Investigating the Impact of Conceptual Metaphors on LLM-based NLI through Shapley Interactions

Meghdut Sengupta[1], Maximilian Muschalik[2], Fabian Fumagalli[3], Barbara Hammer[3], Eyke Hüllermeier[2], Debanjan Ghosh[4], Henning Wachsmuth[1]
[1]Leibniz University Hannover, Institute of Artificial Intelligence [2]LMU Munich, MCML
[3]Bielefeld University, CITEC [4]Analog Devices
m.sengupta@ai.uni-hannover.de

Abstract

Metaphorical language is prevalent in everyday communication, often used unconsciously, as in "*rising* crime." While LLMs excel at identifying metaphors in text, they struggle with downstream tasks that implicitly require correct metaphor interpretation, such as natural language inference (NLI). This work explores how LLMs perform on NLI with metaphorical input. Particularly, we investigate whether incorporating conceptual metaphors (source and target domains) enhances performance in zero-shot and few-shot settings. Our contributions are two-fold: (1) We create a new dataset, FLUTE.st, extending metaphorical texts in an existing NLI corpus by annotations of source and target domains; and (2) we conduct an ablation study using Shapley values and interactions to assess the extent to which LLMs interpret metaphorical language correctly in NLI. Our results indicate that incorporating conceptual metaphors often improves task performance.

1 Introduction

Metaphorical language is pervasive in the flow of everyday life conversations. While metaphors can be observed in explicit cases such as "she was the *light* of my life", the meaning manifestation of conventionalized metaphors such as "*tax* the rich" is more fundamentally grounded in language (Kövecses, 2010; Gábor, 2014). According to Lakoff and Johnson (2003), metaphorical meaning construction is the result of a conceptual metaphor mapping that connects one domain to another: A concept is taken from a *source domain* to explain a *target domain*. For example, in "Gun addicts increasingly realize that society is rejecting them", the source domain is *addiction* and the target domain *guns*.

Despite progress in computational metaphor interpretation (Leong et al., 2018; Tong et al., 2021), large language models (LLMs) underperform on downstream tasks that require correct interpretation of figurative language such as metaphors (Dmitrijev et al., 2024; Gallipoli and Cagliero, 2025). Skrynnikova (2024) highlight that LLMs imitate data rather than reasoning analogically, a key requirement for successful metaphor comprehension. Similarly, Comșa et al. (2022) attribute this struggle to dependency on contextual variables. Hence, inspired by advances on Shapley-based analyses in explainable AI (Muschalik et al., 2024; Fumagalli et al., 2025), we study two research questions on LLM interpretability in NLP tasks on metaphorical input: **(1)** How well do LLMs handle implicit metaphorical language in downstream tasks such as natural language inference (NLI)? **(2)** Does providing source and target domains enhance NLI performance, and how does this information interact with metaphorical texts?

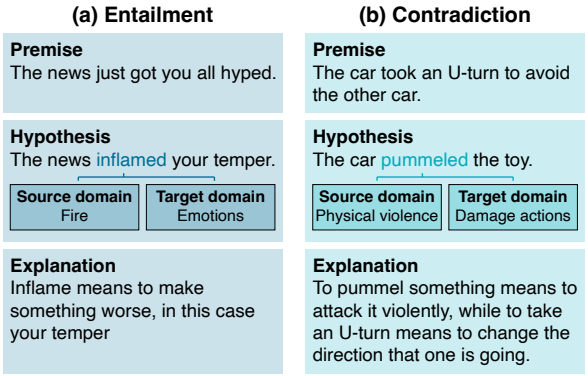

Figure 1: Examples of (a) *entailment* and (b) *contradiction* NLI instances after the re-annotation of parts of the FLUTE dataset with source and target domains. Only the hypotheses are metaphorical, not the premises.

To this end, we make two main contributions: **(1)** FLUTE.st:[1] We extend all metaphorical samples in the FLUTE corpus (Chakrabarty et al., 2022) by annotations of source and target domains (where *s* and *t* stand for source domain and target domain

[1]FLUTE.st is available at: https://github.com/webis-de/emnlp25-shapley-interactions/tree/master/data/flute/csv

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17393–17403
November 4-9, 2025

respectively). **(2)** We evaluate the impact of these domains on five LLMs in zero-shot and few-shot NLI for given pairs of premise and metaphorical hypothesis (see the examples in Figure 1). Analyzing the results using Shapley values and interactions (Muschalik et al., 2024), we find that incorporating conceptual metaphors improves LLM performance in 70% of all experiments.

2 Related Work

Existing NLP methods dealing with metaphors interpret them by cognitively decoding conceptual metaphors. Shutova et al. (2013) applied hierarchical clustering and conceptual mappings, while Stowe et al. (2021) paraphrased metaphors using FrameNet-based source and target domains. Chakrabarty et al. (2021) leveraged these domains for textual entailment. The actual inference of conceptual metaphors was first studied by Sengupta et al. (2022) who used contrastive learning to infer source domains from metaphorical texts. Sengupta et al. (2023) extended this by multitask learning to jointly predict source domains and the aspects highlighted by metaphors.

Unlike prior work, we focus on interpreting *implicit* metaphors, beyond their *explicit* detection. We use NLI, a core NLP reasoning task (Storks et al., 2020), where metaphorical meaning is embedded naturally, as in everyday language. This enables a more realistic test of LLMs' metaphor understanding, viewed here as synthesizing cognitively grounded facts (Sanchez-Bayona and Agerri, 2024; Stowe et al., 2022). We thus analyze LLMs' metaphor interpretation via NLI. We build on the FLUTE dataset (Chakrabarty et al., 2022), where many NLI hypotheses are metaphorical. Prior work benchmarked transformers on FLUTE, revealing the difficulty of implicit figurative tasks. We emphasize the importance of context and conceptual metaphor interactions, addressing this by analyzing LLM performance using Shapley-based methods (Muschalik et al., 2024).

3 Data

In this section, we present our extension, FLUTE.st, of the FLUTE dataset by source and target domains.

3.1 Metaphorical Texts

FLUTE is a benchmark to understand figurative language (Chakrabarty et al., 2022). Consisting of 9000 pairs of literal and figurative sentences, it comes with labels indicating whether the sentences entail or contradict each other, along with textual explanations for the labels (Figure 1). This information contextualizes figurative expressions in the data, facilitating meaning comprehension. While FLUTE covers four figurative categories (sarcasm, simile, metaphor, and idiom), we focus solely on the metaphorical samples. Chakrabarty et al. (2022) prompted GPT-3 to generate paraphrases and contradictions for each metaphor, resulting in 750 entailment and 750 contradiction sentence pairs.

3.2 Annotation of Source and Target Domains

We extended all 1500 metaphorical samples from FLUTE by annotations of source and target domains via in-context learning (Tan et al., 2024) in an AI-supported human annotation process.

First, we prompted GPT-4 (Achiam et al., 2023) and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) in a 10-shot setup using three experimental settings, resulting in six different annotation variations: **(1)** without the metaphorical span in the text, **(2)** with the metaphorical span, and **(3)** with the metaphorical span and the explanation of why the source and target domains are determined as they are. To inform the prompts, we used source and the target domains from the corpus of Gordon et al. (2015). See Appendix A.1.1 for details on the prompts.

Then, we recruited two human annotators with background in metaphor research to evaluate the same random subset of 200 samples for each variation. In line with Chakrabarty et al. (2022), we asked them to give a *yes* for an optimal answer, a *weak yes* for a near-optimal answer, a *weak no* for a suboptimal answer, and a *no* for a wrong answer. For our *main* experiments, we chose the annotated source and target domains for all FLUTE samples from the variation with highest Cohen's κ agreement, namely GPT-4 without the metaphor or the explanation in the prompt: $\kappa = 0.325$ for source domains, and $\kappa = 0.323$ for target domains. For the subjective task of identifying source and target domains of metaphors, similar agreement values have been observed in past NLP research (Sengupta et al., 2024). In line with previous research, we argue that this rather indicates the hardness of the task at hand for both humans and AI.

4 Approach

We now describe the NLI experiments as well as the Shapley-based analysis carried out on their results.

Model	Mistral-7B		Llama-3.2-3B		Gemma-2-27B-IT		DeepSeek-RDQ-32B		Llama-3.3-70B-I	
	0-shot	10-shot	0-shot	10-shot	0-shot	10-shot	0-shot	10-shot	0-shot	10-shot
Premise+Hypothesis	0.632	0.784	0.654	0.723	0.913	**0.869**	**0.799**	0.889	0.903	0.927
+ Source Domain	0.547	0.689	0.648	0.677	0.890	0.864	0.632	0.829	0.814	0.897
+ Target Domain	0.681	0.766	0.652	0.711	0.895	0.815	0.714	0.860	0.793	0.887
+ Explanation	0.839	0.858	**0.695**	0.835	0.873	0.688	0.569	0.863	0.849	0.912
+ Source D. + Target D.	0.643	0.739	0.645	0.676	0.894	0.813	0.685	0.813	0.786	0.886
+ Source D. + Explanation	0.803	**0.865**	0.670	0.842	0.877	0.670	0.520	0.843	0.794	0.894
+ Target D. + Explanation	0.833	0.857	0.670	**0.852**	0.848	0.724	0.740	0.812	0.586	0.841
+ All	**0.845**	0.855	0.651	0.834	**0.947**	0.603	0.745	**0.932***	**0.946***	**0.969***

Table 1: F_1-scores of all five LLMs in 0-shot and 10-shot NLI using only *premise+hypothesis* or combined with all possible subsets of *source domain*, *target domain*, and *explanation*. Adding an explanation consistently improves performance, but the best results are achieved with *all* in addition (* marks $p < .05$ in a Wilcoxon rank-sum test).

4.1 Natural Language Inference

For NLI, we prompted five different LLMs in zero-shot and 10-shot prompting settings on the entire FLUTE.st dataset: Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Llama-3.2-3B-Instruct (Dubey et al., 2024), Gemma-2-27B-IT (Team, 2024), DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI, 2025), and Llama-3.3-70B-Instruct (Grattafiori et al., 2024). The prompts covered subsets of five input components: *premise phrases*, *hypothesis phrases*, the *source domain*, the *target domain*, and the *explanation*. Details on the prompts in Appendix A.1.2.

4.2 Shapley-based Analysis

Shapley values (Shapley, 1953) are used to quantify the influence of entities, such as features or data points (Muschalik et al., 2024). They are based on a set of players $N := \{1, \ldots, n\}$ together with a *cooperative game* $\nu : 2^N \to \mathbb{R}$, mapping a coalition $S \subseteq N$ to a real-valued *payout* $\nu(S)$.

We define a *local* and a *global* explanation game. The local game is used to explain the LLM's NLI prediction for a single instance, by capturing the prediction across all possible combinations of input components. The set of entities N_l contains all five components of the given input. The payout for a subset $S \subseteq N_l$ is given by the prediction of the LLM, where all remaining components in $N_l \setminus S$ are masked. In contrast, the global explanation game is used to explain the performance of the LLM provided different input components. Here, we require that the premise and hypothesis are always present to the LLM, that is, N_g is reduced to the source domain, target domain, and explanation. Finally, the payout for a coalition of $S \subseteq N_g$ is the F_1-score of the LLMs across all instances.

Shapley Value and Interactions We use the Shapley value (SV) and Shapley interactions (SIs) to analyze the impact of components in N_l and N_g. Grounded in mathematical axioms, SV (Shapley, 1953) assigns *fair* contributions to individual components. SIs (Lundberg et al., 2020) further assign contributions to interactions between two or more components (Bordt and von Luxburg, 2023). The efficiency axiom ensures that the sum of all contributions equals the difference in the overall payout $\nu(N)$ and the baseline payout $\nu(\emptyset)$, so the prediction of the LLM (local) or the F_1-score (global).

For the local explanation game, a positive SV indicates that a component contributed to the prediction of *entailment* (on average), a negative SV a contribution to *contradiction*. For the global explanation game, a positive SV indicates that the presence of the respective component yields a gain in F_1-score, a negative value a decrease. Moreover, a positive SI indicates that the LLM obtains additional information from the joint presence of both components, and a negative SI that both components contain redundant information for the LLM.

Phrase Level Chunking Since computing SVs and SIs for n input tokens requires evaluating 2^n coalitions, we chunk the premises and hypotheses into noun, verb, and prepositional phrases (instead of individual tokens) using the *nltk.org* chunker. Phrase chunking allows us to inspect how the context grouped into phrases affects LLM performance, as far as local explanations are concerned. Given the average number of phrases per premise-hypothesis pair in FLUTE.st is 9.5, we chose $n = 9$ phrases as the maximum number of input components. Together with the two domains and the explanation, the maximum total number is thus 12 for the task, with 690 samples for each experiment.

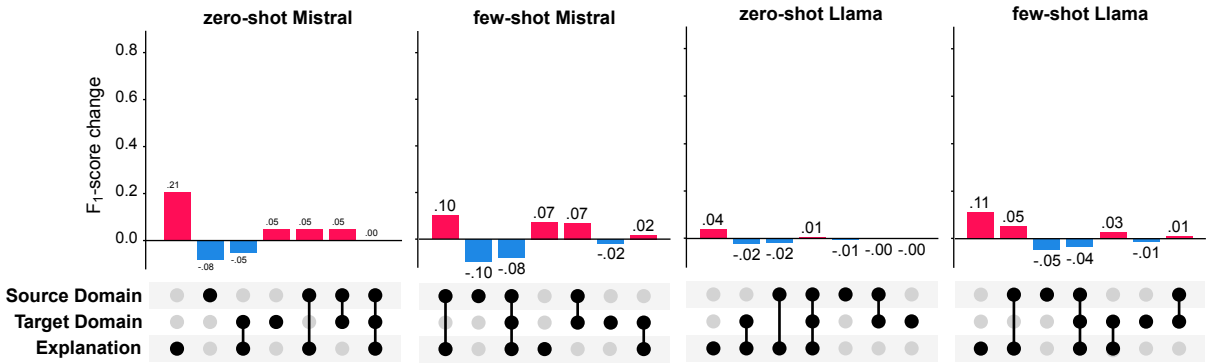

Figure 2: UpSet plots (Lex et al., 2014) of Shapley interactions for the global explanation games, showing performance changes when adding components. Particularly explanations (second bar) notably increase F_1-score, but for the majority of experiments, the best F_1-score is obtained by also adding source and/or target domains.

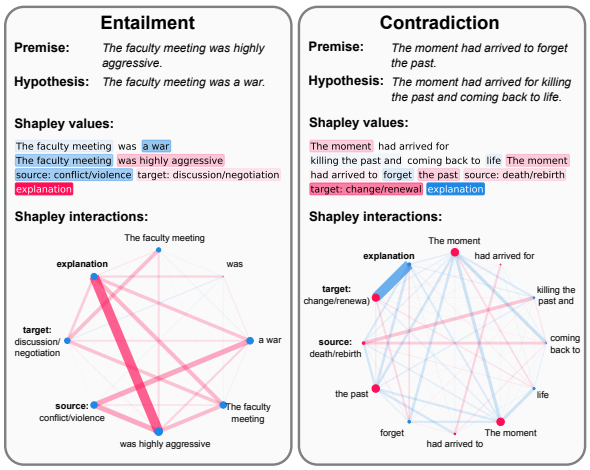

Figure 3: Shapley value (SV) and Shapley interaction (SI) explanations for entailment and contradictions, as predicted by few-shot Mistral (positive contributions in red and negative contributions in blue).

5 Results and Analysis

Table 1 shows the NLI results using all possible subsets of components added to the baseline input (premise and hypothesis only), including significance results from a Wilcoxon rank-sum test with Bonferroni-adjusted p-values. The explanation improves F_1-scores in all settings. However, the best performance is obtained when adding the source and/or target domain, in 70% of all experiments. While the largest (DeepSeek 32B, Llama 70B) perform strongest, for computational limitations, we restrict the following Shapley-based analyses to the smallest models (Mistral 7B, Llama 3B).

Local Explanations Figure 3 shows local SV and SI explanations of two example instances correctly predicted as entailment and contradiction by few-shot Mistral. From the SVs, we observe mixed but high contributions of the source domain, target domain, and explanation. In both cases, the explanation has a high influence toward the predicted class. The SI explanations suggest that higher-order effects greatly influence the predictions. Individually, the components point toward the wrong class. However, the plot reveals a high Shapley Value (SV) for the metaphor "a *war*" and its source domain *conflict/violence* in the entailment case, and for "*killing* the past" and its source domain *death/rebirth* in the contradiction case. From the perspective of Conceptual Metaphor Theory (CMT), this indicates that the model leverages the systematic mappings between source and target domains to anchor its reasoning. Rather than treating these metaphors as isolated lexical items, the model appears to recognize their entailments within the broader conceptual structures they evoke—for example, understanding *war* in terms of struggle or opposition, and *death/rebirth* in terms of negation and renewal. The elevated SVs highlight that these mappings are not peripheral but central to the model's decision-making process, shaping the direction of inference. This suggests that the model is engaging in a form of metaphorical reasoning that parallels human interpretive strategies, constructing meaning by projecting structure from concrete domains onto abstract targets. Consequently, the results provide evidence that the model's predictions in figurative NLI may arise from mechanisms resembling conceptual integration and reasoning, rather than simple pattern imitation.

Global Explanations Figure 2 shows the contributions of the domains and the explanation in the global explanation game. We see a strong benefit of the explanation, which substantially increases

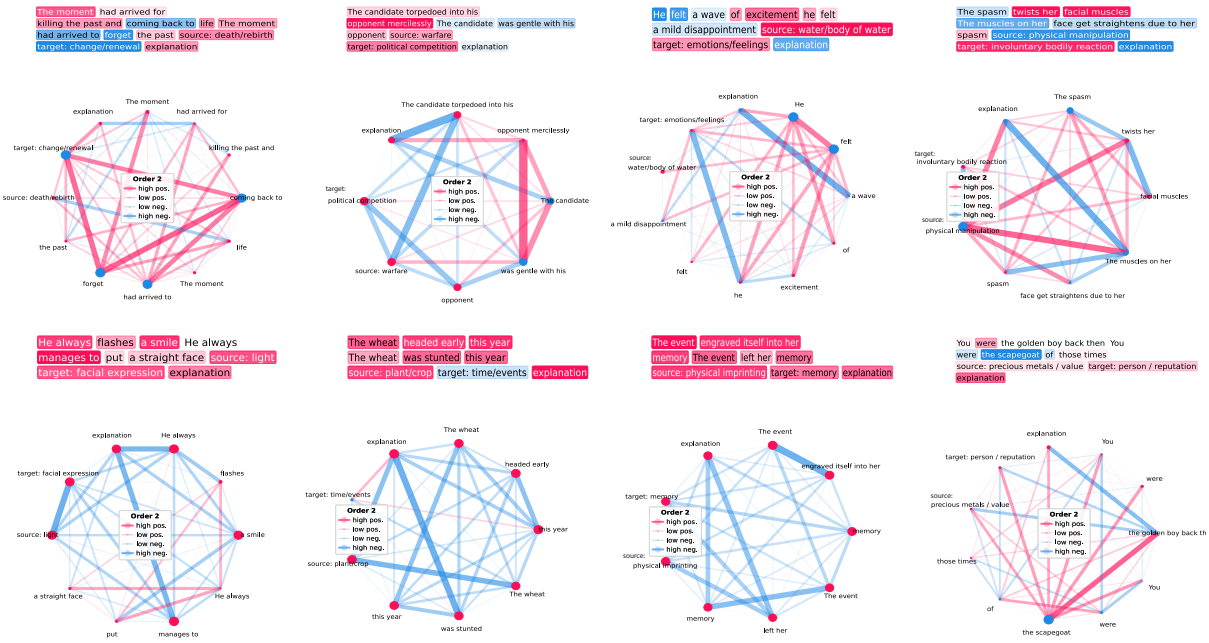

Figure 4: This figure presents a sample of the failed predictions by the Llama-3.2-3B-Instruct model. The shapley-based local explanations indicate that in the zero-shot setting, the small model confuses how to process the different information provided in the input sequence, and particularly the explanation contributes negatively.

performance. Although the domains receive small contributions alone, performance is improved when using both of them (analysis of the zero-shot Llama setting is provided in Section 5.1). In line with the local explanations, this supports our hypothesis of the improvement of model performance with the additional information of the conceptual metaphors.

Summary Our experiments show that LLMs benefit from conceptual metaphors, and strategically incorporating them can further enhance performance both individually and in combination with other components.

5.1 Error Analysis

This section analyzes failure cases of the Llama-3.2-3B-Instruct model in the zero-shot setting. We find that both source and target domains negatively affect performance, with local explanations showing a consistent pattern of moderately high to strong negative influence on the premise–hypothesis interactions (see Figure 4). For example, in the second case (Premise: "The candidate was gentle with his opponent"; Hypothesis: "The candidate torpedoed onto his opponent mercilessly") and the third case (Premise: "He felt a mild disappointment"; Hypothesis: "He felt a wave of excitement"), the explanation interacts with the metaphor in the hypothesis in a way that reduces model performance, contrary to the expected positive contribution. Similar patterns occur across other examples, where varying contributions from different components (premise, hypothesis, explanations, and domains) consistently result in negative effects when metaphors appear in the hypothesis. We hypothesize that in the zero-shot setting, the lack of examples leaves the model uncertain about using input information, causing explanations with metaphors to contribute negatively. This indicates that small models' zero-shot performance in figurative NLI warrants further study.

6 Conclusion

This paper studied the role of conceptual metaphors (source and target domains) in LLM performance on natural language inference (NLI) in case of texts with implicit metaphorical language. To this end, we extended the FLUTE dataset with source and target domain annotations in our new dataset FLUTE.st. Based on the dataset, we ran zero-shot and few-shot ablations across five LLMs. Our results suggest that explanations of NLI combined with conceptual metaphors yield the best performance, supported by a Shapley-based analysis. Future work should explore advanced techniques leveraging metaphorical knowledge to improve LLMs' understanding of implicit metaphors.

7 Limitations

Shapley-based analyses, although proven to be effective in existing research on explainable AI, are computationally costly. This is because for an input length of n, the total number of coalitions to be tested is 2^n, meaning the model will have 2^n queries for that input.

To keep the computational effort manageable, we limited our approach in multiple ways: We decomposed premises and hypotheses into phrases instead of tokens), restricted the maximum number of phrases to 9, carried out most experiments with relatively small LLMs only, and used the fast-inference vLLM for our pipeline.[2] For our experiments, we split our dataset into four parts, with each part having approximately 172 instances. With an A100 GPU and 80 GB RAM, it requires approximately 40 hours for a model to run just any one part of that dataset, when the additional components for the ablation study are included in the input sequence. We stress that, by neither performing experiments on longer input sequences nor employing larger LLMs we could not observe the impact of conceptual metaphors on input sequences of all possible lengths with all possible sizes of LLMs.

The broader underlying goal of the work is uncover the blackbox in inferential LLMs while the research regarding them still stays in a nascent stage. Shapley interaction values have proved to be quite a fundamental research tool for post-hoc explanations. Specifically, when it comes to investigating contributions of individual components of input data towards model performances (which is exactly what we are looking for in order to answer our research questions), it was a rather obvious choice in our work. While the environmental impact is a valid concern, we note that that the Shapley analysis is a one-time endeavor and not part of the approach itself, which is why we deem it acceptable. Still, we will continue to consider more environment-friendly post-hoc explanations algorithms with their performances and usability at per with Shapley interaction values.

8 Ethical Statement

We do not perceive any potential negative impacts of our work. Both our evaluators were hired via Upwork[3] and were paid $15 per hour.

[2] https://docs.vllm.ai/en/latest/index.html

[3] https://www.upwork.com/

9 Acknowledgment

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number TRR 318/1 2021 – 438445824. We thank the anonymous reviewers for their helpful feedback.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sebastian Bordt and Ulrike von Luxburg. 2023. From Shapley Values to Generalized Additive Models and back. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 709–745.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. 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.

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.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Alexander Vladislavovitch Dmitrijev, Elena Sergeevna Krupnova, and Anastasia Aleksandrovna Protopopova. 2024. Metaphors and analogies in the context of large language models. In *International Conference on Professional Culture of the Specialist of the Future*, pages 326–341. Springer.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fabian Fumagalli, Maximilian Muschalik, Eyke Hüllermeier, Barbara Hammer, and Julia Herbinger. 2025. Unifying Feature-Based Explanations with Functional ANOVA and Cooperative Game Theory. In *28th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Torda Gábor. 2014. Metaphors in everyday english.

Giuseppe Gallipoli and Luca Cagliero. 2025. It is not a piece of cake for GPT: Explaining textual entailment recognition in the presence of figurative language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9656–9674, Abu Dhabi, UAE. Association for Computational Linguistics.

Jonathan Gordon, Jerry Hobbs, Jonathan May, Michael Mohler, Fabrizio Morbini, Bryan Rink, Marc Tomlinson, and Suzanne Wertheim. 2015. A corpus of rich metaphor annotation. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 56–66, Denver, Colorado. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zoltán Kövecses. 2010. Metaphor, language, and culture. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 26:739–757.

George Lakoff and Mark Johnson. 2003. *Metaphors We Live By*. University of Chicago Press.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.

Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. 2014. Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis)*, 20(12):1983–1992.

Scott M. Lundberg, Gabriel G. Erion, Hugh Chen, Alex J. DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67.

Maximilian Muschalik, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. 2024. shapiq: Shapley interactions for machine learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

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

Meghdut Sengupta, Milad Alshomary, Ingrid Scharlau, and Henning Wachsmuth. 2023. Modeling highlighting of metaphors in multitask contrastive learning paradigms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4636–4659, Singapore. Association for Computational Linguistics.

Meghdut Sengupta, Milad Alshomary, and Henning Wachsmuth. 2022. Back to the roots: Predicting the source domain of metaphors using contrastive learning. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 137–142, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Meghdut Sengupta, Roxanne El Baff, Milad Alshomary, and Henning Wachsmuth. 2024. Analyzing the use of metaphors in news editorials for political framing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3621–3631, Mexico City, Mexico. Association for Computational Linguistics.

L. S. Shapley. 1953. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press.

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

Inna V Skrynnikova. 2024. Interpreting metaphorical language: A challenge to artificial intelligence. *Bulletin of Volgograd State University. Series 2: Linguistics*, 23(5):99–107.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2020. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *Preprint*, arXiv:1904.01172.

Kevin Stowe, Nils Beck, and Iryna Gurevych. 2021. Exploring metaphoric paraphrase generation. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 323–336, Online. Association for Computational Linguistics.

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.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Gemma Team. 2024. Gemma.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online. Association for Computational Linguistics.

A Appendix

A.1 Prompts

This section contains the example prompts for all the experiments

A.1.1 Annotations

Prompts without Metaphor You are a linguistics professor. Your task is to identify the metaphor, source, and target domain of the given metaphorical sentences. In case of multiple possible source and target domains, they are separated by '/'. Please see the following examples, where each metaphorical sentence, its metaphor, and its source and target domains are separated by '#':
"### metaphorical sentence: My soul was a lampless sea and she was the tempest. # source domain: Body of Water # target domain: emotions"
"### metaphorical sentence: I embarked on life thirty years ago. # source domain: movement # target domain: life"
"### metaphorical sentence: The politician clawed his rival. # source domain: physical harm # target domain: human being"
"### metaphorical sentence: 'We've had more paperwork, not less, since the trust took over' Buzz said tartly. # source domain: institution # target domain: bureaucracy"
"### metaphorical sentence: He crushed remorse and sunk down despair. # source domain: water # target domain: emotions/feelings"
"### metaphorical sentence: Berry's songs are plausible emblems of rock'n'roll rebellion or, at any rate, youthful hedonism.# source domain: institution # target domain: art"
"### metaphorical sentence: The critics paid tribute to this broadway production. # source domain: money # target domain: art"
"### metaphorical sentence: I cannot digest all this information. # source domain: food # target domain: information"
"### metaphorical sentence: I wrestled with this decision for years. # source domain: a fight # target domain: life"
"### metaphorical sentence: The seeds of change were planted in 1943. # source domain: plant/crop # target domain: a change"
"What would be the source and target domains of this metaphorical sentence: '{}'?"
"Answer in the same manner as shown in the examples above."

A.1.2 Zero-Shot and Few-Shot Experiments

Prompts with metaphor "You are a linguistic professor. Your task is to identify the metaphor, source and target domain of the given metaphorical sentences. In case of multiple possible source and target domains they are separated by '/'. Please see the following examples of a metaphorical sentence and their metaphor, source and target domains all separated by a '#':
"### metaphorical sentence: My soul was a lampless sea and she was the tempest. # metaphor: lampless sea # source domain: Body of Water # target domain: emotions"
"### metaphorical sentence: I embarked on life thirty years ago. # metaphor: embarked on # source domain: movement # target domain: life"
"### metaphorical sentence: The politician clawed his rival. # metaphor: clawed # source domain: physical harm # target domain: human being"
"### metaphorical sentence: 'We've had more paperwork, not less, since the trust took over' Buzz said tartly. # metaphor: trust # source domain: institution # target domain: bureaucracy"
"### metaphorical sentence: He crushed remorse and sunk down despair. # metaphor: sunk down # source domain: water # target domain: emotions/feelings"
"### metaphorical sentence: Berry's songs are plausible emblems of rock'n'roll rebellion or, at any rate, youthful hedonism. # metaphor: emblems # source domain: institution # target domain: art"
"### metaphorical sentence: The critics paid tribute to this broadway production. # metaphor: paid tribute # source domain: money # target domain: art"
"### metaphorical sentence: I cannot digest all this information. # metaphor: digest # source domain: food # target domain: information"
"### metaphorical sentence: I wrestled with this

decision for years. # metaphor: wrestled # source domain: a fight # target domain: life"
"### metaphorical sentence: The seeds of change were planted in 1943. # metaphor: seeds of change # source domain: plant/crop # target domain: a change"
"What would be the source and target domains of this metaphorical sentence: ''?"
"Answer in the same manner as shown in the examples above."

Prompts with Metaphor and Explanation
"You are a linguistic professor. Your task is to identify the metaphor, source, target domain of the given metaphorical sentences. Additionally you need toprovide the explanation of why you choose the source and target domains accordingly, as shown in the examples below. In case of multiple possible source and target domains they are separated by '/'. Please see the following examples of metaphorical sentences and their metaphors, source domains, target domains, and explanations of why a source domain and target domain are the respective source and target domains in a metaphorical sentence in the given context, all separated by a '#':
metaphorical sentence: My soul was a lampless sea and she was the tempest. # metaphor: lampless sea # source domain: Body of Water # target domain: emotions # explanation: The concept sea pertains to the domain of awater body and hence it's taken from that concept domain. In this sentence the metaphor lampless sea explains the emotions manifested by the speaker by person, who is referred to as the tempest."
"### metaphorical sentence: I embarked on life thirty years ago. # metaphor: embarked # source domain: movement # target domain: life # explanation: Embarking pertains to the start of a journey hence here the concept domain from where the meaning is taken is that of 'movement', while it projects this meaning onto life
"### metaphorical sentence: The politician clawed his rival. # metaphor: clawed # source domain: physical harm # target domain: human being # explanation: clawing includes the involvement of nails which causes physical harm and in this case the sentence describes how the politician verbally attacked his rival, who is a human being"
"### metaphorical sentence: 'We've had more paperwork, not less, since the trust took over' Buzz said tartly. # metaphor: trust # source domain: institution # target domain: bureaucracy # explanation: In this case trust is a constitution of people and the sentence explains the problems of bureaucracy"
"### metaphorical sentence: He crushed remorse and sunk down despair. # metaphor: sunk # source domain: water # target domain: emotions/feelings # explanation: sinking is a concept associated typically with water and hence that is the source domain, while the target domain is emotions or feelings of a human being as it pertains to sadness"
"### metaphorical sentence: Berry's songs are plausible emblems of rock'n'roll rebellion or, at any rate, youthful hedonism. # metaphor: emblems # source domain: institution # target domain: art # explanation: an emblem is a heraldic device or symbolic object as a distinctive badge of a nation, organization, or family and hence the source domain is that of an institution, while the target domain of music, which is an art form is described
metaphorical sentence: The critics paid tribute to this broadway production. # metaphor: paid # source domain: money # target domain: art # explanation: an emblem is a heraldic device or symbolic object as a distinctive badge of a nation, organization, or family and hence the source domain is that of an institution, while the target domain of music, which is an art form is described here
metaphorical sentence: I cannot digest all this information. # metaphor: digest # source domain: food # target domain: information # explanation: Digestion is used as a metaphor from the concept domain of food and in this sentence the difficulty to retain a lot of information is described, making it the target domain
metaphorical sentence: I wrestled" "with this decision for years. # metaphor: wrestled # source domain: a fight # target domain: life # explanation: Wrestling is used metaphorically pertaining to thesource domain of fight, and the sentence explains the struggle to make a decision of life - making life the target domain
"### metaphorical sentence: The seeds of change were planted in 1943. # metaphor: seeds # source domain: plant/crop # target domain: a change # explanation: Seeds of change is taken from the source domain of plants or crops indicating the origin of something new, where the target domain of a change across time is described, incidentally making time the target domain What would be the metaphor, source, target domains, and explanation of this metaphorical sentence:"++"?
Answer in the same manner as shown in the exam-

ples above."

Zero-shot "You are a linguistics professor. Predict if the first sentence - the hypothesis - entails or contradicts the second sentence - the premise.
Furthermore, the source and target domains of the metaphors present in the hypothesis are provided. A source_domain is the concept domain from where the meaning of the metaphors are taken from. A target_domain is the concept domain which is explained by the metaphor.
Additionally, an explanation of why the hypothesis entails or contradicts the premise is also provided. The hypothesis, the premise, the source_domain, the target_domain, and the explanation are presented in this format: hypothesis premise source_domain target_domain explanation.
So the predicted label will just be 'contradiction' or 'entailment'.
Do not provide any explanations and be very precise. Do not include any .
The hypothesis and the premise are:
Answer: "

Few-shot "You are a linguistics professor. Predict if the first sentence - the hypothesis - entails or contradicts the second sentence - the premise. Furthermore, the source and target domains of the metaphors present in the hypothesis are provided. A source_domain is the concept domain from where the meaning of the metaphors are taken from. A target_domain is the concept domain which is explained by the metaphor. Additionally, an explanation of why the hypothesis entails or contradicts the premise is also provided. The hypothesis, the premise, the source_domain, the target_domain, and the explanation are presented in this format: hypothesis premise source_domain target_domain explanation. So the predicted label will just be 'contradiction' or 'entailment'.
Here are a few examples:
#1 The company fired him after many years of service The company released him after many years of service captivity/animal employment The company released him after many years of service means that they no longer needed his services and so they let him go. Answer: entailment
#2 He choked on the martinis and became unconscious He downed three martinis before dinner drinking consuming alcohol The word downed here connotes that someone has consumed or finished something quickly, while choked implies that someone has had difficulty consuming or finishing something. Answer: contradiction
#3 His body was horribly disfigured by the bacteria known as leprosy His body was twisted by leprosy physical distortion health/disease Leprosy is a bacteria that can cause deformities in the body. Answer: entailment
#4 I was really agonizing over this decision for years I wrestled with this decision for years a fight decision-making process Wrestling with a decision means that you are struggling to make a choice. Answer: entailment
#5 FISA was relatively unimportant at the time, organized European and other championships and participating in the running of Olympic regattas Fisa was then a relatively unimportant body which hosted european and other championships and participated in the running of olympic regattas human/organization institution Here the word host mean FISA was the organizer of the programs Answer: entailment
#6 For summer and his pleasures stop on thee For summer and his pleasures take flight on thee bird/flight time/seasons Here one sentence is saying that summer and its pleasures are fleeting and will soon be gone, while the other sentence is saying that summer and its pleasures will last. Answer: contradiction
#7 We laid in the field of green grass and relaxed We laid in fields of gold wealth/value nature/happiness Laying in fields of gold would suggest that someone is relaxing and enjoying themselves in a field of grass. Answer: entailment
#8 We have to be self-reliant and not rely on others We can lean on this man physical support trust/reliability To lean on someone means to rely on them for support, while to be self-reliant means to be able to rely on oneself. Answer: contradiction
#9 Julia was an small pizza with welcoming toppings, and frankly, I was too hungry Julia was an overbearing pizza with condescending toppings, and frankly, I was on a diet food person/relationship An overbearing pizza with condescending toppings would be one that is large and overwhelming at the same time - unappetizing or off-putting, while a small pizza with welcoming topping mean one that is more manageable and less imposing while being inviting and enticing. Answer: contradiction
#10 A bunch of clouds randomly spinning around in the sky The clouds twirled each other around in the sky Dance Weather/Nature The clouds were spinning around in the sky and it looked like they were dancing. Answer: entailment

Do not provide any explanations and be very precise. Do not include any. The hypothesis and the premise are: Answer: "

A.2 Hyperparameter details

We use the following hyperparameters for all of our experiments, with our pipeline implemented with langchain[4]: max_new_tokens=500, top_k=10, top_p=0.95, temperature=0.8

[4]https://python.langchain.com/docs/integrations/llms/vllm/