Unveiling LLMs' Metaphorical Understanding: Exploring Conceptual Irrelevance, Context Leveraging and Syntactic Influence

Anonymous ACL submission

Abstract

Metaphor analysis is a complex linguistic phenomenon shaped by context and external factors. While Large Language Models (LLMs) demonstrate advanced capabilities in knowledge integration, contextual reasoning, and creative generation, their mechanisms for metaphor comprehension remain insufficiently explored. This study examines LLMs' metaphor-processing abilities from three perspectives: (1) Concept Mapping: using embedding space projections to evaluate how LLMs map concepts in target domains (e.g., misinterpreting "fall in love" as "drop down from love"); (2) Metaphor-Literal Repository: analyzing metaphorical words and their literal counterparts to identify inherent metaphorical knowledge; and (3) Syntactic Sensitivity: assessing how metaphorical syntactic structures influence LLMs' performance. Our findings reveal that LLMs generate 15%-25% conceptually irrelevant interpretations, depend on metaphorical indicators in training data rather than contextual cues, and are more sensitive to syntactic irregularities than to structural comprehension. These insights underline the limitations of LLMs in metaphor analysis and call for more robust computational approaches.

1 Introduction

Metaphor is a pervasive and sophisticated aspect of human language (Gibbs Jr, 2008). However, defining and analyzing metaphors poses significant challenges in both computational and traditional linguistics. Several linguistic frameworks have been proposed to approach metaphor analysis: Selection Preference Violation (SPV) and the Metaphor Identification Procedure (MIP), focus on detecting semantic deviations from conventional linguistic patterns (Wilks, 1975; Group, 2007). However, Conceptual Metaphor Theory (CMT) views metaphors as cross-domain mappings between a source domain, which describes tangible objects, and a target domain, which represents abstract ideas (Lakoff and Johnson, 1980; Croft, 1993). These domains indicate conceptual frameworks behind specific words that can also be implied in other expressions. Thus, **a pair of source and target domains can encompass multiple metaphorical mappings**: The mapping "thinking is linguistic activity" can be the generic concept of "memorization is writing" or "ideas are words" (Do Dinh et al., 2018). This uncertainty of mapping results in difficulty translating linguistic theories into computational procedures.

With impressive capabilities in text comprehension and vast knowledge bases (Yang et al., 2024b), LLMs have been widely applied to metaphor computation on cultural contexts (Ichien et al., 2024), cross-lingual interpretation and detection (Shao et al., 2024) and genre-specific studies (Toker et al., 2024). However, recent studies highlight limitations in LLMs' metaphor comprehension. Wachowiak and Gromann (2023) have discovered "**trigger word**" error that hinders sufficient context analysis: LLMs predict wrong source domains that were not metaphorically related, because models only infer from the trigger word instead of considering context. (e.g., "The arms race." is falsely identified as COMPETITION IS WAR, due to the commonly co-occurrence of "arm" and "war" in daily discourse). Also, the robustness of LLMs in multi-choice prompting of proper metaphor interpretation remains under debate–studies imply that this setting can be less effective due to LLMs' hallucination (Li et al., 2024; Zhao et al., 2021).

It is crucial to investigate LLMs' inherent metaphor understanding for future enhancement. This study primarily focuses on three perspectives of LLMs in metaphorical analysis. 1) We explore the **concept-irrelevant error**: the interpretation generated by LLM failed to clarify the actual target domain that source domain intended to map (e.g., interpreting *falling in love* as *physically falling* rather than *entering an emotional state*). 2) Extend from the "**trigger word**" scenario, we inves-

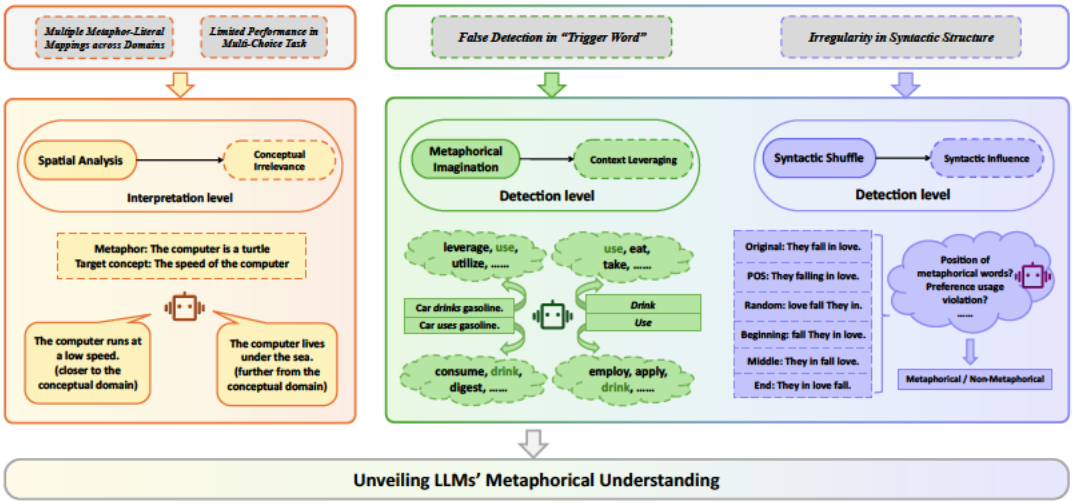

Figure 1: The overview of experiment framework. Spatial analysis addresses the limitations of multi-choice task and multiple metaphorical mapping problem. For "trigger word" error, metaphorical imagination investigate context leveraging and the existence of a metaphor-literal repository within LLMs. Syntactic shuffle identifies the influence of syntax in metaphor analysis.

tigate the inherent metaphor-literal repository of LLMs that indicates limited **leveraging of context** in metaphor analysis and 3) examine how **syntactic structures** influence metaphor detection. Our contributions are as follows:

- We are the first to investigate concept-irrelevant error in metaphor interpretation from high-dimensional space projections derived through dimensionality reduction.
- We discuss the existence of a metaphor-literal repository within LLMs by analyzing contextualized and de-contextualized imagination, demonstrating how context and genre influence metaphor analysis.
- We explore the syntactic structure impact: random word order, part-of-speech (POS) and positional placement of metaphorical words on LLMs' metaphor analysis.
- We reveal that LLMs can produce concept-irrelevant error in metaphor interpretation that may relate to limited ability to leverage contextual and syntactic information.

In Appendix 7.1, we present a detailed compilation of terminologies and definitions in the paper.

2 Related Work

2.1 Metaphor Detection

Metaphor detection involves the computational task of identifying whether a given input contains metaphorical expressions. Early approaches relied on rule-based linguistic frameworks (Group, 2007; Dodge et al., 2015). With the advent of neural networks, metaphor detection has reformulated as a classification task (Rai and Chakraverty, 2020). Wan et al. (2020) incorporated linguistic conceptual norms into machine learning algorithms, which leverage word norms, the informal associations between closely connected words, such as *nurse and doctor* (Church and Hanks, 1990). These norms are influenced by factors such as language (Montefinese et al., 2014), genre (Brysbaert and New, 2009), social context (Cheung et al., 2009) and cultural background (Janschewitz, 2008), leading to domain-specific variations in metaphor detection. Pretrained language models encode metaphorical knowledge within representations and have demonstrated promising results in various tasks (Mao et al., 2024; Ge et al., 2022; Choi et al., 2021).

2.2 Metaphor Interpretation

Metaphor interpretation focuses on unveiling metaphorical mappings and their associated conceptual domains. Sullivan (2013) highlights the constructions of metaphors and their reliance on cross-domain mappings. A key computational method of metaphor interpretation is metaphor component recognition, which identifies and analyzes the source and target domains of metaphors. Sengupta et al. (2024) developed a dataset annotated with conceptual mappings from news editorials to examine political framing. Similarly,

Ge et al. (2022) proposed a concept projection model designed for metaphor detection and component recognition. In pretrained models, frameworks such as Chain-of-Thought (CoT) reasoning combined with knowledge graphs have been introduced to improve metaphor interpretation (Tian et al., 2024). Another approach involves explaining metaphors with literal expressions. Mao et al. (2024) implemented contrastive learning based on SPV, though challenges remain in cross-lingual metaphor interpretation and long-text analysis.

2.3 Metaphor in LLMs

Research in metaphor processing with LLMs has produced exciting findings. Jie Wang and Zhang (2024) proposed a multi-stage generative heuristic-enhanced prompt framework to introduce conceptual background for Chinese metaphor interpretation. Studies also have explored LLMs' associative capabilities to investigate creativity and novelty in metaphor generation (DiStefano et al., 2024; Su et al., 2025). For metaphor detection, Xu et al. (2024) applied CoT reasoning to improve LLMs' performance in multi-modal settings. Aghazadeh et al. (2022) discovered that pretrained models encode metaphorical knowledge within their representations, providing a basis for embedding-level analysis. Even without explicit fine-tuning, LLMs have demonstrated some ability in cross-lingual metaphor detection (Wachowiak and Gromann, 2023). However, current research lacks standardized evaluation metrics (Ge et al., 2023), making it difficult to quantify LLMs' performance across varying aspects of metaphors. Therefore, further studies are needed to investigate LLMs' metaphor-processing capabilities for future refinements.

3 Methodology

Experiments are conducted at both interpretation and detection levels to explore LLMs' metaphorical understanding. The overall experiment framework is illustrated in Figure 1.

Interpretation Level: While multi-choice experiments in metaphor interpretation can be affected by the hallucination, **multiple metaphorical mappings across domains** (e.g., *"thinking is linguistic activity"* can be the generic concept of *"memorization is writing"* or *"ideas are words"*) can also complicate metaphor analysis, which highlights the necessity of analyzing metaphors at conceptual level. Thus, we introduce the **spatial analysis** framework to evaluate conceptual irrelevance in metaphor interpretation. This approach maps LLM-generated interpretations into a high-dimensional embedding space, quantifying concept-irrelevant information of the generated outputs.

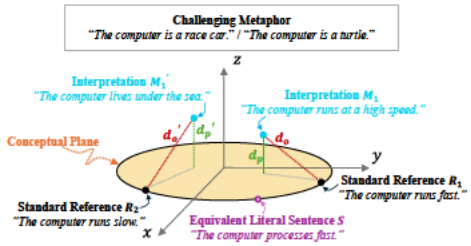

Figure 2: Illustration of spatial analysis in three-dimensional space. d_{pi} is the perpendicular distance from M_i to the conceptual plane. d_{oi} denotes the distance from M_i to R_i.

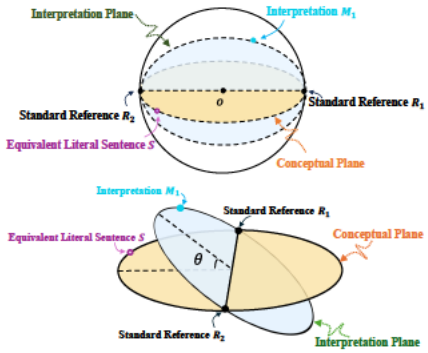

Figure 3: Illustration of angle θ between the conceptual plane and the interpretation plane. The conceptual plane is defined by the representations of R_1, R_2 and S, while the interpretation plane is defined by R_1, R_2 and M_i.

Detection Level: Building on trigger word error identified by Wachowiak and Gromann (2023) and the impact of irregularity in syntactic structure on metaphor detection, we propose two complementary experiments. **Metaphor Imagination**: LLMs are tasked to imagine corresponding literal or metaphorical lexical sets both in context and without context. By comparing the overlap among different settings, we can investigate the influence of context in metaphor analysis. A high overlap implies LLMs inherently connect metaphorical words with commonly co-occurred literal expression–metaphor-literal repository that result in metaphor detection based on specific words regardless of context. **Syntactic Shuffle**: We disrupt the semantic coherence and syntax of metaphor to investigate whether syntactic structures influence metaphor detection. This experiment tests whether LLMs rely on syntactic information for metaphor analysis.

3.1 Spatial Analysis

Problem Definition Inspired by similarity-based framework proposed in style evaluation (Wegmann and Nguyen, 2021), we extend the concept-irrelevance, the comparison of included conceptual mappings between human interpretation of the metaphor and LLM-generated interpretation from two-dimensional to high-dimensional space, analyzing the geometric relationships between sentences. We treat LLM-generated interpretations of a metaphor m_i as representation M_i, and compare it with the representation of actual concept mapping to the metaphor: 1) We embed $\{R_1, R_2\}$ (human interpretations) and S (LLM-generated equivalent interpretation) to collectively construct a **conceptual plane** γ_i–a semantic space delineates the conceptual information in the three sentences, representing the ideal concept mapping the metaphor intended to convey. 2) We measure the projection and discrepancy between M_i interpreted by LLMs and the conceptual plane γ_i to reveal the corresponding concept-irrelevant error.

Methodology As shown in Figure 2 and 3, two measures are defined in the spatial analysis:

1. d_p: The perpendicular distance from generated outputs to the conceptual plane γ_i, indicating conceptual irrelevance.

2. $\cos\theta$: The cosine similarity between conceptual plane γ_i and interpretation plane β_i defined by LLM-generated interpretation M_i and the two human-annotated interpretations $\{R_1, R_2\}$. Representing the concept-irrelevant offset compared with standard conceptual plane.

d_p measures the distance from M_i to the conceptual plane defined by $\{R_1, R_2, S\}$, capturing sentence-plane differences. Meanwhile, $\cos\theta$ quantifies inter-planar divergence: Ideal alignment is achieved when $\cos\theta = 1$, indicating negligible irrelevant information in the linguistic encoding. Together, d_p and $\cos\theta$ provide complementary perspectives for concept-irrelevant information in generated interpretations.

3.2 Metaphorical Imagination

The trigger word error suggests that LLMs may prioritize word-to-word mapping over contextual reasoning in metaphor analysis. Instead of understanding metaphorical relations based on context, LLMs might rely on inherent metaphor and literal connection to align with source and target domains. Therefore, LLMs may have constructed a **metaphor-literal repository**, where words commonly co-occurring in metaphors are directly treated as indicators of conceptual mapping, even in the absence of contextual information. To test this hypothesis, we design a quadripartite experimental paradigm with two perspectives: 1) Literal-Metaphor: LLMs are asked to generate literal words based on contextualized metaphorical expressions and discrete metaphorical words; 2) Metaphor-Literal: LLMs are prompted to generate metaphorical words based on contextually grounded literal substitutes and discrete words. This examines whether LLMs rely on contextual analysis or default to certain fixed associations, discusses their ability to process metaphors beyond word alignment.

3.3 Syntactic Shuffle

Metaphors often exhibit specific syntactic structures (Sullivan, 2013). For accurate metaphor detection, a model also needs to comprehend sentence meanings based on proper syntactic structure. To evaluate the role of syntactic coherence in metaphor detection, we manipulate sentence structures in three ways: 1) **Random Shuffle**: Sentence words are randomly reordered, disrupting both semantic coherence and syntactic structure, creating sequences of unrelated words, removing meaningful sentence patterns. 2) **POS Shuffle**: Preserving the overall meaning of metaphors and altering the specific metaphorical words with synonyms of the same meaning but different POS. This tests the impact of syntactic irregularities on LLMs' metaphor comprehension while minimizing changes to semantic content. 3) **Metaphorical word reposition**: The position of metaphorical word is rearranged to the beginning, a random intermediate location (excluding the beginning, end, and original position), or the end of the sentence, to examine LLMs' focus on the metaphorical word or the overall sentence structure or meaning. Thereby, we investigate how syntactic coherence and word positioning contribute to LLMs' understanding.

4 Experiment

4.1 Datasets

Table 1 provides details of datasets in our experiments. For spatial analysis experiment, we employed Fig-QA, a human-annotated resource for Winograd-style non-literal language understanding tasks (Liu et al., 2022). Fig-QA consists of two pairs of metaphors $\{m_1, m_2\}$ and correspond-

Datasets	Sentences	Example
Fig-QA	2.6k	The computer **is a race car**. | The computer runs fast. The computer **is a tortoise**. | The computer runs slow.
MUNCH (Context)	2.9k	The council **appealed** by ... | The council **petitioned** by ...
MUNCH (Word)	2.9k	**appealed** | **petitioned**
MUNCH (Original)	2.9k	The council **appealed** by case stated.
MUNCH (POS)	1.3k	The council **complainant** (n.) by case stated.
MUNCH (Random)	2.9k	council case **appealed** stated by The.
MUNCH (Beginning)	2.9k	**appealed** The council by case stated.
MUNCH (Middle)	2.9k	The council by case **appealed** stated.
MUNCH (End)	2.9k	The council by case stated **appealed**.

Table 1: The sizes of the Fig-QA (spatial analysis) and MUNCH (metaphor imagination and syntactic shuffle) datasets and examples of sample sentences processed in this study.

ing literal meanings $\{R_1, R_2\}$ grouped into sets of four, sharing the same concept mapping but conveying divergent meanings. To ensure the experiment solely focus on metaphorical expression, minimizing the influence of literal components, metaphor interpretations are required to only alter concept-relevant parts annotated by GPT-4o within sentences (bold in Fig-QA examples).

For metaphorical imagination and syntactic shuffle experiments, we introduced Metaphor Understanding Challenge Dataset (MUNCH) annotated by linguists (Tong et al., 2024), derived from VU Amsterdam Metaphor Corpus (Steen et al., 2010), one of the most widely cited metaphor corpora. MUNCH is designed to challenge LLMs with complex metaphors and we extract data from the paraphrase generation task of MUNCH. In the syntactic shuffle, tokenization protocols and systematic extraction of divergent lexical substitutions are implemented with WordNet 2020 (McCrae et al., 2020), an updated version of English lexical database.

4.2 Models

Our experiments were conducted using following models: DeepSeek-V3-671B (V3-671B) (Liu et al., 2024), Qwen-Turbo (Qwen-T) (Yang et al., 2024a), GPT-4 (Achiam et al., 2023), GPT-4o(Hurst et al., 2024), o3-mini and DeepSeek-R1-671B (R1-671B) (Guo et al., 2025), LLaMA-3.1-8B (LMA-3.1-8B) (Grattafiori et al.). For spatial analysis experiment, we harmonize the interpretation embeddings across models by using text-embedding-3-small model of OpenAI to encode generated outputs. Experiments were conducted on a T4 GPU of Google Colab. For all generations, we set the temperature parameter to 0 for consistent performance.

4.3 Implementation Details

Spatial Analysis Since pretrained models are proven to encode metaphorical knowledge within representations (Aghazadeh et al., 2022), we aim to utilize sentence embeddings to analyze conceptual irrelevance within interpretations. The concept information enriched in the conceptual plane γ_i is derived from R_1, R_2 and S. While R_1 and R_2 are standard references annotated in Fig-QA. To construct the conceptual plane, LLMs are tasked to generate an equivalent literal sentence S that representing the identical mapping of R_1 and R_2. We further project the LLM-generated interpretation onto the concept-irrelevant direction to quantify misinterpretations at concept-level.

Since the major difference of the sentences are the implicit metaphorical concepts, we collect the embeddings of generated interpretations M_i as well as R_1, R_2 and S. To compute the perpendicular distance d_p from M_i to the conceptual plane γ_i, and the $\cos\theta$ between γ_i and interpretation plane β_i (defined by M_i, R_1 and R_2), we use Singular Value Decomposition (SVD) to obtain the subspace bases of the high-dimensional embeddings. Defining a matrix A, where each row corresponds to the embedding of the sentence (e.g., interpretation plane composed by M_i, R_1 and R_2, or conceptual plane composed by R_1, R_2 and S). Using SVD, we decompose A into three components as presented in Eq. (1):

$$A = U\Sigma V^T \tag{1}$$

U is an orthogonal matrix of left singular vectors, Σ is a diagonal matrix containing singular values in descending order, and V^T is an orthogonal matrix of right singular vectors. By selecting the top singular values from Σ along with corresponding

	V3-671B	Qwen-T	GPT-4	GPT-4o	o3-mini	R1-671B	LMA-3.1-8B
d_{p_M}	0.1903	0.2319	0.2267	**0.1772**	0.2020	0.2063	0.2866
$\cos\theta_M$	**0.8207**	0.7835	0.7940	0.7526	0.7931	0.7804	0.7396
$d_{p_{SD}}$	0.2194	0.2386	0.2342	**0.2182**	0.2343	0.2204	0.2649
$\cos\theta_{SD}$	**0.2531**	0.2742	0.2669	0.2905	0.2703	0.2698	0.2918

Table 2: Average d_p and $\cos\theta$ metrics for each model. M and SD represent the mean and standard deviation, respectively. The best values are in bold, and the second-best values are underlined.

singular vectors from U and V^T, a subspace bases column vectors A_r is created, as Eq. (2):

$$A_r = U_k \Sigma_k V_k^T \tag{2}$$

The subspace bases of these two planes encapsulate their respective conceptual information, enabling d_p and $\cos\theta$ to measure the difference between LLM-generated M_i and the conceptual references R_1, R_2 and S from both the sentence and plane perspectives. A larger distance d_p or a smaller $\cos\theta$ (ranging from 0 to 1) indicates a greater deviation of the M_i from the ideal concept.

Metaphorical Imagination The MUNCH dataset include metaphors with single metaphorical word paired with literal substitutions annotated by linguists. LLMs are required to produce twenty substitutions of the target words, reflecting literal or metaphorical semantic concept. If the metaphor-literal repository exists, generated contextual and de-contextualized predictions should demonstrate substantial overlap or similarity. Therefore, the **overlap ratio** is defined to measure the inherent metaphorical imagination regardless of context: overlap ratio equals to one when shared words occur between comparative sets. For non-overlapping sets, compute the overlap ratio using maximum cosine similarity based on 300-dimensional GloVe embeddings (Pennington et al., 2014). Additionally, we analyze the results across different genres.

Syntactic Shuffle To assess the influence of syntactic structure on metaphor understanding, we apply the six types of sentence disruptions described above (random shuffle, POS shuffle, and metaphorical word reposition). The accuracy of metaphor detection is evaluated across these syntactically disrupted sentences to determine whether syntactic coherence affect LLMs' metaphor comprehension.

5 Results & Analysis

The complete experimental and analytical results of each model can be found in the Appendix 7.

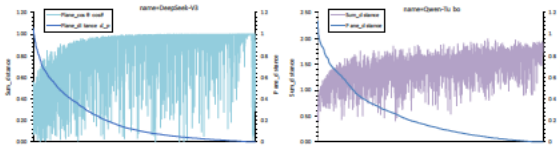

Figure 4: Distributions of (d_p, Ad) and (d_p, $\cos\theta$) for V3-671B and Qwen-T. Fluctuations are attributed to the variance of non-metaphorical parts in the sentences.

5.1 Conceptual Irrelevance

Spatial analysis aims to avoid the potential uncertainty of multiple-choice experiment in LLMs' metaphor interpretation. To validate this problem, we conducted a multiple-choice task: LLMs were tasked with generating two sentences $\{L_{21}, L_{22}\}$ that matched the meanings of references R_2 and metaphor m_2. $\{R_1, L_{21}, L_{22}, R_2\}$, were grouped in random order as possible answers for interpretation of m_1. LLMs were required to select the correct answer. As shown in Appendix 7.2, all seven models yielded an accuracy of approximately 50%, indicating multiple-choice tasks can be unreliable in evaluating LLMs in metaphor interpretation.

To validate the effectiveness of d_p and $\cos\theta$, we introduce an additional measure Ad, the sum of cosine similarities between the metaphor interpretation M and references $\{R_1, R_2\}$. Intuitively, interpretations exhibiting high semantic congruence with both references are expected to have higher respective similarities, Ad, with greater probability to be closer to the concept mapping (lower d_p). For V3-671B and Qwen-T, Figure 4 demonstrated that the d_p from M to conceptual plane, along with Ad, exhibited a mean Spearman's correlation (Spearman, 1961) of ρ_1= -0.62. The $\cos\theta$ metric, computing the angle between interpretation plane and the conceptual plane, exhibited a correlation of ρ_2= -0.64 with d_p, collectively validate d_p and $\cos\theta$ as effective quantifiers of conceptual irrelevance.

As summarized in Table 2, GPT-4o achieved the lowest d_p value, indicating its superior ability to suppress irrelevant errors in interpretation

R_1	R_2	*Metaphor*	S	M_i	d_p	$\lvert\cos\theta\rvert$
The monks were very honorable.	The monks were not honorable.	The monks had the honor of a knight.	... were highly respected.	... had a prestigious recognition.	0.1153	0.9034
		The monks had the honor of a lawyer.	... were highly respected.	*... had the privilege of legal representation.*	*0.7913*	*0.2609*
I can eat a lot.	I eat little.	I have the appetite of an elephant.	I consume a moderate amount of food.	I have a very large appetite.	0.1367	0.9784
		I have the appetite of a chipmunk.	I consume a moderate amount of food.	I have a very small appetite.	0.1573	0.9646

Table 3: Example cases in spatial analysis. The interpretation with less concept-relevance is in *italic*.

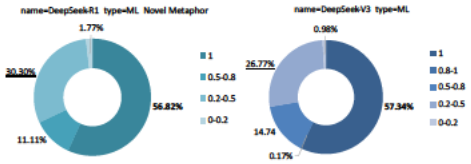

Figure 5: Overlap ratio distributions of ML (novel metaphor) imagined by R1-671B and LM by V3-671B. The best values are in bold, the second are underlined.

at the single-utterance level. Meanwhile, $\cos\theta$ highlighted holistic discrepancies between interpretation plane and conceptual plane, with V3-671B achieving the best performance. However, it revealed that concept-irrelevant information persists in 15%–25% of the output across all tested LLMs. While LLMs can generate interpretations of metaphors, there are still significant challenges in achieving deeper conceptual domain understanding. This further underscores the need for advancements in metaphor processing, particularly in aligning outputs at concept-level rather than superficial lexical mappings. In Table 3, we present several metaphor interpretations generated by LLMs, including those that accurately capture the concept mapping as well as those exhibiting noticeable deviations with references. Notably, this analysis explicitly excluded preset or fine-tuned models, focusing on the native cognitive capacities of LLMs. A consistent embedding model encoding outputs of all LLMs ensures the variance can only attributed to interpretations.

5.2 Context Leveraging

Overlap ratio in metaphorical imagination are summarized in Table 4. GPT-4, o3-mini, and R1-671B achieved strong performance. Notably, there are 65%–80% of overlap between contextualized and de-contextualized outputs, regardless of scenarios, suggesting that LLMs are likely to interpret concept mapping from individual words. Interestingly, imagination from metaphor to literal (ML) almost consistently achieved higher overlap than from literal to metaphor (LM), aligns with the discovery of Liu et al. (2022): matching metaphorical expressions to literal phrases is generally more challenging than the reverse. Performance across genres are various. With more robust inferential understanding of daily language, o3-mini and R1-671B significantly outperformed in conversational metaphors, while GPT-4 achieved the best result in both LM and ML scenarios of NEWS. We assume this variance stems from differences in genre-specific training data, as models tend to maintain consistency within a single genre in both ML and LM tasks.

Metaphors in MUNCH with a novelty score higher than 0.3 are identified as novel metaphors, impose higher demands on LLMs' contextual reasoning ability compared to conventionalized metaphors (e.g., fall in love). Therefore, we additionally analyzed the ML overlap ratio (74.89%) distribution of novel metaphor under R1-671B, presented alongside the ML results under V3-671B in Figure 5. While there is an observed drop in novel metaphor compared with overall result, it remains to be consistent that the overlap ratio equals to 1 over 50% of cases (generating identical words with and without context), indicating LLMs primarily rely on inherent metaphor-literal associations rather than leveraging contextual information. Ratios ranging from 0.2 to 0.5 formed the secondary proportion, suggesting that only a subset of words can be identified independently of context; the rest exhibit weak association. These findings support that LLMs possess a metaphor-literal repository and their ability to utilize context remains limited.

5.3 Syntactic Influence

We evaluated metaphor detection accuracy across six distinct syntactic scenarios, as detailed in Table

	V3-671B	Qwen-T	GPT-4	GPT-4o	o3-mini	R1-671B	LMA-3.1-8B
LM	73.30	73.45	**78.92**	76.28	72.25	78.11	65.09
ML	76.96	75.17	79.22	78.01	**81.55**	80.87	72.86
News (LM)	74.43	74.91	**80.96**	76.90	72.76	78.00	66.28
News (ML)	77.82	75.23	**83.77**	79.32	81.27	80.68	72.11
Fiction (LM)	73.95	73.55	76.06	**79.30**	68.20	76.22	64.32
Fiction (ML)	75.21	73.19	74.73	75.70	**78.73**	77.97	67.79
Academic (LM)	75.02	76.11	**82.71**	77.15	71.09	80.37	65.46
Academic (ML)	80.69	79.69	84.93	81.84	**85.25**	83.62	75.09
Conversation (LM)	66.80	66.17	69.74	71.20	73.22	**75.00**	61.27
Conversation (ML)	73.35	71.13	71.16	73.86	**81.87**	81.11	73.94

Table 4: The overlap ratio of metaphorical imagination from literal to metaphor (LM) and from metaphor to literal (ML). *News*, *Fiction*, *Academic* and *Conversation* are four included genres. The best values are in bold and the second-best are underlined.

	Original	Random	POS	Beginning	Middle	End
V3-671B	18.31	22.71	**23.59**	19.14	22.10	22.63
Qwen-T	30.37	1.17	**33.59**	25.46	27.15	27.50
GPT-4	34.73	12.93	**43.74**	36.07	37.92	37.60
GPT-4o	28.89	7.78	**36.87**	30.92	30.84	29.98
o3-mini	29.87	5.40	**38.63**	25.27	25.85	25.58
R1-671B	28.68	12.22	**46.41**	39.25	30.88	36.03
LMA-3.1-8B	53.36	50.33	**53.81**	51.75	53.08	53.67

Table 5: The accuracy of metaphor detection with shuffled sentences. The best values are in bold.

5. LLaMA-3.1-8B demonstrated consistent accuracy of around 50% across all tasks. Combined with its low overlap ratio in metaphorical imagination, suggesting a chance-level performance on the dataset and a fundamental lack of metaphorical understanding. According to our assumption, if syntactic irregularities contribute to metaphor detection, models should exhibit improved performance under certain conditions. As observation, LLMs perform better in POS shuffle than the original metaphors. This reflects alignment with the SPV theory, where irregular language usage is treated as the indicator of metaphorical expression. Unlike random shuffle, POS shuffle preserves the overall meaning of metaphors while altering syntactic structure, which resembles SPV scenarios more closely. The consistent performance across three positional paradigms (beginning, middle, and end shuffles) suggests that LLMs tend to rephrase abnormal expressions, interpreting these deviations as cognitive markers of metaphor. For V3-671B, even random shuffle scenario outperformed original sentences, implying that the model may prioritize metaphorical words over syntactic structures. Cross-model performance comparisons reveal significant heterogeneity, reaffirming that MUNCH is a challenging dataset for LLMs to understand. These results highlight discrepancies in metaphor detection across syntactic structures, emphasizing the need for models to better integrate syntactic information into metaphor analysis.

6 Conclusion

This study investigates LLMs' metaphorical understanding through conceptual irrelevance, context leveraging and syntactic influence. Corresponding to three experiments: spatial analysis, metaphorical imagination and syntactic shuffle: Spatial analysis highlights the persistent concept-irrelevant errors in interpretations, while metaphorical imagination suggests LLMs rely on inherent metaphor-literal repository rather than effectively leveraging context for in-depth analysis. Syntactic shuffle presents limitation in LLMs' syntactic comprehension, with models often prioritizing irregular language usage over structural understanding. While LLMs demonstrate some surface-level competence, their performance remains inconsistent, underscoring the need for advancements in conceptual alignment, contextual reasoning and syntactic integration.

Limitations

In the spatial analysis experiment, an appropriate dimensionality reduction method could yield more explicit results in spatial analysis, but the effectiveness of the ideal approach is hard to define. Additionally, the cross-linguistic challenges in metaphor comprehension are closely connected with translation tasks, yet the capacity of LLMs to understand metaphors across multiple languages remains unexplored.

Since our experiments are implemented on LLMs without fine-tuning or few-shot prompting, future work could explore the growth of learning ability under these scenarios. In our work, we focus on LLMs' understanding of metaphor, while pretrained models like BERT and RoBERTa have also demonstrated promising results in this field. However, these models exhibit domain-limited analytical capabilities due to the constraints of their training datasets. Therefore, to enhance LLMs' performance in general metaphor, we may further focus on the alignment between metaphor understanding of LLMs and human conceptual mapping processes of metaphor.

A minimal proportion of the source dataset (constituting <0.1% of total volume) contained inputs violating LLM content moderation protocols. Biases, violence, and similar issues may be implicitly embedded in the data. Since our experiments were conducted solely on English datasets, the cultural contexts representation are also relatively homogeneous.

Ethics Statement

This research is conducted on two public datasets Fig-QA (Liu et al., 2022) and MUNCH (Tong et al., 2024). The corpora have solely been deployed in the probing experiments on LLMs. The experiments strictly excluded any materials associated with personal identifiers or sensitive data categories.

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*.

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

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Man Yee Cheung, Chuan Luo, Choon Ling Sia, and Huaping Chen. 2009. Credibility of electronic word-of-mouth: Informational and normative determinants of on-line consumer recommendations. *International journal of electronic commerce*, 13(4):9–38.

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.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

William Croft. 1993. *The role of domains in the interpretation of metaphors and metonymies*. Walter de Gruyter, Berlin/New York Berlin, New York.

Paul V DiStefano, John D Patterson, and Roger E Beaty. 2024. Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal*, pages 1–15.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

Ellen Dodge, Jisup Hong, and Elise Stickles. 2015. MetaNet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49, Denver, Colorado. Association for Computational Linguistics.

Mengshi Ge, Rui Mao, and Erik Cambria. 2022. Explainable metaphor identification inspired by conceptual metaphor theory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10681–10689.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, 56(Suppl 2):1829–1895.

Raymond W Gibbs Jr. 2008. *The Cambridge handbook of metaphor and thought*. Cambridge University Press.

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

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Nicholas Ichien, Dušan Stamenković, and Keith J Holyoak. 2024. Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol*, 39(4):296–309.

Kristin Janschewitz. 2008. Taboo, emotionally valenced, and emotionally neutral word norms. *Behavior research methods*, 40(4):1065–1074.

Jin Wang Jie Wang and Xuejie Zhang. 2024. Chinese metaphor recognition using a multi-stage prompting large language model. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 234–246. Springer.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. Chicago University Press.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2819–2834, Torino, Italia. ELRA and ICCL.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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

Rui Mao, Kai He, Claudia Ong, Qian Liu, and Erik Cambria. 2024. MetaPro 2.0: Computational metaphor processing on the effectiveness of anomalous language modeling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9891–9908, Bangkok, Thailand. Association for Computational Linguistics.

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 14–19, Marseille, France. The European Language Resources Association (ELRA).

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the affective norms for english words (ANEW) for italian. *Behavior research methods*, 46:887–903.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.

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.

Yujie Shao, Xinrong Yao, Xingwei Qu, Chenghua Lin, Shi Wang, Wenhao Huang, Ge Zhang, and Jie Fu. 2024. CMDAG: A Chinese metaphor dataset with annotated grounds as CoT for boosting metaphor generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3357–3366, Torino, Italia. ELRA and ICCL.

Charles Spearman. 1961. "General intelligence" objectively determined and measured.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, Tina Krennmayr, and Tryntje Pasma. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins Publishing Company.

Chang Su, Xingyue Wang, Yongzhu Chang, Kechun Wu, and Yijiang Chen. 2025. Metaphor generation

based on noval evaluation method. *Neurocomputing*, 611:128651.

Karen Sullivan. 2013. *Frames and Constructions in Metaphoric Language*, volume 14. John Benjamins Publishing.

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.

Michael Toker, Oren Mishali, Ophir Münz-Manor, Benny Kimelfeld, and Yonatan Belinkov. 2024. A dataset for metaphor detection in early medieval Hebrew poetry. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 443–453, St. Julian's, Malta. Association for Computational Linguistics.

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

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.

Mingyu Wan, Kathleen Ahrens, Emmanuele Chersoni, Menghan Jiang, Qi Su, Rong Xiang, and Chu-Ren Huang. 2020. Using conceptual norms for metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 104–109, Online. Association for Computational Linguistics.

Anna Wegmann and Dong Nguyen. 2021. Does it capture STEL? a modular, similarity-based linguistic style evaluation framework. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7109–7130, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024. Exploring chain-of-thought for multimodal metaphor detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024b. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

7 Appendix

In this section, we provide the terminologies and definitions included in this paper, the specific prompt settings, as well as the visualization results for each model and experiment.

7.1 Definition and Terminology

This paper requires multiple definitions and terminologies to explain the motivation and support our experimental design. Thus, in Table 6 and Table 7, we present a detailed compilation of terminologies and definitions mentioned in the paper.

7.2 Multi-choice Validation Experiment

In order to explore the effectiveness of multi-choice experimental settings in metaphor understanding, we conduct the validation experiment. In Fig-QA dataset, the two metaphors $\{m_1, m_2\}$ in one group describe the identical concept with different meanings. Therefore, from $\{R_1, L_{21}, L_{22}, R_2\}$ (L_{21} and L_{22} are two sentences generated by LLMs with the identical meaning of R_2), we deploy LLMs to choose the most appropriate interpretation for m_1. The example of experiment and the accuracy of multi-choice validation are shown in Table 8.

7.3 Prompts of Spatial Analysis

Two prompts in Figure 6 are designed for the spatial analysis experiment. Prompt 1 is intended to generate a synonymous sentence S aligning with predefined standard literal sentences; prompt 2 instructs models to generate the literal interpretation M_i of given metaphor m_i, by substituting metaphorical component with literal expression.

7.4 Prompts of Metaphorical Imagination

Prompts in Figure 7 are designed for metaphorical imagination. LLMs are required to perform based on the given prompts, to imagine both with and without context: 1) the literal words corresponding to given metaphorical words, and 2) the metaphorical words corresponding to given literal words. Prompts are subdivided into two types, depending on whether contexts are provided in the experiment.

7.5 The Prompt of Syntactic Shuffle

LLMs detect metaphors through the prompt in Figure 8 in syntactic shuffle experiment.

7.6 Distribution of (d_p, $\cos\theta$) and (d_p, Ad)

We present the distance d_p between generated interpretation M_i and conceptual plane γ_i of each metaphor, respectively correlated with $\cos\theta$ of interpretation and conceptual plane $\{\beta_i, \gamma_i\}$, and the sum Ad of similarity between interpretation and standard references in Figure 9–10.

7.7 Overlap Ratio of Imagination Word Sets

In the metaphorical imagination experiment, we analyze the overlap ratio distribution within imagination word sets, the distribution of overlap ratio under each task and each genre is shown in Figure 11–20. The portion of different overlap ratios reveals the metaphorical imagination results of LLMs, where a higher overlap ratio (especially 1) suggests that the LLM's metaphor analysis is detached from the context.

Terminology	Definition	Notes/Examples
Selection Preference Violation (SPV) (Wilks, 1975)	The disparity between the context of a word within a sentence and its frequently used contexts is an indicator of this word's metaphorical usage (Tian et al., 2024).	I **drank** a bottle of water. Cars **drink** gasoline. (Drinking usually connected with human, not car.)
Metaphor Identification Procedure (MIP) (Group, 2007)	A metaphor is identified if the contextual meaning of the word differs from its basic meaning (Tian et al., 2024).	They **fall** in love. Metaphorical: obsessed in feelings Literal: dropping down
Conceptual Metaphor Theory (CMT) (Lakoff and Johnson, 1980)	Metaphors are mappings between source domain (describes tangible objects or concepts) and target domain (represents abstract ideas).	Metaphor: The arm race. Source domain: COMPETITION Target domain: ARMS BUILDUP Conceptual mapping: COMPETITION IS ARMS BUILDUP
Multiple Metaphor-Literal Mappings across Domains	A pair of source and target domains can encompass multiple metaphorical mappings. The same meaning in metaphor can be expressed through multiple comparison of metaphorical and literal indication.	The mapping "thinking is linguistic activity" can be the generic concept of "memorization is writing" or "ideas are word".
Trigger Word Error (Wachowiak and Gromann, 2023)	The model predict wrong source domains that were not metaphorically related, because models only infer from the words that are commonly co-occurred instead of considering context.	The conceptual mapping of "The arms race." is falsely identified as COMPETITION IS WAR, due to the commonly co-occurrence of "arm" and "war" in daily discourse.
Concept-irrelevant Error/Conceptual Irrelevance	The interpretation generated by LLMs failed to clarify the actual target domain that source domain intended to map.	Interpreting falling in love as physically falling rather than entering an emotional state.
Context Leveraging	Leveraging holistic context when analyzing metaphor rather than focusing on specific words.	Metaphor: The arm race. Conceptual mapping: COMPETITION IS ARMS BUILDUP

Table 6: Terminology and definitions related to metaphor identification and context analysis in this study (Part 1).

Terminology	Definition	Notes/Examples
Conceptual Plane γ_i	Constructed by the embeddings of three sentences R_1, R_2 and S, implying correct concept mapping. Representing the ideal concept the metaphor intended to convey.	R_1: The computer runs fast. R_2: The computer runs slow. S: The computer processes fast. The conceptual plane include the concept of the speed of computers.
Interpretation Plane β_i	The plane defined by LLM-generated interpretation M_i and the two human-annotated interpretations R_1 and R_2 to evalutate the diviation of interpertations.	R_1: The computer runs fast. R_2: The computer runs slow. M: The computer runs at a high speed.
Metaphor-Literal Repository	An abstract repository inheretly exists in LLMs. When analyzing metaphor, specific word will activate this repository and result in resoponse without effectively analyzing context.	The conceptual mapping of "The arms race." is falsely identified as COMPETITION IS WAR. Indicating "arm" and "war" may be inheretly paired in metaphor understanding of LLMs.
Syntactic Influence	The influence of metaphorical syntactic structures in metaphor analysis.	The accuracy of metaphor detection varies depending on different syntactic irregularity settings.
Random Shuffle	Sentence words are randomly reordered, disrupting both semantic coherence and syntactic structure, creating sequences of unrelated words, removing meaningful sentence patterns.	council case **appealed** stated by The.
Part-of-speech (POS) Shuffle	Preserving the overall meaning of metaphors and altering the specific metaphorical words with synonyms of the same meaning but different POS.	The council **complainant** (n.) by case stated.
Metaphorical Word Reposition	The position of metaphorical word is rearranged to the beginning, a random intermediate location (excluding the beginning, end, and original position), or the end of the sentence.	Beggining: appealed The council by case stated. Middle: The council by case appealed stated. End: The council by case stated appealed.

Table 7: Terminology and definitions related to conceptual planes and metaphor processing in this study (Part 2).

m_1: **This blanket is as insulating as a wet tissue.**	**Accuracy of Multi-choice Validation**			
L_{21}: The blanket keeps me really cozy.	V3-671B	Qwen-T	GPT-4	GPT-4o
R_1: The blanket does not keep me warm.	50.89	51.69	50.77	50.92
L_{22}: The blanket makes me feel quite warm.	o3-mini	R1-671B	LLaMA-3.1-8B	
R_2: The blanket keeps me very warm.	50.85	47.31	46.04	

Table 8: The example and accuracy of multi-choice validation.

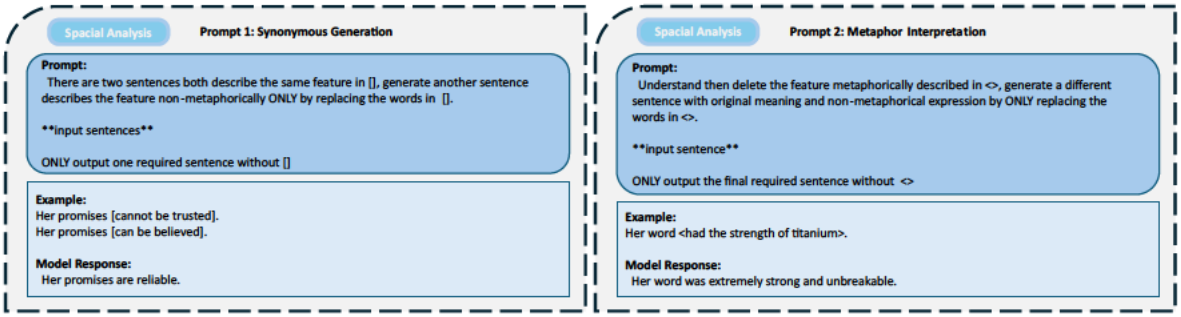

Figure 6: The prompts of Spatial Analysis.

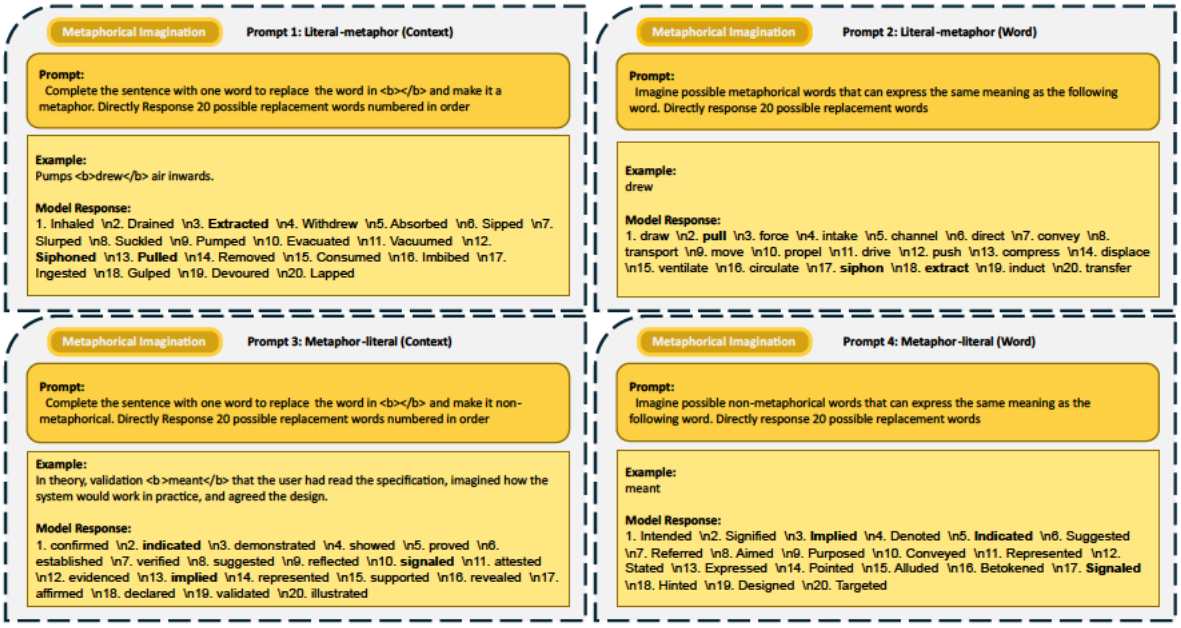

Figure 7: The prompts of Metaphorical Imagination.

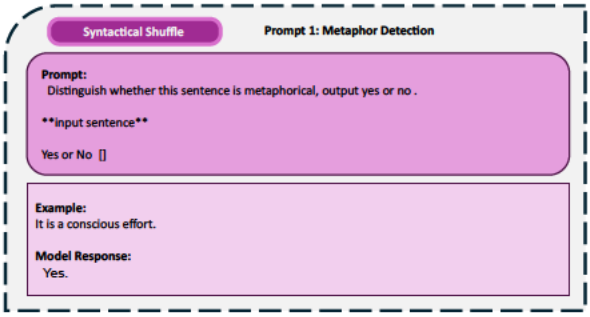

Figure 8: The prompt of Syntactic Shuffle.

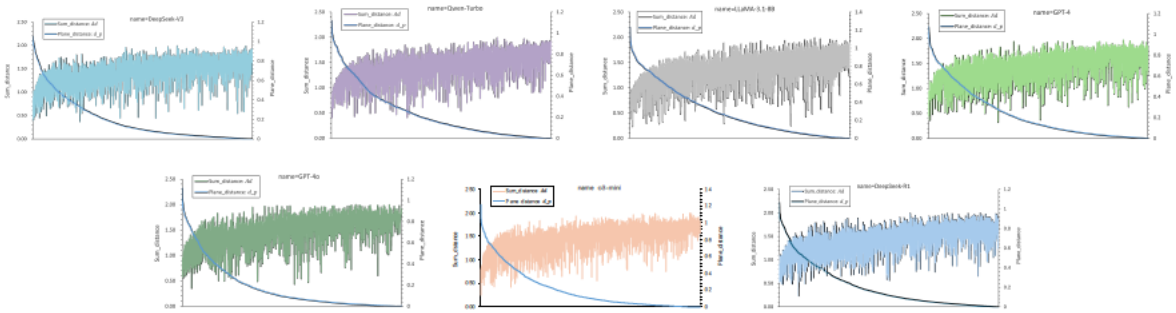

Figure 9: The $(d_p,\ Ad)$ distribution of every model. Sort d_p in decreasing order. Significant fluctuation can be observed due to the variance in the non-metaphorical part of sentences.

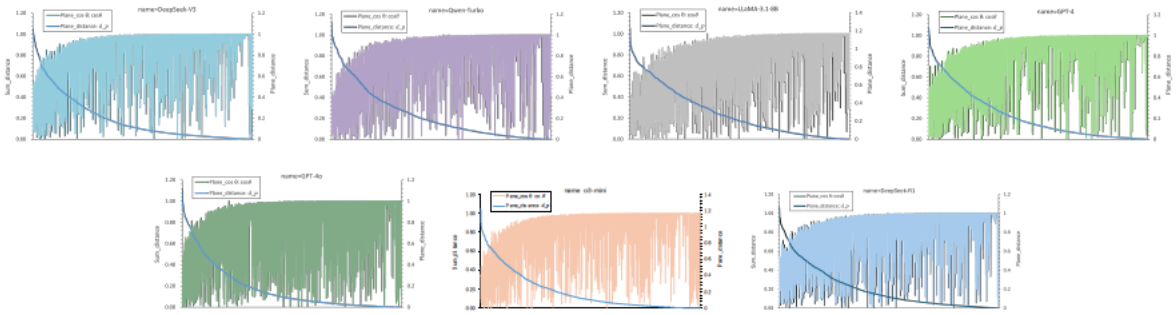

Figure 10: The $(d_p,\ \cos\theta)$ distribution of every model. Sort d_p in decreasing order. Significant fluctuation can be observed due to the variance in the non-metaphorical part of sentences.

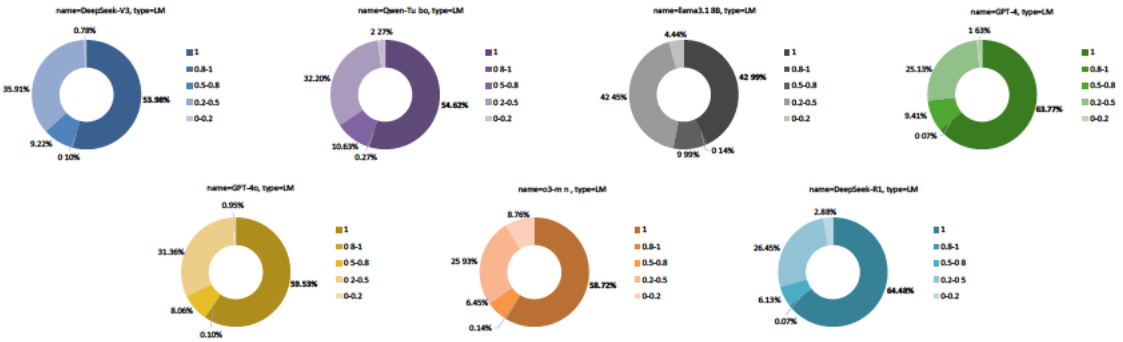

Figure 11: The overlap ratio distributions of literal-metaphor (LM) word imagination task on every model (the largest portion is in bold and the second largest is underlined).

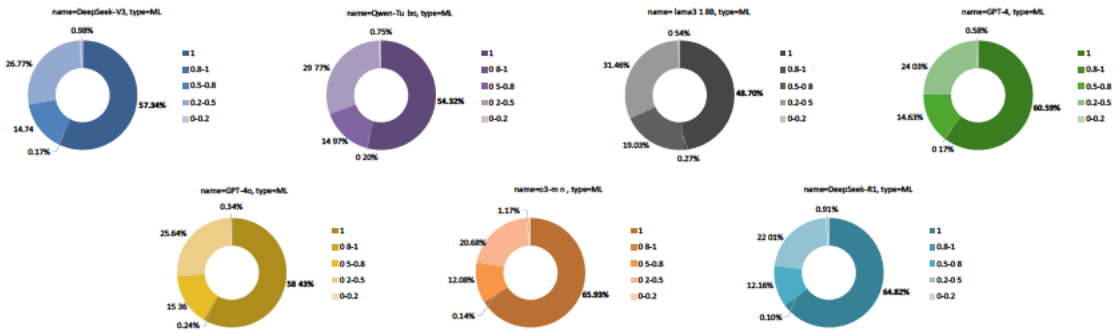

Figure 12: The overlap ratio distributions of metaphor-literal (ML) word imagination task on every model (the largest portion is in bold and the second largest is underlined).

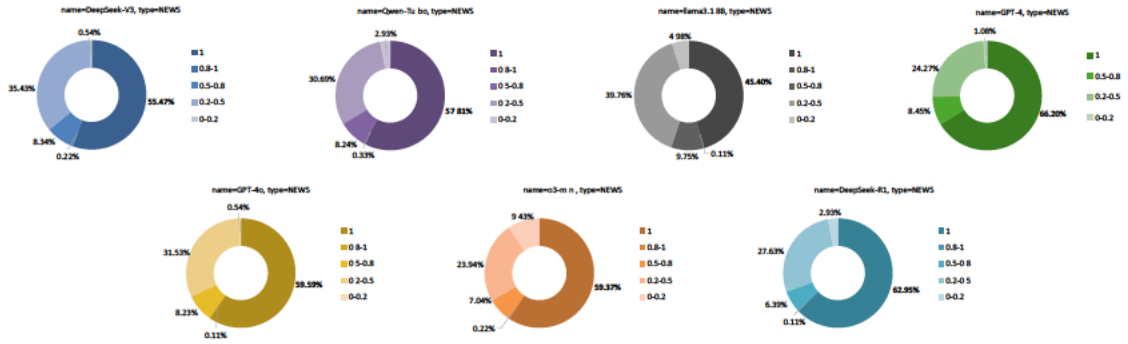

Figure 13: The overlap ratio distributions of literal-metaphor (LM) word imagination task with sentences in NEWS on every model (the largest portion is in bold and the second largest is underlined).

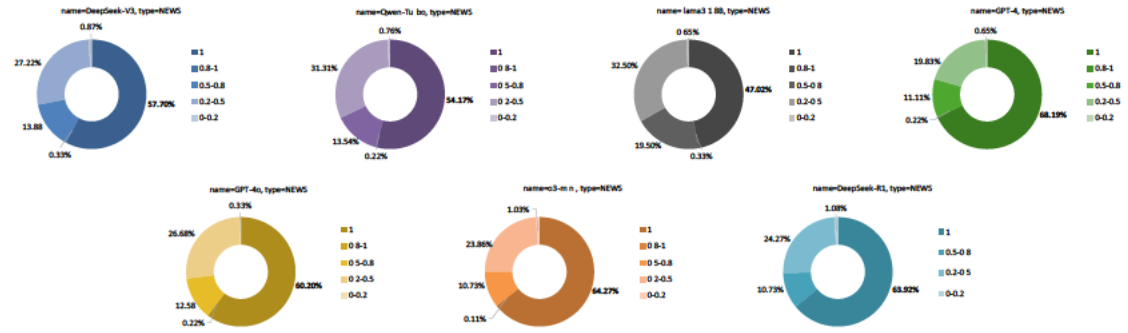

Figure 14: The overlap ratio distributions of metaphor-literal (ML) word imagination task with sentences in NEWS on every model (the largest portion is in bold and the second largest is underlined).

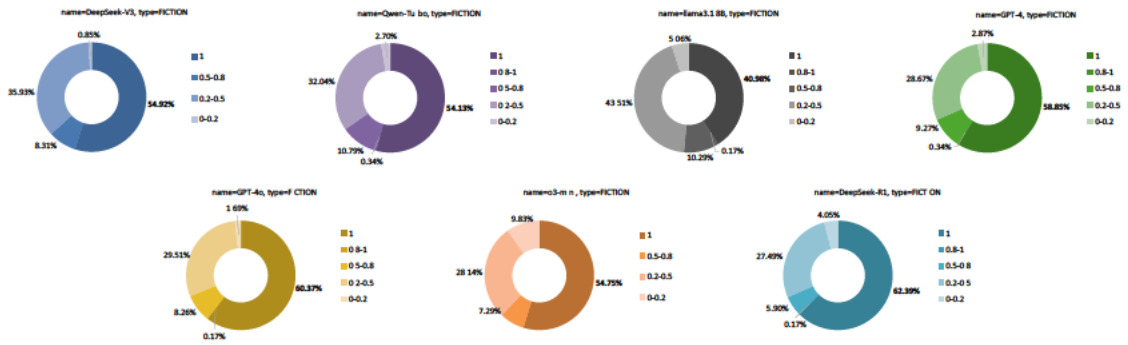

Figure 15: The overlap ratio distributions of literal-metaphor (LM) word imagination task with sentences in FICTION on every model (the largest portion is in bold and the second largest is underlined).

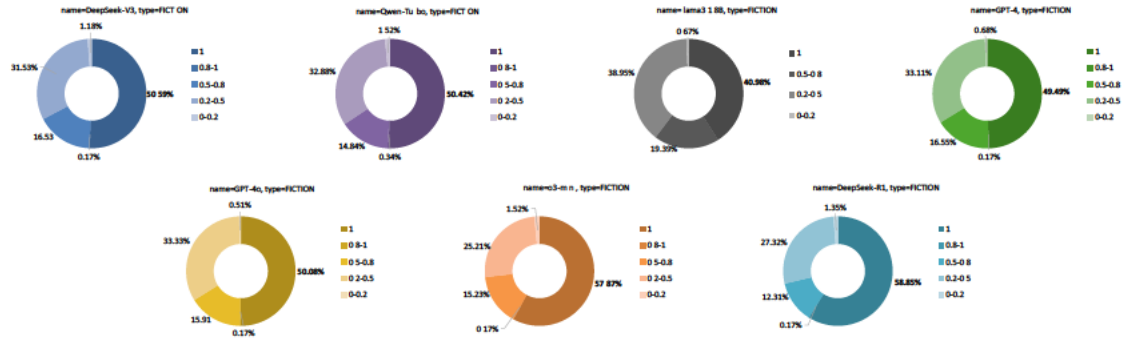

Figure 16: The overlap ratio distributions of metaphor-literal (ML) word imagination task with sentences in FICTION on every model (the largest portion is in bold and the second largest is underlined).

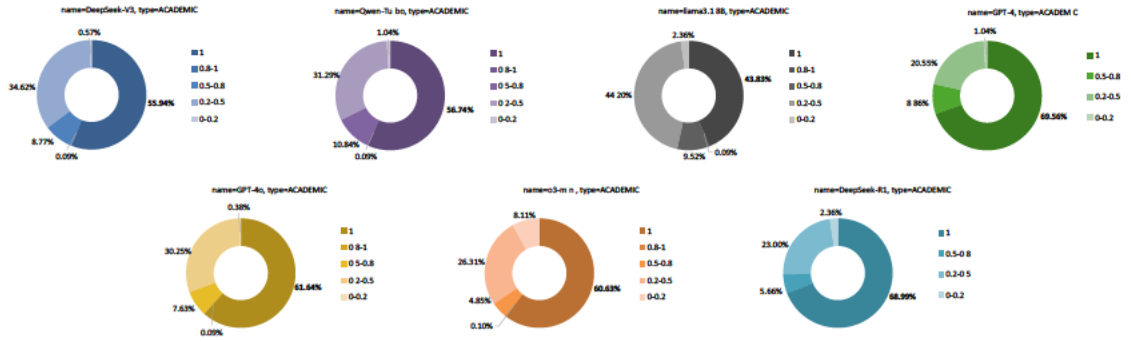

Figure 17: The overlap ratio distributions of literal-metaphor (LM) word imagination task with sentences in ACADEMIC on every model (the largest portion is in bold and the second largest is underlined).

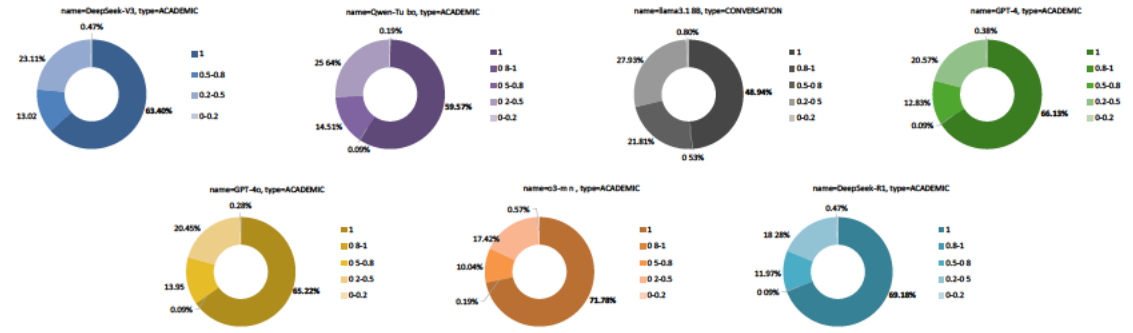

Figure 18: The overlap ratio distributions of metaphor-literal (ML) word imagination task with sentences in ACADEMIC on every model (the largest portion is in bold and the second largest is underlined).

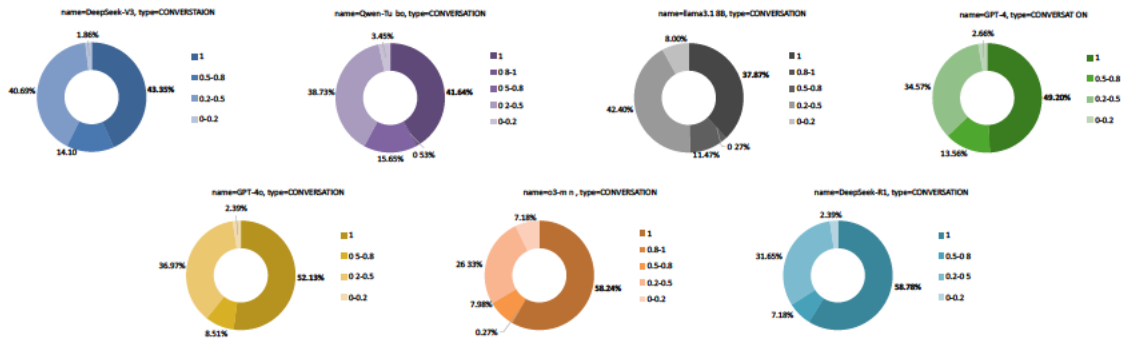

Figure 19: The overlap ratio distributions of literal-metaphor (LM) word imagination task with sentences in CONVERSATION on every model (the largest portion is in bold and the second largest is underlined).

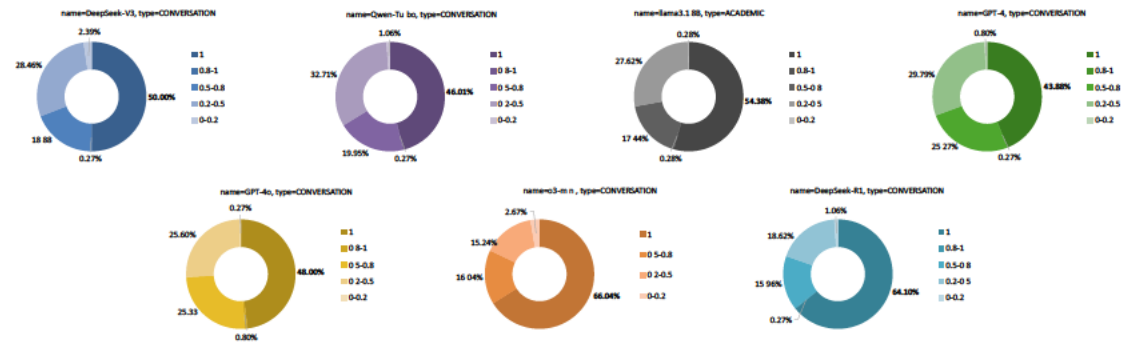

Figure 20: The overlap ratio distributions of metaphor-literal (ML) word imagination task with sentences in CONVERSATION on every model (the largest portion is in bold and the second largest is underlined).