Using Large Language Models to Perform MIPVU-Inspired Automatic Metaphor Detection

Sebastian Reimann and Tatjana Scheffler
Ruhr University Bochum
Department for German Language and Literature
Bochum, Germany
{sebastian.reimann,tatjana.scheffler}@rub.de

Abstract

Automatic metaphor detection has often been inspired by linguistic procedures for manual metaphor identification. In this work, we test how closely the steps required by the Metaphor Identification Procedure VU Amsterdam (MIPVU) can be translated into prompts for generative Large Language Models (LLMs) and how well three commonly used LLMs are able to perform these steps. We find that while the procedure itself can be modeled with only a few compromises, neither language model is able to match the performance of supervised, fine-tuned methods for metaphor detection. All models failed to sufficiently filter out literal examples, where no contrast between the contextual and a more basic or concrete meaning was present. Both versions of LLaMa however signaled interesting potentials in detecting similarities between literal and metaphoric meanings that may be exploited in further work.

1 Introduction

Metaphors, according to the Conceptual Metaphor Theory (CMT) of Lakoff and Johnson (1980), fundamentally shape the way humans perceive the world. Metaphorical language like considering claims to be *indefensible* is, according to CMT, a way how conceptual mappings in human cognition may be observed on the surface. Several procedures have been develop to systematically identify such linguistic metaphors or metaphor-related words (MRWs) in text, the most famous out of which are the Metaphor Identification Procedure (MIP) by the Pragglejaz Group (2007) and its modified version, the Metaphor Identification Procedure VU Amsterdam (MIPVU) of Steen et al. (2010).

The automatic detection of metaphors has already received considerable interest in NLP, with two shared tasks (Leong et al., 2018, 2020) specifically devoted to it. One line of work in automatic metaphor detection is rooted in modeling aspects of guidelines for manual metaphor identification with the help of (contextualized) language models (Mao et al., 2019; Choi et al., 2021; Babieno et al., 2022; Zhang and Liu, 2022). More recently, efforts were made to integrate the current generation of generative large language models (LLMs) such as GPT (Brown et al., 2020) and LLaMa (Touvron et al., 2023) into annotation processes (Tan et al., 2024). While previous MIP(VU)-motivated approaches to automatic metaphor detection only approximated the instructions of manual metaphor identification methods, prompting LLMs would theoretically allow for the direct instruction of the language models to conduct the steps required by metaphor identification procedures.

Moreover, the analogical reasoning capabilities of LMs have been a recent topic of interest, ranging from verbal, word-level analogies of the type "lawyer:defending :: teacher:educating" (Stevenson et al., 2023) to drawing analogies between high-level messages and more complex narratives (Sourati et al., 2024). Analogy and metaphor are strongly intertwined, since metaphor is often considered a subtype of analogy (Bowdle and Gentner, 2005) and even the most complex type of analogy (Wijesiriwardene et al., 2023). Determining similarities and analogies between two senses is also a key part of MIPVU since, in the MIPVU manual, Steen et al. (2010) explicitly state that two distinct senses are considered metaphor-related if they "capitalize on external or functional resemblances (attributes and relations) between the functions they designate". Consequently, a detailed evaluation of the performance of LLMs on MIPVU, with a particular focus on the role of similarity, would provide further insights into the reasoning capacities of LLMs on complex analogies.

In this paper, we make the following contributions:

1. We present the first effort to adapt MIPVU for

Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II), pages 10–21

its application with generative LLMs, which we achieve with only minor tweaks to the procedure.

2. We evaluate state-of-the-art generative LLM families on two large-scale datasets annotated with MIPVU, where LLaMa performed best but none of the models matched the performance of supervised, fine-tuned approaches.

3. We present an error analysis concentrating on the step where the models made the most mistakes. We find that judging the concreteness and distinctness of two senses is a larger problem than reasoning about similarity, where LLaMa showed interesting potential.

2 Previous Work

2.1 MIPVU

The starting point for our approach is the Metaphor Identification Procedure VU Amsterdam (MIPVU) by Steen et al. (2010), which we will outline in this section. After reading and understanding the entire text (Step 1) and dividing the text into lexical units (Step 2), MIPVU asks the annotator on a word by word basis to **identify the contextual meaning** (Step 3a) of each word. This meaning may or may not be found in a dictionary.

After the contextual meaning has been identified, MIPVU requires the annotator to **identify whether a more basic meaning exists** (Step 3b). Here, Steen et al. (2010) ask the annotator to consult the dictionary entry since they stress that a more basic meaning according to their definition is always conventionalized enough to be found in a dictionary. Moreover, they consider a meaning to be *more basic* if it is more concrete, more specific, or more human-oriented.

If such a basic meaning is found, then MIPVU asks the annotator to **decide if the two meanings are sufficiently distinct** and if they are **related by similarity** (Step 3c). As a shortcut for distinctness, Steen et al. (2010) consider two senses with two different numbered sense descriptions in a dictionary always to be sufficiently distinct. If only one sense description is available or if the contextual meaning is not represented in the dictionary (such as in the case of novel metaphors), this decision is up to the annotator.

Similarity is described by Steen et al. (2010) as "sharing external or functional resemblances (attributes and relations)". They stress that similarity distinguishes metaphor from metonymies like *The White House* for the US government, which expresses a part-whole relationship. If the two senses are similar, then the word should be marked as an MRW. Additionally, MIPVU leaves room for special cases, like MRWs that are part of direct comparisons (*[...]like an eagle*; "direct MRWs"), the replacement of MRWs by pronouns, personifications and borderline cases.

2.2 Automatic Metaphor Identification inspired by MIP(VU)

Mao et al. (2019) aimed to model MIP and Selectional Preference Violations (SPV) by Wilks (1975), which detect metaphors through clashes with their context. The MIP model uses GloVe and ELMo embeddings and a BiLSTM, whose hidden states represent the contextual meaning. The basic meaning is represented by the GloVe embedding only and a concatenation of contextual and basic meaning representation serves as input to a classifier. SPV is modeled with the same architecture and a concatenation of left- and right-context representations. Both models achieved F1-scores of around 74 points on the VUA18 dataset (Leong et al., 2018) and outperformed other metaphor identification approaches at the time.

Inspired by Mao et al. (2019), Choi et al. (2021) developed MelBERT, which aims to model MIP and SPV with the help of contextual BERT embeddings. It uses two encoders, one for the entire sentence, and one for the word in isolation. MelBERT models MIP through a concatenation of the contextual embedding and the embedding of the word in isolation. SPV, on the other hand, is imitated through a concatenation of the sentence embedding and the contextual word embedding. The output of the SPV and MIP layers is concatenated and fed into a linear classifier. Evaluated on VUA18, it outperformed Mao et al. (2019) with an F1-score of 78.5. On VUA20 (Leong et al., 2020), it achieved an F1-score of 73.9.

Several other authors presented further improvements of MelBERT, mainly concentrating on the representation of the basic meaning. MIss-RoBERTaWiLDE (Babieno et al., 2022) uses a Wiktionary entry to represent the basic meaning, MisNet (Zhang and Liu, 2022) encodes an example sentence from the first sense in the dictionary entry of the target word and uses its contextual word embedding as basic meaning. Li et al. (2023) search the training set for non-metaphoric literal examples,

encode them and average the obtained embeddings to represent the basic meaning.

2.3 LLMs and Metaphor

One early contribution to evaluate the metaphor understanding of generative LLMs was made by Wachowiak and Gromann (2023). They asked GPT-3 to provide the source for examples from the Master Metaphor List[1] and the English and Spanish sections of the LCC Corpus (Mohler et al., 2016) with prompts that include the target domain of the metaphor and completed source-target mappings. They also included non-metaphoric examples from VUA to test if the model is able to distinguish metaphoric and non-metaphoric examples. For the simple sentences from the Master Metaphor List, the model predicted the correct source domain with an accuracy of 81.33%, which however went down drastically for the more complex LCC examples. Moreover, the non-metaphoric examples were only singled out with an accuracy of 42.11%.

Schuster and Markert (2023) investigated cross-lingual detection of metaphoric adjective-noun pairs. They compared the zero-shot cross-lingual transfer performance of various BERT and fastText-based classifiers with the performance of ChatGPT when given various prompts. Here, ChatGPT performed best when including the MIP guidelines in the prompt. However, already with little target language data, smaller models were outperforming ChatGPT.

Chen et al. (2024) extended automatic detection with an additional reasoning task, where the models are asked to justify why they considered a token literal or metaphoric. They prompted two variants of LLaMA3 (8B and 70B), Gemma-7B, and ChatGPT3.5 on detecting metaphoric tokens and reasoning. The metaphor detection performance mostly falls short in comparison to supervised approaches, particularly on VUA, where they only achieve F1-scores between 27 and 44 points on the binary metaphor detection.

TSI by Tian et al. (2024) represents the most elaborate approach to metaphor identification to date. They prompted GPT-3.5 with a series of questions inspired by CMT, MIP and SPV and filled knowledge graphs with the answers. If the graph surpasses a final comparison with the ideal knowledge graph pattern, the example is labeled as metaphoric. Their CMT approach performed best and even outperformed several fine-tuned BERT approaches with an F1 of 82.59 on the MOH-X dataset and 66.07 on the TroFi dataset, with MIP ranking second with F1-scores of 79.39 and 65.60, respectively.

The approach of Boisson et al. (2025) also extracted entire source-target mappings with generative LLMs. They represented metaphors as mappings involving two concept terms from each a source and a target domain and aimed to extract such mappings from a collection of novels, poems, songs and speeches. For this, they provided LLMs (GPT-4, Llama-3 and Mixtral) with prompts containing a text and one of the four terms representing the different concepts. They reported mostly satisfactory accuracies, often over 60%.

3 Experiments

3.1 Adapting MIPVU for LLMs

In the previous section, we have already seen two approaches that integrated MIP(VU) into LLM prompts and achieved satisfactory performance on a small amount of data. However, in both Schuster and Markert (2023) and Tian et al. (2024) the metaphor identification procedure was implemented in a simple and rather counterintuitive way. Schuster and Markert (2023) only provided the model with the respective adjective-noun-pair without context, and Tian et al. (2024) left out similarity as a criterion.

For our application of MIPVU, we aim to replicate the steps outlined in Section 2.1 as closely as possible. We, however, do not take extra steps for any of the aforementioned special cases. We already deal with tokenized text, thus no prompts for the steps 1 and 2 are needed. Step 3a, identifying the contextual meaning, can be achieved in a straightforward manner by simply providing the LLM with the word in question and the text containing the word and prompting it to provide the meaning of the word in the given context.

Step 3b, the identification of a more basic meaning, on the other hand, is more complex. It is first necessary to decide on a resource of possible senses. For this, we ask the model to provide us with an entire dictionary entry for the word in question. After extracting the senses, we additionally prompt the model to identify if any of the present senses can be considered more concrete than the contextual meaning.

Here, we needed to make compromises. MIPVU

[1] https://www.lang.osaka-u.ac.jp/sugimoto/MasterMetaphorList/metaphors/index.html

technically asks for *a more basic meaning* instead of *the most basic meaning*. Thus, in theory, we would need to conduct the next steps for all potential meanings that fulfill the "more basic" criterion, which would be very resource-demanding. Consequently, we ask the model to only extract the most concrete candidate and proceed with this as the more basic meaning. If no basic meaning is available, the word is considered non-metaphoric.

If the model identified a more basic meaning, then the two meanings need to be checked for sufficient distinctness (first part of step 3c). We simply ask the model whether the contextual and more basic meanings refer to the same concept. If yes, the example is considered non-metaphoric.

Otherwise, we proceed with the crucial question on similarity between the two senses (second part of step 3c). For this, we ask the model if the two senses in question share aspects, functions or features (the criteria for "similarity" as outlined by Steen et al. (2010)) and for a short explanation. We test a zero-shot prompt that only asks for similarity according to Steen et al. (2010) as well as a one-shot prompt that illustrates similarity according to Steen et al. (2010) with examples 1 and 2, the two senses for *journey* provided in the Longman Dictionary of Contemporary English (Longman, 2023).

(1) an occasion when you travel from one place to another, especially over a long distance

(2) a long and often difficult process by which someone or something changes and develops

However, given that word-sense disambiguation (WSD) represents an NLP task which non-generative language models perform well (Bevilacqua et al., 2021), we also test the combination of a BERT-based WSD model and a generative LLMs. For this, we used the fine-tuned model presented in Yap et al. (2020). Here, for step 3b of MIPVU, we thus extract the sense keys and their respective glosses from WordNet (Miller, 1994) instead of generating an entire dictionary entry and then ask the LLM to provide us with the more basic meaning among the extracted senses.

The fine-tuned WSD model comes into play when checking for sufficient distinctness. We have the WSD model predict the sense. If this predicted sense by the WSD model and the predicted more basic meaning by the LLM are different, then they are considered sufficiently distinct. The final question for similarity is then asked in the same way as in the procedure without WSD. An overview over all prompts that we used is provided in Appendix B.

3.2 Models

We evaluate three commonly used families of LLMs in our experiments: LLaMa, Mistral and GPT. For LLaMa, we specifically use the 8B and 70B instruction-tuned versions of LLaMa 3.1. For Mistral, we use the also instruction-tuned Mistral-Small-Instruct-2409 with 22 billion parameters. We obtain the LLaMa and Mistral models via HuggingFace (Wolf et al., 2020). For GPT, given financial constraints, we only use the lightweight GPT-4o-mini, which we access via the OpenAI API. We used the default hyperparameters of all the models. We ran the 8B version of LLaMa 3.1 on NVIDIA A40 GPUs, the 70B version of LLaMa 3.1 on NVIDIA H100 SXM5 GPUs and the Mistral model was run on NVIDIA A30 GPUs.

3.3 Data

For evaluation purposes, we use two larger metaphor datasets where the annotation followed MIPVU very closely. On the one hand, we use the VUA dataset in the version that was used in the 2020 Metaphor Detection Shared Task (Leong et al., 2020) and which is based on the original application of MIPVU to the British National Corpus by Steen et al. (2010).

Moreover, we use the metaphor dataset of Reimann and Scheffler (2024) ("R&S" in the following), which consists of posts from Christian subreddits annotated for metaphor via MIPVU. Due to the limits of the OpenAI API and financial considerations, we only use a fraction of the test data, namely two reddit threads from R&S and one fragment from VUA in the experiments involving GPT-4o-mini. Table 1 presents an overview of the data that we used.

Dataset	Tokens	MRWs
VUA	22196	3982
VUA (short)	3960	821
R&S	14437	3170
R&S (short)	3562	555

Table 1: Overview of the data.

Model	Setup	R & S				VUA			
		P	R	F1	Acc.	P	R	F1	Acc.
LLaMa 3.1 8B	0-Shot	25	72	38	47	21	59	31	50
	1-Shot	25	79	38	44	21	64	32	53
	0-Shot + WSD	*32*	47	38	68	*32*	63	*43*	64
	1-Shot + WSD	*32*	63	*43*	64	28	59	38	65
LLaMa 3.1 70B	0-Shot	26	82	39	44	21	69	33	49
	1-Shot	26	*84*	40	43	22	*70*	33	50
	0-Shot + WSD	27	63	38	56	26	68	38	60
	1-Shot + WSD	28	67	39	56	26	68	38	60
GPT-4o-mini	Zero	20	19	20	76	19	22	21	60
	One	20	11	14	**79**	22	18	20	67
	0-Shot + WSD	18	11	13	*78*	27	10	15	**74**
	1-Shot + WSD	17	8	11	**79**	29	13	18	*73*
Mistral-Small	Zero	22	**96**	36	29	19	**95**	32	27
	One	20	72	32	35	18	74	29	35
	0-Shot + WSD	31	54	40	65	29	55	38	68
	1-Shot + WSD	29	45	35	65	28	49	36	69
MelBERT (Choi et al., 2021)		**76**	69	**72**	-	**68**	60	**64**	-

Table 2: Precision, recall and F1 for the metaphor class and accuracy on the two datasets for all LLM setupts as well as the results of the supervised MelBERT approach reported in Choi et al. (2021) and Reimann and Scheffler (2024) for comparison. Best result for each metric in bold, second best in italics.

4 Results and Error Analysis

Table 2 shows our results. Overall, we can see that neither of the LLM is actually able to achieve satisfactory performance in any setting, with GPT-4o-mini in particular trailing behind the other two models in all metrics except for accuracy. The data is imbalanced (around 80% of tokens non-metaphorical), which means that if the model considers fewer tokens to be MRWs, then will automatically be higher.

A general trend for LLaMa is that model size had a positive impact on recall. The same can be said for providing LLaMa with an example in the last prompt since this resulted in increased recall for almost all setup scenarios. GPT-4o-mini, on the other hand, was not at all able to profit from the inclusion of an example and for the Mistral model, the inclusion of an example in the prompt even had the opposite effect and led to drastic drops in performance.

4.1 False Positives

One more particularly striking pattern in Table 2 is the fact that both LLaMa and especially Mistral produce a large amount of false positives, i.e. non-metaphoric examples that were considered to be MRWs, as seen in the low values for precision. The way MIPVU functions may provide an explanation: the initial steps of MIPVU are designed to weed out cases where no meaning can be considered "more basic" according to the procedure or where the contextual and the more basic meaning are too similar. This ensures that in the last step, only two distinct meanings are compared. However, when the model fails to complete the previous steps correctly, it will eventually compare two meanings that refer to the same concept to each other and, as a logical choice, understandably consider them to be similar.

The example of *conversation* in Table 3 illustrates such cases, which represent the most common type of false positive. Here, both the explanation of the contextual meaning and the "more basic" meaning refer to the same concept; however, the model did not judge the meanings to be distinct. They were then inevitably judged as related by similarity and received the label *MRW*.

However, some cases of false positives were also similar to the example of *night* in Table 3. Here, the model already misinterpreted the contextual meaning of *night* (which was used literally as the time of day in the original sentence) as nonliteral. Based on this propagated error, the meanings were then judged as sufficiently distinct and, consequently, as *MRW*. The wider context for both examples in Table 3 is provided in Appendix A.

Word	Contextual	More Basic	Similarity?
conversation	In the context of this post, the word "conversation" refers to a meaningful and impactful exchange of thoughts, ideas, and emotions between two individuals, often leading to a deeper understanding, personal growth, and a sense of connection or revelation.	a talk between two or more people, especially one that is friendly and informal	The similarity between the two senses is that they both refer to a talk or exchange between two or more people. Sense 1 emphasizes the depth and impact of the conversation, while Sense 2 provides a more general definition that focuses on the formality and tone of the conversation.
night	In this context, the word "night" refers to a period of time, specifically a dark and potentially difficult or vulnerable moment, often symbolizing a state of need or crisis.	The period of time from sunset to sunrise, when it is dark outside.	The similarity between the two senses of 'night' is that they both refer to a period of time, specifically a time when it is dark outside. Sense 1 describes 'night' as a moment of need or crisis, while Sense 2 defines it as the time from sunset to sunrise.

Table 3: Examples for false positives, produced by the 8B version of LLaMa.

Replacing the prompt-based judgment on sufficient distinctness with BERT-based word sense disambiguation overall resulted in a stricter application of the *MRW* label. However, for all models, the improvements for precision are much smaller than the drops in recall. This may be because the glosses for WordNet senses may sometimes be not informative enough or two glosses may appear too similar for the model.

(3) water falling in drops from vapor condensed in the atmosphere

(4) drops of fresh water that fall as precipitation from clouds

The examples 3 and 4 for *rain* illustrate this. They are the glosses for two different senses in WordNet, however, it may be argued that they denote the same concept. The LLaMa models selected 4 as the more basic meaning and 3 was selected by the WSD model to be the contextual meaning. This led the model to not dismiss the example as metaphoric and in the last step, the meanings were considered similar and thus wrongly labeled as *MRW*.

4.2 False Negatives

This implementation of MIPVU with generative LLMs gives us, in the case of false negatives, the opportunity to track exactly where the decisive error was made. We make use of this to better interpret the results of Table 2. The results of this analysis are provided in Table 4.

Consistently, for both instances of LLaMa, deciding on distinctness appears to be the biggest problem, followed by deciding on similarity. In contrast, the small GPT model already produces the most false positives when prompted to decide on a more basic meaning. Surprisingly, the model notably produced more false negatives in the distinctness step when evaluated on VUA, compared to the evaluation on R & S. Looking into the most frequent false negatives for VUA, we find a wide range of heavily conventionalized MRWs such as *make* among the most common false negatives. For these examples, the contextual and more basic meaning may appear too similar, which possibly explains why the models considered them to be not sufficiently distinct.

Mistral without the added WSD model on the other hand in general produced not many false negatives, which further highlights that it was not sufficiently strict to carry out MIPVU. A striking result, however, is that when provided with the WordNet glosses, the model wrongly sorted many MRWs out when checking for a more basic meaning. This may again be because the WordNet glosses were shorter and less informative than the generated definitions.

Table 4 also further exemplifies the improvement in recall for LLaMa. Here, we can see that for all LLaMa models, the number of wrong classifications as non-metaphorical in the last step drops notably in the one shot scenario. This suggests that LLaMa indeed was able to better reason on metaphoric similarity when provided with an example.

This impression is also confirmed by looking at the first two examples in Table 5, which compares the output of several models for the zero- and few-shot prompts for the last MIPVU step. The example of *sheep*, by the 70B version of LLaMa is particularly interesting as the model, when asked to describe the contextual meaning of the word, already mentions that it is used metaphorically. However,

Model		R & S				VUA			
		Basic	Dist.	Sim. (0-shot)	Sim. (1-shot)	Basic	Dist.	Sim. (0-shot)	Sim. (1-shot)
LLaMa-3.1-8B-Instruct	w/o WSD	193	366	250	53	170	1202	277	45
	w/ WSD	153	943	551	31	139	1203	574	45
LLaMa-3.1-70B-Instruct	w/o WSD	85	304	163	86	372	743	130	99
	w/ WSD	60	766	264	177	97	769	172	176
GPT-4o-mini	w/o WSD	345	83	19	66	526	92	5	55
	w/ WSD	417	64	9	24	549	106	3	27
Mistral-Small-Instruct-2409	w/o WSD	53	71	6	769	106	105	6	805
	w/ WSD	1405	0	8	296	1568	0	217	6

Table 4: Number of MRWs wrongly considered to be non-metaphorical across different steps.

when asked to reason about similarities between the contextual and basic usage of *sheep* in a zero-shot manner, it denies the question. A relation by similarity is often considered a defining feature of metaphor (Steen et al., 2010), which makes the model output thus contradictory. The model given the one-shot similarity prompt, on the other hand, answers with *yes* and provides extensive reasoning on the similarities between the metaphorical *sheep* in the sense of believers and the animal. The metaphorical example of *light* in Table 5 also illustrates how the few-shot prompt had the completely opposite effect for Mistral as it reasoned correctly without the example senses. However, when provided with an example, it still understood that one sense is abstract and one physical but failed to point out the similarity. The context for all examples discussed in Table 5 is provided in Appendix A.

4.3 Deliberate and Direct Metaphor

Finally, we look at different subtypes of metaphors and see if they were harder to detect for the models. For this, we use the more detailed annotations of Reimann and Scheffler (2024), which also indicate if an MRW is direct or indirect and if it is used in a potentially deliberate way (i.e. used "as metaphor") according to the Deliberate Metaphor Identification Procedure (DMIP) by Reijnierse et al. (2018). In addition to direct metaphor, this also includes novel metaphor and extended metaphor. Table 6 shows the recall for all potentially deliberate metaphors and direct metaphors in the most successful scenario with respect to the overall recall (no external WSD model and a one-shot prompt for similarity).

Interestingly, finding potentially deliberate metaphors did not pose particular problems for the LLaMa and Mistral models, which contrasts with the results of supervised BERT-based approaches, that struggled with several types of potentially deliberate metaphors. This also extends to direct metaphor in particular, which is surprising at first glance, given that our prompts do not cover direct metaphor, as outlined in Section 2.1.

(5) I just have to learn to wait on the lord as I feel like a *car* that was *Idling* for so long that my *engine* is just *ready* to *go* , and I need to be calm and wait on the lord .

(6) In this context, a "car" is used metaphorically to describe a person's eagerness and readiness to take action, likening their built-up energy and anticipation to a car's engine idling and ready to move.

The direct MRWs in example 5 are in bold. When asked for the contextual meaning of the word *car*, the large LLaMa model provided the explanation given in Example 6. Here, although not explicitly asked, the model already generates an explanation for the direct metaphor. This makes the contrast between contextual and basic meaning in further comparisons apparent and demonstrates why direct MRWs do not seem to pose problems for this procedure.

5 Discussion

Putting our results into the wider context of automatic metaphor detection, we can, on the one hand, see that they matched or outperformed for the most part the approach of Chen et al. (2024), which also

MRW	Contextual	More Basic	Expl. 0-Shot	Expl. 1-Shot
deeper	In the context of the given post, "deeper" refers to a more profound or underlying level of understanding, issue, or problem, often requiring a more nuanced or complex analysis.	very wide and not very shallow	There is no similarity between the two senses. The first sense refers to a more in-depth or thorough exploration or understanding, often implying a greater level of insight or knowledge. The second sense refers to the physical characteristics of a river or lake, specifically its width and depth.	Sense 1 and Sense 2 are similar because in both senses, the word "deeper" refers to a greater extent or dimension, whether it's a more in-depth understanding or a wider and less shallow body of water.
sheep	In this context, "sheep" is a metaphorical representation of a lost or misguided individual who has strayed from their spiritual path and is being called back to a higher power or sense of self-worth.	a farm animal that produces wool and is often kept for its meat	Sense 1 is a metaphorical representation, while Sense 2 refers to the literal animal. There is no similarity in their meanings, as one is a figurative concept and the other is a physical entity.	Sense 1 and Sense 2 are similar because in both senses, the term 'sheep' refers to an entity that can be guided or directed, whether it's a farm animal being herded or a metaphorical individual being called back to a spiritual path.
light	"Light" in this context refers to one's positive influence, joy, or faith that can be shared with others regardless of circumstances.	the natural brightness that comes from the sun, moon, or stars	Yes, both senses of "light" involve the idea of illumination and positivity. In Sense 1, it's about illuminating someone's spirit or life with positivity, while in Sense 2, it's about illuminating the physical environment with natural brightness.	No. Sense 1 refers to an abstract, emotional concept, while Sense 2 refers to a physical phenomenon. They do not share any aspects, functions, or features.

Table 5: Examples wrongly considered to be non-metaphorical by the 8B (first), 70B (second) versions of LLaMa and Mistral (third) when prompted in a zero-shot fasion and considered to be MRWs when given an example in the prompt on similarity.

Model	R (Pot. Delib.)	R (Direct)
LlaMa 3.1 8B	78	72
LlaMa 3.1 70B	87	88
GPT-4o-mini	10	13
Mistrall-Small	67	68

Table 6: Recall for potentially deliberate and direct metaphors when using the models in the scenario involving a few-shot prompt and without WSD.

used generative LLMs and which evaluated all of the VUA data. The approaches of Schuster and Markert (2023) and Tian et al. (2024) are much harder to compare with ours, as they used much smaller and more balanced data sets.

However, it also becomes clear that our LLMs emulating MIPVU massively underperform previous approaches with fine-tuned variants of BERT. Our results demonstrate that, as of now, identifying metaphors with a series of prompts inspired by the steps of MIPVU is not a realistic alternative for automatic metaphor detection. In order to correctly identify metaphor, the last step of MIPVU is heavily dependent on the correct implication of the previous two steps and propagated errors may result in heavy overuse of the *MRW* label.

Moreover, the models from the three families performed wildly differently. This makes it harder to draw final conclusions about the suitability of LLMs for this task. GPT-4o-mini overall failed, with neither recall nor precision achieving satisfactory results, and Mistral was unable to sufficiently filter out negative examples. However, the results of LLaMa, while far from perfect, show encouraging tendencies.

Neither LLaMa version was able to filter out negative examples in a satisfactory way due to the aforementioned reasons, even though the picture here is more nuanced in comparison to Mistral. However, the improvements for both versions when provided with an example and a qualitative inspection of the output suggested some capacities to reason on similarity and analogy in order to correctly answer the last step of MIPVU. Paired with better identification of the basic meaning and better filtering of negative examples due to insufficient distinctness, this may present potential to even-

tually be able to conduct MIPVU via generative LLMs.

6 Conclusion and Future Work

We replicated MIPVU with LLMs as closely as possible in two different ways. On the one hand, with a series of prompts and only using the respective LLM and, on the other hand, via a combination with WordNet as an external resource. We evaluated four LLMs from three popular model families on two large, realistic datasets that were manually annotated via MIPVU and conducted an extensive evaluation and error analysis of the output.

We found that our approach achieves competitive performance with other LLM-based approaches on the VUA data with LLaMa. However, it still falls short in comparison to models that were specifically fine-tuned on the metaphor detection task in either setup. In general, the models also differed in their behavior, making interpretation and final conclusions difficult. However, one point that all models had in common were struggles to select a basic sense and to decide on distinctness of word senses. In contrast to the others, the LLaMa models performed satisfactory on the last question of analogical similarity between literal and contextual word senses, especially when provided with an example mapping.

For future work, we suggest further investigation into the steps that were problematic for the models, namely perceiving concreteness and distinctness of word senses. For this, and for a better evaluation of the suitability of LLMs for MIPVU in general, we would need more gold labels and human judgements for all the steps of MIPVU instead of only the final labels. Our prompt and MIP(VU) in general is also relatively vague on its definition of *more concrete*. Thus, extending the concreteness prompt with a more comprehensive definition of concreteness or concreteness ratings such as Brysbaert et al. (2014) would be worth trying out. Moreover, as the one-shot prompt already led to some improvements, we suggest providing more examples and testing if different examples for the few-shot experiment would lead to different results.

Finally, we suggest exploring the analogical reasoning capabilities of LLaMa further in the context of automatic metaphor identification: In this work, we used a relatively simple prompt that asks for *similarity* according to the definition of Steen et al. (2010) only. However, the output for this question could be evaluated in an even more systematic manner by, for example, explicitly providing the model with source and target domain terms or, for the last MIPVU question, additionally asking it to identify source and target terms like Boisson et al. (2025) did.

7 Acknowledgements

We thank the Paderborn Center for Parallel Computing (PC2) and HPC@RUB for granting us compute time on their HPC clusters and we thank the anonymous reviewers for their valuable comments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1475 – Project ID 441126958

References

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. Miss roberta wilde: Metaphor identification using masked language model with wiktionary lexical definitions. *Applied Sciences*, 12(4).

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent Trends in Word Sense Disambiguation: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization.

Joanne Boisson, Zara Siddique, Hsuvas Borkakoty, Dimosthenis Antypas, Luis Espinosa Anke, and Jose Camacho-Collados. 2025. Automatic extraction of metaphoric analogies from literary texts: Task formulation, dataset construction, and evaluation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6692–6704, Abu Dhabi, UAE. Association for Computational Linguistics.

Brian F Bowdle and Dedre Gentner. 2005. The career of metaphor. *Psychological review*, 112(1):193.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand

generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Puli Chen, Cheng Yang, and Qingbao Huang. 2024. Merely judging metaphor is not enough: Research on reasonable metaphor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5850–5860, Miami, Florida, USA. Association for Computational Linguistics.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. Univ. of Chicago Press, Chicago [u.a.].

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

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.

Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.

Longman. 2023. *Longman Dictionary of Contemporary English (Online Edition)*. Pearson Education Limited.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Pragglejaz Group. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2018. DMIP: A method for identifying potentially deliberate metaphor in language use. *Corpus Pragmatics*, 2(2):129–147.

Sebastian Reimann and Tatjana Scheffler. 2024. Metaphors in online religious communication: A detailed dataset and cross-genre metaphor detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11236–11246, Torino, Italia. ELRA and ICCL.

Jakob Schuster and Katja Markert. 2023. Nut-cracking sledgehammers: Prioritizing target language data over bigger language models for cross-lingual metaphor detection. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 98–106, Gothenburg, Sweden. Association for Computational Linguistics.

Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. 2024. ARN: Analogical reasoning on narratives. *Transactions of the Association for Computational Linguistics*, 12:1063–1086.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Claire E. Stevenson, Mathilde ter Veen, Rochelle Choenni, Han L. J. van der Maas, and Ekaterina Shutova. 2023. Do large language models solve verbal analogies like children do? *Preprint*, arXiv:2310.20384.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Yuan Tian, Nan Xu, and Wenji Mao. 2024. A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning. 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 7738–7755, Mexico City, Mexico. 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.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549, Toronto, Canada. Association for Computational Linguistics.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1):53–74.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. Adapting BERT for word sense disambiguation with gloss selection objective and example sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 41–46, Online. Association for Computational Linguistics.

Shenglong Zhang and Ying Liu. 2022. Metaphor detection via linguistics enhanced Siamese network. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Posts

Table 7 provides the post containing the examples presented in Tables 3 and 5.

B Used Prompts

We provide the LLM prompts in Table 8.

I recently spent a night of " homelessness " the details of which do n't matter. All I know is walking barefoot , alone , in the **night** , in need of help , during christmas. The only person that found me in the middle of the night or would help , was a homeless man. He gave me his shoes , and a sweater , he walked with me for miles , and talked with me. He gave me aid and comfort , and told me truth about my life that he could n't of known. I thought I was insane , I was n't . (because the things he told me he could n't know) I have more peace now than I ever had in my life and I just need to figure out how to help. I also know I am a Christian and can never deny Christ in my life , I know every blessing I have is not mine , its simply for me to use to help others. I went back to church , and told my tale to the pastor , and that very day I went out to help again and was opposed. Opposed in just the right way to make me fall back into my depression and hide away from the world. The only reason I did n't was a christian brother was there with me and saw everything , he prayed and told me exactly what I needed to hear. because what the Homeless person screamed at me , was exactly every doubt I had about going and telling the pastor my tale , she hit every insecurity I had about the **conversation**. My friend simply said at the time " thats how you know your doing the right thing" So , I have discovered I have a heart for the homeless.
So from there , my standard for evaluating anything became , " Does this facilitate wholeness in someone 's life? " That 's the lens I use to view LGBT issues , and after breaking away from some of the Evangelical propoganda , I 've realized that whole , healthy homosexual relationships are 100 % possible and not at all uncommon. I believe there are also examples where transitioning is the healthiest choice for someone , but I feel like there is a ton of **deeper** rooted identity dysfunction present within the ideologies accompanying that " community / movement / not - sure - the - right - term " , but I have zero interest in dictating anyone 's behavior or telling anyone how they should or should n't live their lives. I feel like our purpose as people is just to love ourselves and everyone else , live whole , healthy , happy lives , and help those around us.
Sadly marriage wo n't solve your issues with self - worth which is the root of your hook up lifestyle. You 're looking for value in relationships when Jesus has already paid the highest price for you. He loves you and wants you to return to Him. You 're a daughter in His eyes. Go read Luke 15. You 're the lost **sheep** / coin / son. God wants you to come back to Him. Your created value comes from Him alone , not what others think of you. The guys you 're hooking up with you just want what you can give them and see no value in you. To them you 're a means to an end , which is why you may find value in the moment but regret it after. It 's a vicious cycle. At least you recognize you ca n't have one or the other.
But thanks be to His Word , first spoken , then written , and then affirmed , confirmed , and fulfilled by The Living Word , one can truly state that they know " of " our Father Love The Adam and Eve , the only 2 flesh humans that had a clear recollection of their true celestial origin / heritage in The **Light**

Table 7: Posts containing the examples discussed in Tables 3 and 5 with the respective MRWs in bold.

Step	Prompt
get the contextual meaning	In one sentence, describe the meaning of the given word in the context of the given post as general as possible. Word: [WORD], Post: [POST]
generate the dictionary entry	Write a dictionary entry in the style of the Longman Dictionary of Contemporary English that provides all possible senses of the given word with the given Part of Speech. Word: [WORD], Part-of-Speech: [POS]
decide on basic meaning	Decide if any of the dictionary senses can be considered more concrete than the example definition. If yes, output 'yes' and in a new line provide only the respective sense. If no, then just provide 'no'. Dictionary Senses: [SENSES]
decide on sufficient distinctness	Do the two senses express the same meaning or is sense 2 only a more specific version of sense 1? Answer with 'yes' or 'no' followed by a brief explanation. Sense 1: [CONT. SENSE], Sense 2: [MORE BASIC SENSE]
decide on similarity (0-Shot)	Can you see a similarity between the senses 1 and 2? 'Similarity' means that the two senses denote distinct concepts that share certain aspects, functions or features. Answer with 'yes' or 'no' followed by a brief explanation. Sense 1: [CONT. SENSE], Sense 2: [MORE BASIC SENSE]
decide on similarity (1-Shot)	Can you see a similarity between the senses 1 and 2? 'Similarity' may also mean that the two senses denote distinct concepts that share certain aspects, functions or features. The following example for the word 'journey' illustrates this: journey: Sense 1: "an occasion when you travel from one place to another, especially over a long distance" Sense 2: "a long and often difficult process by which someone or something changes and develops" Answer: Yes. Sense 1 and Sense 2 are similar because in both senses refer to something that takes a longer period of time. Answer with 'yes' or 'no' followed by a brief explanation. Sense 1: [CONT. SENSE], Sense 2: [MORE BASIC SENSE]

Table 8: Overview over the used prompts for each step.