la pena avere uno show televisivo così succinto per i nostri fan. **Output**: Sì, l'analisi rivela che la frase incorpora le seguenti espressioni idiomatiche: vale la pena.

Table A.1: Examples from the instruction dataset. 'L' refers to the LLM-oriented instruction type, and 'H' indicates the human-oriented instruction type.