Metaphor and Large Language Models: When Surface Features Matter More than Deep Understanding

Elisa Sanchez-Bayona
HiTZ Center - Ixa
University of the Basque
Country UPV/EHU
elisa.sanchez@ehu.eus

Rodrigo Agerri
HiTZ Center - Ixa
University of the Basque
Country UPV/EHU
rodrigo.agerri@ehu.eus

Abstract

This paper presents a comprehensive evaluation of the capabilities of Large Language Models (LLMs) in metaphor interpretation across multiple datasets, tasks, and prompt configurations. Although metaphor processing has gained significant attention in Natural Language Processing (NLP), previous research has been limited to single-dataset evaluations and specific task settings, often using artificially constructed data through lexical replacement. We address these limitations by conducting extensive experiments using diverse publicly available datasets with inference and metaphor annotations, focusing on Natural Language Inference (NLI) and Question Answering (QA) tasks. The results indicate that LLMs' performance is more influenced by features like lexical overlap and sentence length than by metaphorical content, demonstrating that any alleged emergent abilities of LLMs to understand metaphorical language are the result of a combination of surface-level features, in-context learning, and linguistic knowledge. This work provides critical insights into the current capabilities and limitations of LLMs in processing figurative language, highlighting the need for more realistic evaluation frameworks in metaphor interpretation tasks. Data and code publicly available.[1]

1 Introduction

Figurative language is a recurrent element in our daily communication. It reshapes our perception and understanding of knowledge, allowing us to better comprehend and transmit abstract concepts from a more concrete domain. Lakoff and Johnson (1980) defined these mental associations as **conceptual mappings**, which are verbalized through language into **linguistic metaphors**, subject of study of our work.

The widespread use of metaphors in everyday language has boosted the popularity of research on this type of figurative language within the field of NLP. Large Language Models (LLMs, to refer to decoder-only models) are now widely available, not only for NLP researchers but also for all kinds of users in chatbot assistant forms. Moreover, figurative language, metaphors specifically, are key for other NLP tasks, such as hate speech detection (Lemmens et al., 2021), political discourse analysis (Baleato Rodríguez et al., 2023), or mental illness detection (Zhang et al., 2021; Kurian et al., 2025).

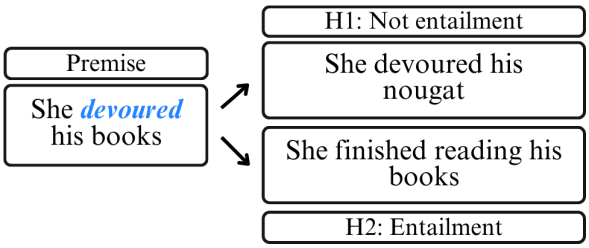

Figure 1: Example from IMPLI (Stowe et al., 2022) dataset with one premise and two hypotheses. The premise contains the verb *devour* used metaphorically, equivalent to 'to read vividly'. Note that the inference relation is affected by the metaphorical expression.

For this reason, it is essential to critically assess the capabilities of LLMs and their applications, particularly their ability *to understand complex cognitive-linguistic phenomena like metaphorical expressions*.

This work focuses specifically on **metaphor interpretation**. Initial publications on this topic were based on statistics and machine learning (Agerri, 2008; Mohler et al., 2013; Shutova, 2010). The introduction of Transformer models (Devlin et al., 2019) marked a major milestone, providing a more accessible and open framework for evaluating and probing language models' abilities to process metaphorical expressions (Mao et al., 2021; Pedinotti et al., 2021).

Although there has been some recent interest in researching metaphorical analogies or concep-

[1] https://github.com/elisanchez-beep/metaphorLLM

Findings of the Association for Computational Linguistics: ACL 2025, pages 17462–17477

tual metaphors (Tong et al., 2024; Boisson et al., 2024; Dmitrijev et al., 2024), our work is centered on linguistic metaphors. On this particular topic, while interesting, previous work studying the capabilities of LLMs to understand linguistic metaphors has been hindered by several shortcomings. From a task perspective, the data used in these approaches has typically been developed through lexical substitution. This means that there is a metaphor in a premise and that a literal version, used as the hypothesis, is generated by replacing the metaphor with its literal sense (Chakrabarty et al., 2022, 2021a; Stowe et al., 2022). As a result, these resources are not representative of metaphor occurrence in natural language utterances and incorporate biases, namely lexical overlap, that may affect performance in the NLI or QA tasks (Naik et al., 2018; Stowe et al., 2022; Liu et al., 2022).

Moreover, most of the previous work tests LLMs only on a single dataset prepared ad-hoc for each specific experimentation. Furthermore, the evaluation scenarios are limited to one setting or task. Thus, there is a lack of comprehensive, cross-dataset, and multi-prompt evaluation of the interpretation abilities of LLMs on diverse and natural language corpora. As a consequence, results have been rather mixed, with some studies claiming that LLMs achieve understanding over chance but below human performance (Liu et al., 2022) while others conclude that LLMs can accurately establish entailment relations between the figurative and their literal counterparts (Stowe et al., 2022).

In order to address these issues, this work provides comprehensive experimentation to evaluate the ability of LLMs to understand and interpret metaphorical language across several datasets and tasks. More specifically:

- RQ1: Does the presence of metaphors in the text impact LLMs' ability to perform the correct inference?
- RQ2: Do LLMs exhibit generalization or even emergent capabilities in the understanding of metaphorical language across tasks, prompt verbalizations, and datasets?
- RQ3: Do datasets generated through lexical replacement introduce biases that may explain the high performance of LLMs in metaphor understanding?

To tackle these questions, we gathered the most recent publicly available datasets with inference and metaphor annotations. Then, we conducted comprehensive experiments on these data on evaluation scenarios based on NLI and QA (Agerri, 2008; Comșa et al., 2022; Bollegala and Shutova, 2013a). The results and subsequent analyses led to the following key contributions:

- We present the first study in which the performance of LLMs on metaphor interpretation is evaluated in a multi-dataset, multi-task, and multi-prompt setting.
- We conduct quantitative and qualitative analyses that show, on the one hand, that LLMs' performance is more sensitive to lexical overlap and sentence length than to metaphor presence when extracting the inference. Thus, rather than an emergent ability, performance appears to result from a combination of surface-level features, in-context learning, and linguistic knowledge.
- We show that few-shot and chain-of-thought prompt (CoT) setups outperform the performance of fine-tuned encoders (Devlin et al., 2019; Liu et al., 2019; Conneau et al., 2020). This reduces the demand for large amounts of manually annotated data specifically crafted for the task.

In the next sections, we present previous work on metaphor interpretation (Section 2) and describe the data used for evaluation (Section 3). In Section 4, we describe the experimental settings as well as the generation of an adversarial, literal paraphrased version of the datasets. Subsequently, in Section 5, we report the results and perform a series of quantitative and qualitative analyses (Section 6) to conclude our work (Section 7).

2 Related Work

As mentioned in the previous section, metaphor interpretation has been formulated through other pivot tasks from Natural Language Understanding (NLU), such as NLI (Agerri, 2008; Mohler et al., 2013), QA (Comșa et al., 2022) or paraphrasing (Bizzoni and Lappin, 2018; Bollegala and Shutova, 2013b). Initially, supervised and unsupervised deep learning techniques were employed (Shutova et al., 2013, 2012), using datasets specifically designed for the task (Mohammad et al., 2016) or other external resources with exploitable linguistic information (Zayed et al., 2020).

Along with Transformers' emergence, metaphor interpretation was approached as a sequence classification task to evaluate Masked Language Models

(encoders). Mao et al. (2021) evaluate BERT's capability to generate literal substitutes by leveraging metaphor detection VUAM (Steen, 2010) and MOH-X (Mohammad et al., 2016) datasets. Similarly, Pedinotti et al. (2021) develop their evaluation corpus with conventional and novel metaphors to test if BERT distinguishes between metaphor and nonsense (Griciūtė et al., 2022).

Recent studies framed metaphor interpretation through NLI or QA to evaluate LLMs such as GPT (Brown et al., 2020) and LLaMA (Touvron et al., 2023). The work of Comșa et al. (2022) introduces a set of 300 metaphorical questions and paired implications to ask models whether these implications are true or false. In addition to English resources, Kabra et al. (2023) publish a dataset with seven underrepresented languages with a high number of speakers to assess the impact of socio-cultural characteristics on model performance.

Datasets specifically designed to evaluate metaphor interpretation through inference are relatively scarce and based on lexical replacement (Liu et al., 2022; Stowe et al., 2022). They usually include different annotations, such as metaphoricity, acceptability, or correctness of the paraphrases. However, for our study, we focus on datasets that include inference labels that are affected by metaphorical expressions. In the following, we will provide a more detailed overview of recent publications that present both metaphor and inference labels, which we will leverage for our multi-task, multi-prompt, and cross-dataset evaluation (Aghazadeh et al., 2022) of the capabilities of LLMs to understand metaphorical language.

3 Evaluation Datasets

In this section, we will describe the main features of the datasets used for the experiments. We collected these resources because they are labeled for NLI and include metaphorical language. All of these corpora are manually validated. Except for Meta4XNLI, the majority of them were developed through lexical replacement. Thus, results might be biased by their templatic nature and lexical artifacts (Boisson et al., 2023). In addition, we describe the process of generating adversarial literal examples to test whether correct inferences were due to the understanding of metaphorical expressions by LLMs or related to more surface-level features. We include the distribution of the data in Table 1. Examples of each dataset can be checked in Table 6 in the Appendix Sec. A.

Dataset	#Test	Labels	Met loc.	Lex Sub.
Figurative-NLI	613	E, NE	P	✓
IMPLI	668	E, NE	P	✓
FLUTE	248	E, C	H	✓
Fig-QA	2188	E, NE	P	✓
Meta4XNLI	598	E, C, N	P, H	✗

Table 1: Distribution of corpora used in evaluation experiments. **#Test** refers to the number of paired instances in the test set. In **Labels**, inference tags from the original dataset: E for *entailment*, NE: *not_entailment*, C: *contradiction*, and N: *neutral*. C and N tags were merged into NE. **Met loc.** indicates if metaphors are present in the premises (P) or hypotheses (H). In **Lex Sub.**, if datasets were developed through lexical substitution.

Figurative-NLI (Chakrabarty et al., 2021a) This test set contains 12,500 instances for Recognizing Textual Entailment (RTE), a.k.a. NLI, with simile, metaphor, and irony examples. For its collection, they leveraged five existing datasets, although we will focus only on metaphors. The metaphor subset comprises a total of 300 instances. They used 150 literal sentences from the Gutenberg Poetry corpus (Jacobs, 2018) subsequently curated in the work of Chakrabarty et al. (2021b). Chakrabarty et al. (2021a) created the metaphorical sentences for the entailment relation by replacing a literal verb with a metaphorical one. To develop non-entailment examples, they swapped the verb from the literal sentences with its antonym. Therefore, metaphors always occur in premises.

IMPLI (Stowe et al., 2022) This dataset is a compilation of 24k silver and 1.8k gold NLI pairs with metaphorical and idiomatic language. In this work, we will use the 668 instances with metaphors from the gold subset. To develop gold pairs, Stowe et al. (2022) collected metaphorical sentences from the VUAM corpus (Steen, 2010), Gutenberg Poetry corpus (Jacobs, 2018), and from Mohammad et al. (2016). Annotators were asked to rephrase the metaphorical sentence by removing the figurative expression. To create not-entailed hypotheses, annotators rewrote premises, adding elements to change their meaning but maintaining as much as possible the lexical overlap.

FLUTE (Chakrabarty et al., 2022) This benchmark provides 9000 NLI pairs with metaphor, simile, sarcasm, and idioms. Regarding metaphors, they collected 750 metaphoric sentences from exist-

ing datasets, namely, Figurative-NLI, IMPLI, and Srivastava et al. (2023). They prompted GPT-3 with metaphorical sentences to generate a literal paraphrase as entailment. For contradictions, they used the GPT-3 generated literal sentences and prompted the model to invert the sentence and contradict the metaphor itself. Results were reviewed and post-edited when needed by annotators. In total, they obtained 1500 pairs, from which we leveraged the 248 pairs belonging to the gold test set.

Fig-QA (Liu et al., 2022) The dataset follows the Winograd schema format (Levesque et al., 2012), wherein human annotators created paired sentences that share identical opening segments but conclude with contrasting metaphorical meanings. Each sentence pair is accompanied by two corresponding hypotheses: one that represents an entailed paraphrase and another that is not entailed. For our experimental analysis, we used the 2188 NLI pairs extracted from the development set, as the test set labels were not accessible.

Meta4XNLI (Sanchez-Bayona and Agerri, 2024) This parallel dataset includes NLI instances for Spanish and English. Since other corpora are only available in English, we used only samples in this language. It is a compilation of existing NLI datasets, XNLI (Conneau et al., 2018) and esXNLI (Artetxe et al., 2020). In contrast to the other datasets, it contains spontaneously generated natural language text for NLI tasks which was subsequently annotated for metaphoricity by the Meta4XNLI project. Our experiments use the test set with metaphorical sentences, that is, a total of 598 NLI pairs.

Adversarial Paraphrases In works that approached metaphor interpretation through paraphrasing (Bollegala and Shutova, 2013a; Stowe et al., 2022; Chakrabarty et al., 2022; Liu et al., 2022), the metaphorical premise sentence is usually rephrased, via lexical substitution, into a literal one that serves as the hypothesis. However, instead of using simple lexical substitution, we generate complete paraphrases (striving to preserve the original semantic content) for all sentences containing metaphorical expressions into their literal counterparts to act as adversarial examples. The newly generated paraphrases will allow us to examine potential variations in the performance of LLMs when processing literal versus metaphorical language. Furthermore, rather than relying on manual conversion, literal paraphrases were created using LLMs.

4 Experimental Setup

In this section, we provide technical information about the experiments and the characteristics of each prompt formulation. We will also detail the settings used for the generation of the literal paraphrases.

4.1 Evaluation

We test the datasets labeled for inference with linguistic metaphors through multiple prompt verbalizations and by framing the task of metaphor interpretation as NLI and QA, respectively, in both zero- and few-shot settings.

Prompts We propose diverse prompt configurations to assess how the verbalization, presence of examples, and context affect model performance. We differentiate between two task formulations: NLI (Stowe et al., 2022; Chakrabarty et al., 2021a) and QA (Rakshit and Flanigan, 2023; Comșa et al., 2022). In the NLI formulation, the model is asked to identify the inference relationship, such as entailment, or others ('neutral' and 'contradiction', merged into the 'not_entailment' class due to original dataset labels).

In contrast, the QA setting consists of determining whether the sentences are entailed or not by answering in a yes/no fashion. Thus, while in the case of NLI prompts the valid answers are ["entailment", "other"], in QA the possible responses correspond to ["yes", "no"]. For each setting, we design zero- and few-shot (one example for each inference type) prompts. Finally, we also explore chain-of-thought (CoT) prompting, also framed as a QA task, but with a more detailed context that explains the steps to perform the task in greater depth. Table 7 in the Appendix Section B illustrates the exact prompts used.

Models We evaluated the following large language models: `Llama-3-8B-Instruct`, `Llama-3.3-70B-Instruct` (Dubey et al., 2024), `Mistral-7B-Instruct` (Jiang et al., 2023), `Qwen/Qwen2.5-7B-Instruct`, `Qwen2.5-72B-Instruct` (Team, 2024), `gemma-3-4b-it` and `gemma-3-27b-it` (Team et al., 2024). We used implementation of HuggingFace and vLLM (Kwon et al., 2023) with for

Dataset	Baseline	Mistral-7B		Llama-8B		Llama-70B		Qwen-7B		Qwen-72B		Gemma-4B		Gemma-72B	
		QA-Few	CoT	QA-Few	CoT	QA-Few	CoT	QA-Few	CoT	QA-Few	CoT	QA-Few	CoT	QA-Few	CoT
M4X-met	76.73	72.07	**73.41**	61.04	**78.26**	85.95	**87.96**	89.13	**89.80**	89.97	**90.47**	81.60	**79.93**	88.79	**90.97**
M4X-lit		73.58	72.74	62.04	76.76	84.11	85.45	82.61	86.96	86.79	88.63	79.26	79.10	86.97	88.46
Fig-QA-met	90.32	74.91	76.42	58.04	76.17	89.08	**89.48**	72.21	74.91	85.05	**85.24**	68.46	70.93	81.99	**81.58**
Fig-QA-lit		78.38	**80.67**	62.71	**81.58**	83.36	85.51	73.86	**76.42**	81.17	80.57	73.35	**78.20**	80.62	79.84
Fig-NLI-met	88.09	86.95	**88.25**	62.81	**86.13**	91.35	**92.80**	90.86	**90.21**	94.13	**95.43**	83.85	**86.13**	90.37	**90.37**
FigNLI-lit		83.85	82.22	59.22	78.30	86.79	87.77	84.50	84.18	88.09	88.58	76.83	80.42	84.99	85.81
FLUTE-met	81.80	79.03	82.26	62.50	**83.87**	85.48	85.08	76.21	81.85	87.10	**87.50**	73.39	**79.03**	85.48	**85.08**
FLUTE-lit		75.81	**83.06**	54.43	79.44	86.95	**85.89**	84.83	**84.18**	88.09	86.69	77.82	77.82	85.32	83.48
IMPLI-met	85.55	84.88	**84.28**	60.48	**82.93**	93.54	**94.31**	89.79	**90.24**	93.69	**95.04**	78.22	**84.98**	93.39	**93.99**
IMPLI-lit		80.39	79.79	61.38	81.87	84.13	86.68	84.13	86.98	85.93	87.57	75.60	83.23	87.12	88.77
Avg_met	78.63	79.57	**80.92**	60.97	**81.47**	89.08	**89.93**	86.50	**88.03**	91.22	**92.11**	79.27	**82.52**	89.51	**90.10**
Avg_lit		78.40	79.70	59.96	79.59	85.07	86.26	81.83	82.94	85.82	85.85	75.90	79.92	84.51	84.48
Avg_all		79.44	80.31	60.47	80.75	87.07	88.09	84.16	85.48	88.52	88.98	77.58	81.22	87.01	87.29

Table 2: Accuracy results. Baselines: for Meta4XNLI (Sanchez-Bayona and Agerri, 2024), setup as NLI obtained by fine-tuning XLM-RoBERTa (Conneau et al., 2020) on the Meta4XNLI's training set; Fig-QA setup as in a Winograd-style QA task, results obtained with a fine-tuned RoBERTa-large (Liu et al., 2022); in Figurative-NLI (Chakrabarty et al., 2021a) the baseline was obtained with RoBERTa-large (Liu et al., 2019); for FLUTE (Chakrabarty et al., 2022) the NLI task is addressed with the encoder-decoder T5 (Raffel et al., 2023) fine-tuned on e-SNLI (Camburu et al., 2018); finally, IMPLI is also a NLI benchmark and best previous result obtained with RoBERTa-large (Stowe et al., 2022). In bold, the best result for each version of each dataset with CoT prompt, that is, the original dataset with metaphors or the literal paraphrased version. In underscore, the best model for each evaluation dataset.

every evaluation setting. We set the following hyperparameters to limit the response to a range of selected words: *temperature*=0.3, *max_tokens*=5, and a fixed seed. To compute accuracy, we search for the tokens corresponding to the *valid answers* in the LLMs string response and map them to their corresponding NLI label according to the formulation of the task and the prompt, detailed in Table 7 in the Appendix Section B. If none of the labels appeared in the answer, we assigned a "unk" label.

4.2 Literal Paraphrase Generation

We apply `Mistral-7B-Instruct` (HuggingFace implementation) and `Command R+` through the Cohere's API to generate literal paraphrases of the metaphorical sentences in the datasets. We prompt the models only with those sentences that include metaphors (see the prompt specified in Table 8, Appendix C). The input was the same for both models.

With respect to the parameters, we used the default settings from the API of `Command R+`. We had to adjust *temperature=3* and *max_new_tokens=100* parameters with `Mistral-7B-Instruct`, to limit generation to a single sentence.

As `Command R+`'s paraphrases achieved higher performance in a first evaluation round performed with `Llama-3-8B-Instruct` and `Mistral-7B-Instruct`, `Mistral-7B-Instruct`'s paraphrases were discarded from the final evaluation (but see all results of zero-/few-shot evaluation on literal paraphrases in Appendix Section D Table 9), using only the paraphrases generated with `Command R+`. The number of test instances and inference labels are maintained the same for the evaluation with every dataset.

5 Results

We first report the results of zero- and few-shot experiments in all experimental settings with the original metaphor datasets (Section 5.1, *-met* results in Table 2), while in Section 5.2, we discuss the results obtained with their literal paraphrases. (in Table 2, *-lit* results). We provide the results with QA-Few and CoT prompts in Table 2, and the results of all evaluations in the Appendix Table 10.

5.1 Zero-/Few-shot Evaluation

Experiments demonstrated that zero-shot results of smaller LLMs were close to random, while larger versions of the LLMs fared much better in this particular setting (see Table 10 in Appendix). However, every model behaved much more robustly in few-shot and CoT evaluations, both formulated as a QA task, which is why the main results reported on Table 2 focus on these two evaluation scenarios. Thus, in few-shot settings, where the models are prompted with examples, results are already quite competitive, and improvements with respect to zero-shot are substantial especially for the smaller models, such as `Llama-3-8B-Instruct`,

gemma-3-4b-it and Mistral-7B-Instruct. In other words, adding examples to the prompting of larger models does not have so much effect in terms of accuracy results.

Adding a CoT prompt to the QA task, which offers a more fine-grained explanation of the task together with examples, improves the performance of every model across the board, obtaining the best overall average scores, as shown in Table 2 (*-met* scores). However, performance disparities across datasets are evident across all experimental setups. Thus, the Figurative-NLI dataset stands out as the one with the highest score and Fig-QA with the lowest. Still, the average results of all models exceed 80 points for most datasets with CoT prompt, being Qwen2.5-72B-Instruct the one that achieves the best performance, followed by gemma-3-27b-it and Llama-3.3-70B-Instruct. Mistral-7B-Instruct is the worse performing model, with results 10 points lower in accuracy than the rest of the models. Finally, it is worth mentioning that our in-context learning approach managed to outperform strong baselines often based on fine-tuned encoder and encoder-decoder models, the only exception being Fig-QA.

5.2 Evaluation on Adversarial Literal Paraphrases

In these experiments, we evaluated the performance of LLMs with the automatically generated literal version of the datasets maintaining the same experimental settings. Overall, the trends observed in Tables 2 and 10 (*-lit* scores) align with those found in the evaluation of the original datasets (*-met* scores): in zero-shot settings and smaller models, the scores resemble random predictions; results improve in few-shot settings, and the CoT prompt achieves the best performance. Larger models like Llama-3.3-70B-Instruct, Qwen2.5-72B-Instruct and gemma-3-27b-it keep a stable performance across most evaluation settings, achieving the best scores also with the CoT prompt.

Quite surprisingly, the results in Table 2 seem to suggest that LLMs perform better on the original datasets containing metaphorical expressions, than with their literal paraphrases. These are the only results consistent with previous work comparing metaphorical and literal contexts (Agerri, 2008; Rakshit and Flanigan, 2023; Sanchez-Bayona and Agerri, 2024). Does this mean that LLMs display emergent capabilities to understand metaphorical language? Or is it explained by in-context learning competencies that arise from alternative prompting techniques, lexical overlap between premises, and hypothesis and linguistic knowledge? We will further analyse this behavior in the next Section 6.

6 Analysis of Results

Firstly, to explore the factors contributing to the disparity in results between the datasets, we tried to capture some form of lexical overlap between premises and hypotheses via Levenshtein distance (Levenshtein, 1966), and average sentence length. These analyses aim to provide further insight into how sentence structure and similarity may influence the models' performance (Stowe et al., 2022; Naik et al., 2018). Secondly, we manually inspected some errors from the evaluation with adversarial literal paraphrases. All the examples used for quantitative and qualitative analyses come from the experimental setup that obtained the best average results, that is, CoT prompt with Qwen2.5-72B-Instruct.

6.1 Lexical Overlap and Sentence Length

To approximate some measurement of lexical overlap, we used the Levenshtein distance metric. It quantifies the number of changes in characters (insertions, deletions, or substitutions) required to transform one word into another. That is, the greater the number of changes, the more distinct the two sentences are from one another.

Dataset	Samples	CoT	Levenshtein	Sent_len
Meta4XNLI-met	598	**90.47**	101.44	16.02
Meta4XNLI-lit		88.63	108.42	16.73
Fig-QA-met	2188	**85.24**	27.31	7
Fig-QA-lit		80.57	36.32	7.58
Figurative-NLI-met	613	**95.43**	6.07	7.33
Figurative-NLI-lit		88.58	24.22	8.1
FLUTE-met	248	**87.50**	23.6	8.74
FLUTE-lit		86.69	37.2	9.93
IMPLI-met	668	**95.04**	12.37	9.1
IMPLI-lit		87.57	35.21	9.66

Table 3: Numerical results of quantitative analysis. Column **CoT** is the accuracy score obtained in the evaluation on the datasets from the Qwen2.5-72B-Instruct + CoT prompt experimental setup. **Levenshtein**: distance metric used to measure lexical overlap between hypotheses and premises. The last column refers to the **average sentence length** of premises and hypotheses.

Other metrics such as Jaccard (Jaccard, 1912), BLEU (Papineni et al., 2002), or semantic similarity methods, operate more at meaning level, while

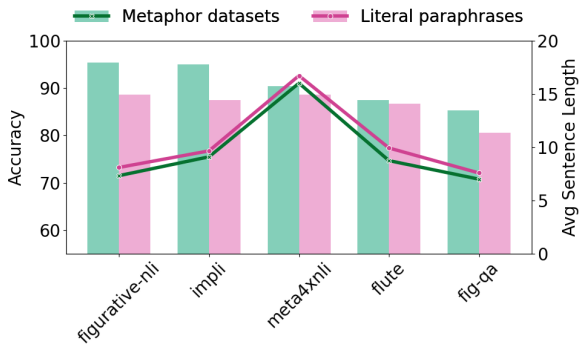

Figure 2: Comparison of the evaluation with original datasets and their literal paraphrases with CoT prompt and `Qwen2.5-72B-Instruct`. Bars represent the accuracy of the models in the y-left axis. Lines represent the **average sentence length** (number of tokens) of each dataset in the y-right axis.

Levenshtein captures surface-level differences directly. Perplexity is also often used to assess how natural or fluent a sentence is (Boisson et al., 2024), but it depends on the language model and its training data. Given the templatic nature of the used datasets, and following the approach of Stowe et al. (2022), we chose Levenshtein distance to measure lexical overlap with the aim of checking the influence of surface features in models' performances.

In our approximation, the higher the Levenshtein distance, the lower the lexical overlap. In this case, we calculated the distance between the premise and the hypothesis sentences to establish whether this is a feature that might correlate with model performance.

As shown in Table 3, LLMs perform best in most cases when evaluated with the original metaphor datasets, which exhibit a higher overlap than their paraphrased versions. This behavious is also observed in absolute terms. Thus, `Qwen2.5-72B-Instruct` achieves the highest results in those datasets that present higher overlap between premises and hypotheses, namely, Figurative-NLI and IMPLI. These results suggest that a substantial degree of overlap influences the models' performance in extracting the inference (Stowe et al., 2022; Naik et al., 2018).

To calculate the average sentence length of each dataset, we computed the arithmetic average of the number of tokens of each sentence. Table 3 and Figure 2 show a noticeable increase in both the number of tokens per sentence and the Levenshtein distance in the literal paraphrases. This seems to agree with the findings from the analysis of the original datasets. As a general trend, longer sentence length and lower lexical overlap tend to co-occur with lower performance of LLMs, and viceversa.

In other words, this quantitative analysis highlights the notable impact that dataset features can have on model performance. Specifically, datasets tailored for the task and developed through lexical substitution present a templatic structure that may explain the high performance of LLMs, despite the presence of metaphorical expressions. Summarizing, LLMs strong performances on the metaphorical pairs may be explained by the high degree of lexical overlap.

6.2 Error Analysis

The aim of conducting a manual error analysis of LLMs' performance is to explore the decrease in accuracy with literal paraphrases, compared to when metaphors are present. We extracted the intersection of correct predictions from the original datasets with those cases that the model fails to predict when evaluated with literal paraphrases. The quantitative information reported in Table 5 is aligned with the distance in performance between original metaphorical datasets and literal paraphrases.

Dataset	Errors	Errors (%)
Meta4XNLI	35	5.85
Fig-QA	193	8.82
Figurative-NLI	57	9.30
FLUTE	13	5.24
IMPLI	77	11.53

Table 5: Intersection of correctly classified pairs from original metaphor datasets and misclassified with paraphrased dataset by `Qwen2.5-72B-Instruct` and CoT prompt. The % represents the percentage of errors with respect to the total number of samples.

Throughout our analysis, we identified several patterns that help explain why evaluating with literal paraphrases leads to poorer performance on the task. We classified the errors into the following categories:

- Paraphrases that still **contain metaphors**: cases where `Command R+` introduced metaphorical expressions in the literal paraphrases.
- Paraphrases that result in a **label shift**: the paraphrase altered the meaning of the original sentence and triggered a change in the inference relationship.
- **Lexical overlap decrease** between premise

Source	Gold	Premise	Hypothesis	Prediction	Error Type
Fig-QA-met	E	Her mind is a steel trap.	She remembers everything, no matter how insignificant.	E	Paraphrase w/ metaphors
Fig-QA-lit		Her mind is very sharp and she has an excellent memory.	She remembers everything, no matter how insignificant.	NE	
FLUTE-met	NE	Came the Spring with all its blunder,	Came the Spring with all its splendor ,	NE	Overlap decreases
FLUTE-lit		Came the Spring with all its blunder,	The season of Spring arrived with its full beauty.	E	Paraphrase w/ metaphors
Meta4XNLI-met	E	NOVEMBER 3, 2000 She was Channel's muse for seven years.	She was Chanel's muse	E	Label shift
Meta4XNLI-lit		NOVEMBER 3, 2000 She worked as a model for Channel for seven years.	She was Chanel's muse	NE	
IMPLI-met	NE	And they felt it rising , rising ,	And they felt the carpet rising , rising	NE	Overlap decreases
IMPLI-lit		They experienced the sensation of something increasing in height.	And they felt the carpet rising , rising	E	

Table 4: Examples of error types found during manual analysis on the evaluation of `Qwen2.5-72B-Instruct` + CoT prompt. **Source** column refers to the source dataset, ***-met** means sentences come from the original metaphoric dataset, while ***-lit** means the sentences come from the literal paraphrases automatically generated. **Gold** column alludes to the inference gold label.

and hypothesis: the paraphrases generate different verbalizations that decrease the lexical overlap (increase editing distance) from the original metaphorical datasets produced by generating them through lexical substitution.

Table 4 provides some examples for each type of error. In the FLUTE example, the only difference between the original premise and hypothesis is the last word "blunder" in the premise and "splendor" in the hypothesis. However, in the generated paraphrase, the similarity between the two sentences decreases, both in terms of length and lexical overlap. Additionally, the paraphrase of the hypothesis introduces the metaphorical verb "to arrive" to refer to the start of the spring season, despite the model being explicitly asked not to do it. As a result, the paraphrase adds an extra difficulty for the model, which fails to predict the correct NLI label, whereas in the original dataset is accurately classified.

Similarly, in the Fig-QA example, the paraphrase of the premise is longer than the original sentence and contains a metaphorical expression ("sharp") to allude to memory.

Another case of the decrease of the lexical overlap is the IMPLI instance. Premise and hypothesis are identical but for the pronoun "it" and the noun "carpet". The paraphrases produces a much longer and distinct premise than the original one, thus the model fails to predict the inference.

In the example of Meta4XNLI, the paraphrased version replaced the metaphor "muse" by "model". In this case, the paraphrase forced a label shift, since being a model does not necessarily imply being a muse, leading to a correct prediction by the model, however, it does not match the original gold_label.

7 Concluding Remarks

The main aim of this work is to test the capabilities of LLMs to understand metaphorical language. In order to do so, we evaluate whether LLMs can predict the inferential relationship between a premise and a hypothesis when metaphorical expressions affect the inference. More specifically, we use multiple available datasets in English, some developed through lexical replacement and others with natural spontaneously generated utterances and framed the task as NLI and QA. In addition, we performed

comprehensive experimentation with various verbalizations and zero- and few-shot settings. Also, we automatically developed a parallel version of the original datasets with literal paraphrases that served as adversarial examples.

The results indicate that LLMs' performance is more influenced by features like lexical overlap and sentence length than by metaphorical content, demonstrating that any alleged emergent abilities of LLMs to understand metaphorical language are the result of a combination of surface-level features, in-context learning, and linguistic knowledge.

Through our experiments, we demonstrate that performance fluctuates remarkably depending on the dataset features, especially lexical overlap (a consequence of data created through lexical substitution) between premise and hypothesis, as well as sentence length. A higher overlap and shorter sentences boost the performance, while naturally occurring sentences, lower overlap, and shorter sentences result in poorer performance. Moreover, LLMs with a smaller number of parameters show almost random performance in zero-shot settings, which can be easily improved with few-shot prompting, while models with more parameters display a more stable performance across prompts. Furthermore, formulating the task as QA and providing few-shot examples enables superior performance, especially when combined with CoT, which helps to outperform any other scenario, including some strong baselines. We hypothesize that this is due to the post-training of the instruct models (Ouyang et al., 2022; Touvron et al., 2023).

Our manual error analysis shows that automatic generation of literal paraphrases requires exhaustive human evaluation, since models still include metaphors in the newly generated sentences. We argue that this behavior reveals LLMs' lack of ability to discriminate between metaphorical and literal expressions, although more research is required, perhaps with manually generated paraphrases in future work.

We believe that this work provides critical insights into LLMs' current capabilities and limitations in processing figurative language, highlighting the need for more realistic evaluation frameworks in metaphor interpretation tasks.

8 Limitations

This work expands the scope of metaphor interpretation evaluation, moving beyond the conventional and limited approaches seen in recent research. We have broadened the evaluation to several models, diverse resources, and various experimental scenarios. While we acknowledge the limitations of our study, future research could benefit from manually inspecting the generated paraphrases. Additionally, the datasets available for assessing metaphor interpretation remain relatively small in size compared to resources for other NLP tasks. Furthermore, extending the analysis to multiple languages would be valuable, but this also requires the existence of open resources with metaphorical data, which is currently limited and scarce. We hope that this comprehensive assessment will encourage the research community to create valuable and diverse resources that enable a reliable assessment of the emergent capabilities of LLMs to understand metaphorical language in multifaceted scenarios.

Acknowledgments

We are grateful to the free credits awarded by the Cohere For AI Research Grant Program[2]. Elisa Sanchez-Bayona is funded by the UPV/EHU PIF20/139 grant.

We would also like to acknowledge the funding received by the following MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and ERDF A way of making Europe; (ii) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR.

References

Rodrigo Agerri. 2008. Metaphor in Textual Entailment. In *COLING*, pages 3–6.

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Daniel Baleato Rodríguez, Verna Dankers, Preslav Nakov, and Ekaterina Shutova. 2023. Paper bullets:

[2] https://cohere.com/research/grants

Modeling propaganda with the help of metaphor. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 472–489, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction artifacts in metaphor identification datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6581–6590, Singapore. Association for Computational Linguistics.

Joanne Boisson, Asahi Ushio, Hsuvas Borkakoty, Kiamehr Rezaee, Dimosthenis Antypas, Zara Siddique, Nina White, and Jose Camacho-Collados. 2024. How are metaphors processed by language models? the case of analogies. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 365–387, Miami, FL, USA. Association for Computational Linguistics.

Danushka Bollegala and Ekaterina Shutova. 2013a. Metaphor interpretation using paraphrases extracted from the web. *PLOS ONE*, 8(9):1–10.

Danushka Bollegala and Ekaterina Shutova. 2013b. Metaphor Interpretation Using Paraphrases Extracted from the Web . *PloS one*, 8(9):e74304.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021a. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021b. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.

Iulia Comșa, Julian Eisenschlos, and Srini Narayanan. 2022. MiQA: A benchmark for inference on metaphorical questions. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 373–381, Online only. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander Vladislavovitch Dmitrijev, Elena Sergeevna Krupnova, and Anastasia Aleksandrovna Protopopova. 2024. Metaphors and analogies in the context of large language models. In *Scenarios, Fictions, and Imagined Possibilities in Science, Engineering, and Education*, pages 326–341, Cham. Springer Nature Switzerland.

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, and 516 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Bernadeta Griciūtė, Marc Tanti, and L. Donatelli. 2022. On the cusp of comprehensibility: Can language models distinguish between metaphors and nonsense?

Proceedings of the 3rd Workshop on Figurative Language Processing (FLP).

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. i. *New Phytologist*, 11(2):37–50.

Arthur M. Jacobs. 2018. The gutenberg english poetry corpus: Exemplary quantitative narrative analyses. *Frontiers in Digital Humanities*, 5.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Ruben Sinu Kurian, Chandramani Chaudhary, Abhay Unni Nambiar, and Abhina Sunny. 2025. Metan: Metaphoric temporal attention network for depression detection on social media. In *Web Information Systems Engineering – WISE 2024*, pages 90–104, Singapore. Springer Nature Singapore.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

George Lakoff and Mark Johnson. 1980. Metaphors We Live By.

Jens Lemmens, Ilia Markov, and Walter Daelemans. 2021. Improving hate speech type and target detection with hateful metaphor features. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 7–16, Online. Association for Computational Linguistics.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

H. J. Levesque, E. Davis, and L. Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning (KR-12)*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Rui Mao, Chenghua Lin, and Frank Guerin. 2021. Interpreting verbal metaphors by paraphrasing. *Preprint*, arXiv:2104.03391.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Michael Mohler, Marc T. Tomlinson, and D. Bracewell. 2013. Applying textual entailment to the interpretation of metaphor. *2013 IEEE Seventh International Conference on Semantic Computing*, pages 118–125.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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

Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021. A howling success or a working sea? testing what BERT knows about metaphors. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Geetanjali Rakshit and Jeffrey Flanigan. 2023. Does the "most sinfully decadent cake ever" taste good? answering yes/no questions from figurative contexts. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 926–936, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2024. Meta4xnli: A crosslingual parallel corpus for metaphor detection and interpretation. *Preprint*, arXiv:2404.07053.

Ekaterina Shutova. 2010. Automatic Metaphor Interpretation as a Paraphrasing Task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037. Association for Computational Linguistics.

Ekaterina Shutova, T. V. D. Cruys, and Anna Korhonen. 2012. Unsupervised metaphor paraphrasing using a vector space model. In *International Conference on Computational Linguistics*.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical Metaphor Processing. *Computational Linguistics*, 39(2):301–353.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

G. Steen. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging evidence in language and communication research. John Benjamins Publishing Company.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. 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)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.

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. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2020. Figure me out: A gold standard dataset for metaphor interpretation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5810–5819, Marseille, France. European Language Resources Association.

Dongyu Zhang, Nan Shi, Ciyuan Peng, Abdul Aziz, Wenhong Zhao, and Feng Xia. 2021. Mam: A metaphor-based approach for mental illness detection. In *Computational Science – ICCS 2021*, pages 570–583, Cham. Springer International Publishing.

A Data Samples

Dataset	Premise	Hypothesis	Label
Fig-QA	The girl had the flightiness of a sparrow	The girl was very fickle.	entailment
	The girl had the flightiness of a sparrow	The girl was very stable.	not_entailment
	The girl had the flightiness of a rock	The girl was very stable.	entailment
	The girl had the flightiness of a rock	The girl was very fickle.	not_entailment
IMPLI	he absorbed the costs for the accident.	he paid for the costs for the accident.	entailment
	he absorbed the costs for the accident.	he absorbed the sunlight after the accident.	not_entailment
	the sales tax is absorbed into the state income tax.	the sales tax is incorporated into the state income tax.	entailment
	the sales tax is absorbed into the state income tax.	the dirty water is absorbed into the clean water.	not_entailment
FLUTE	It is sad to observe the consequences of ignorance.	It is sad to observe the fruits of ignorance.	entailment
	It is amusing to observe the fruits of ignorance.	It is sad to observe the fruits of ignorance.	not_entailment
	That guy knows how to charm and manage people and work his way up the social ladder.	This young man knows how to climb the social ladder.	entailment
	This young man is stuck in the middle of this social ladder.	This young man knows how to climb the social ladder.	not_entailment
Figurative-NLI	The moon winked back at itself from the lake's surface	The moon reflected back at itself from the lake's surface	entailment
	The moon winked back at itself from the lake's surface	The moon absorbed back at itself from the lake's surface	not_entailment
	The company released him after many years of service	The company fired him after many years of service	entailment
	The company released him after many years of service	The company hired him after many years of service	not_entailment
Meta4XNLI	28 In some cases longer outages are needed.	Sometimes longer outages are needed.	entailment
	The Great Depression hit California hard.	California's economy has always thrived.	not_entailment
	My father lived many years in Africa and his stay there had a great impact on me.	My father's stay in Africa conditions my life.	entailment
	You are very outgoing and open with the fans.	You meet with your fans after each concert.	not_entailment

Table 6: Examples from original datasets used in our evaluation.

B Evaluation Prompts

NLI-zero
Say which is the inference relationship between these two sentences. Please, answer between "entailment" or "other". {Premise} -> {Hypothesis}:
NLI-few
Say which is the inference relationship between these two sentences. Please, answer between "entailment" or "other". Here you have some examples: I am open -> I am friendly: entailment. My heart is broken -> I am happy: other. {Premise} -> {Hypothesis}:
QA-zero
Are these two sentences entailed? Please, answer between "yes" or "no". {Premise} -> {Hypothesis}:
QA-few
Are these two sentences entailed? Please, answer between "yes" or "no". Here you have some examples: I am open -> I am friendly: yes. My heart is broken -> I am happy: no. {Premise} -> {Hypothesis}:
CoT
You are an expert linguist and your task is to annotate sentences for the task of Natural Language Inference. This task consists in determining if a first sentence (premise) entails or not the second sentence (hypothesis). Please, limit your answer to "yes" or "no". Here you have a few examples: Premise: I am an open person. Hypothesis: I am friendly. Answer: yes Premise: My heart is broken. Hypothesis: I am happy. Answer: no. {Premise}: {Hypothesis}: {Answer}:
NLI mapping
{"entailment": "entailment", "other": "not_entailment"}
QA mapping
{"yes": "entailment", "no": "not_entailment"}

Table 7: Prompts and response mappings for NLI, QA, CoT, zero- and few-shot evaluation setups.

C Literal Paraphrase Generation Prompt

Prompt: Please, generate a literal paraphrase of this sentence. The sentence contains a metaphorical expression. Your task is to rewrite the sentence so it does not contain any metaphors. The generated sentence must have the same meaning as the original. Please, DO NOT include metaphorical or idiomatic expressions in the generated sentence. Answer only with the literal sentence. Original sentence: [metaphorical_sentence] Paraphrase:

Table 8: Prompt for `Command R+` and `Mistral-7B-Instruct` to generate literal paraphrases from sentences with metaphorical expressions.

D Zero-/Few-shot Evaluation with Literal Paraphrases

			Llama-3-8B-Instruct					Mistral-7B-Instruct				
Dataset	**Samples**	**Baseline**	NLI-Zero	NLI-Few	QA-Zero	QA-Few	CoT	NLI-Zero	NLI-Few	QA-Zero	QA-Few	CoT
Meta4XNLI-Cmdr	598	76.73	53.84	57.20	49.16	62.04	**76.76**	60.37	68.22	45.65	73.58	72.74
Meta4XNLI-Mistral			39.47	62.04	33.44	50.17	**71.91**	49.67	65.55	37.96	60.03	68.30
Fig-QA-Cmdr	2188	61.00	50.91	56.03	61.43	62.71	**81.58**	41.50	68.42	44.88	78.38	80.67
Fig-QA-Mistral			50.73	61.75	48.67	58.87	**79.39**	41.04	66.54	42.60	74.91	77.70
Figurative-NLI-Cmdr	613	88.09	51.39	50.90	54.65	59.22	78.30	37.85	73.90	33.12	83.85	**82.22**
Figurative-NLI-Mistral			48.61	61.01	43.23	58.73	76.35	40.62	71.94	38.66	75.86	**81.24**
FLUTE-Cmdr	248	81.80	53.63	52.01	56.05	54.43	79.44	46.37	69.35	54.03	75.81	**83.06**
FLUTE-Mistral			50.81	50.40	49.19	56.85	75.81	31.05	66.53	35.08	73.79	**77.82**
IMPLI-Cmdr	668	85.55	44.76	45.06	39.22	61.38	**81.87**	33.83	68.86	51.05	80.39	79.79
IMPLI-Mistral			41.44	48.05	25.37	54.35	**79.88**	33.33	66.97	47.6	73.87	79.13
Avg-Cmdr	-	78.63	50.91	52.24	52.10	59.96	**79.59**	43.98	69.75	45.75	78.40	**79.70**
Avg-Mistral	-		46.21	56.65	39.98	55.79	76.67	39.14	67.51	40.38	71.69	76.84

Table 9: Accuracy of evaluation results with automatic literal paraphrases. Meta4XNLI baseline evaluation framed as NLI with XNLI-RoBERTa fine-tuned on Meta4XNLI train set. Fig-QA baseline evaluation framed as Winograd-style QA task with GPT-3 Ada through prompting. Figurative-NLI baseline evaluation framed as NLI with RoBERTa-large. FLUTE baseline framed as NLI with T5 fine-tuned on e-SNLI (Camburu et al., 2018). IMPLI baseline evaluation framed as NLI with gold standard examples and RoBERTa-large. In bold, best model for each evaluation dataset. In underscore, the best result for each version of each dataset, that is, paraphrases generated with `Command R+` (Cmdr) or `Mistral-7B-Instruct` (Mistral).

E Complete Evaluation Results

Prompt	Dataset	Llama-3-Instruct 8B	Llama-3-Instruct 70B	Qwen-2.5-Instruct 7B	Qwen-2.5-Instruct 72B	Gemma-3-it 4B	Gemma-3-it 27B	Mistral-Instruct 7B
NLI-zero	Meta4XNLI-met	**55.18**	**85.12**	**86.12**	**89.80**	**52.00**	**62.21**	59.53
	Meta4XNLI-lit	53.84	82.94	82.94	87.62	51.67	60.03	**60.37**
	Fig-QA-met	50.27	**87.16**	68.55	**77.58**	42.38	**71.80**	41.41
	Fig-QA-lit	**50.91**	83.45	**71.53**	77.15	**46.30**	68.42	**41.50**
	Figurative-NLI-met	49.43	**90.54**	**85.64**	**93.47**	**46.98**	**88.74**	**47.96**
	Figurative-NLI-lit	**51.39**	86.95	78.96	87.60	40.78	70.30	37.85
	FLUTE-met	51.21	83.87	**79.43**	**87.50**	**45.56**	**76.61**	**46.77**
	FLUTE-lit	**53.63**	**87.60**	78.47	87.44	41.13	71.13	46.37
	IMPLI-met	42.37	**94.89**	**89.19**	**93.99**	**47.74**	**89.64**	**37.13**
	IMPLI-lit	**44.76**	86.08	82.93	86.83	39.67	66.61	33.83
NLI-few	Meta4XNLI-met	**60.37**	**85.62**	**88.46**	**90.63**	**78.76**	**90.13**	**70.07**
	Meta4XNLI-lit	57.20	83.44	85.12	87.79	76.92	87.96	68.22
	Fig-QA-met	**60.65**	**85.24**	73.95	**81.03**	66.54	**77.47**	66.32
	Fig-QA-lit	56.03	81.95	**75.87**	79.57	**72.35**	76.64	**68.42**
	Figurative-NLI-met	**64.27**	**91.52**	**92.01**	**94.13**	**81.89**	**84.99**	**75.37**
	Figurative-NLI-lit	50.90	86.79	83.20	88.09	78.14	81.89	73.90
	FLUTE-met	51.61	84.27	80.24	**87.90**	71.77	79.43	66.53
	FLUTE-lit	**52.01**	**86.95**	**83.20**	87.76	**75.81**	**82.05**	**69.35**
	IMPLI-met	**53.59**	**94.29**	**90.99**	**94.14**	**80.78**	**90.39**	**74.55**
	IMPLI-lit	45.06	84.58	85.03	86.98	74.70	84.58	68.86
QA-Zero	Meta4XNLI-met	45.65	**86.79**	**86.96**	**89.13**	**80.77**	**89.30**	**50.84**
	Meta4XNLI-lit	**49.16**	85.62	83.11	87.12	78.93	86.29	45.65
	Fig-QA-met	50.18	**89.21**	69.15	**82.91**	69.38	**81.49**	41.86
	Fig-QA-lit	**61.43**	82.86	**71.80**	79.66	**73.72**	78.15	**44.88**
	Figurative-NLI-met	43.39	**88.42**	**86.95**	**93.80**	**87.44**	**89.40**	**56.28**
	Figurative-NLI-lit	**54.65**	85.15	82.71	87.60	80.91	82.54	33.12
	FLUTE-met	**57.66**	**84.68**	81.45	86.69	75.40	79.43	50.81
	FLUTE-lit	56.05	85.32	**82.54**	**87.60**	**78.63**	**82.71**	**54.03**
	IMPLI-met	**39.37**	**92.94**	**89.64**	**94.59**	**84.23**	**92.34**	**59.61**
	IMPLI-lit	39.22	85.48	84.73	86.98	82.33	86.82	51.05
QA-Few	Meta4XNLI-met	61.04	**85.95**	**89.13**	**89.97**	**81.60**	**88.79**	72.07
	Meta4XNLI-lit	**62.04**	84.11	82.61	86.79	79.26	86.97	**73.58**
	Fig-QA-met	58.04	**89.08**	72.21	**85.05**	68.46	**81.99**	74.91
	Fig-QA-lit	**62.71**	83.36	**73.86**	81.17	**73.35**	80.62	**78.38**
	Figurative-NLI-met	**62.81**	**91.35**	**90.86**	**94.13**	**83.85**	**90.37**	**86.95**
	Figurative-NLI-lit	59.22	86.79	84.50	88.09	76.83	84.99	83.85
	FLUTE-met	**62.50**	85.48	76.21	87.10	73.39	**85.48**	**79.03**
	FLUTE-lit	54.43	**86.95**	**84.83**	**88.09**	**77.82**	85.32	75.81
	IMPLI-met	**60.48**	**93.54**	**89.79**	**93.69**	**78.22**	**93.39**	**84.88**
	IMPLI-lit	61.38	84.13	84.13	85.93	75.60	87.12	80.39
CoT	Meta4XNLI-met	**78.26**	**87.96**	**89.80**	**90.47**	**79.93**	**90.97**	**73.41**
	Meta4XNLI-lit	76.76	85.45	86.96	88.63	79.10	88.46	72.74
	Fig-QA-met	76.17	**89.48**	74.91	**85.24**	70.93	**81.58**	76.42
	Fig-QA-lit	**81.58**	85.51	**76.42**	80.57	**78.20**	79.84	**80.67**
	Figurative-NLI-met	**86.13**	**92.80**	**90.21**	**95.43**	**86.13**	**90.37**	**88.25**
	Figurative-NLI-lit	78.30	87.77	84.18	88.58	80.42	85.81	82.22
	FLUTE-met	**83.87**	85.08	81.85	**87.50**	**79.03**	**85.08**	82.26
	FLUTE-lit	79.44	**85.89**	**84.18**	86.69	77.82	83.48	**83.06**
	IMPLI-met	**82.93**	**94.31**	**90.24**	**95.04**	**84.98**	**93.99**	**84.28**
	IMPLI-lit	81.87	86.68	86.98	87.57	83.23	88.77	79.79

Table 10: Evaluation accuracy scores with all models and prompts. In bold, subset met/lit with best results.