Probing Semantic Alignment, Lexical Invariance, and Syntactic Influence in LLM Metaphor Processing

Fengying Ye[1] **Shanshan Wang**[1] **Lidia S. Chao**[1] **Derek F. Wong**[1*]
[1]NLP[2]CT Lab, Department of Computer and Information Science, University of Macau
nlp2ct.{fengying,shanshan}@gmail.com, {lidiasc,derekfw}@um.edu.mo

Abstract

Large language models (LLMs) achieve strong performance on metaphor detection and interpretation tasks, yet it remains unclear what such behavioral success reveals about metaphor processing. We present a diagnostic analysis that examines the limits of behavioral evidence by probing three complementary dimensions: semantic attribute alignment, lexical invariance, and syntactic sensitivity. Using geometric probing, we assess whether model-generated interpretations align with reference semantic attributes; through context-varying substitution, we analyze the stability of lexical associations between metaphorical and literal expressions; and via controlled syntactic perturbations, we examine sensitivity in metaphor detection. Our analysis reveals that LLM-generated interpretations can exhibit semantic drift relative to reference attributes; stable lexical anchors persist across contextual conditions, potentially supporting conventional metaphors while biasing novel metaphors requiring contextual integration; and detection performance is sensitive to syntactic irregularities. These findings suggest that strong behavioral performance may reflect heterogeneous underlying signals, highlighting the need for caution when interpreting metaphor benchmarks as evidence of robust, integrated semantic understanding.

1 Introduction

Metaphor is a pervasive and sophisticated aspect of human language (Gibbs Jr, 2008). Processing metaphors requires more than recognizing unusual word usage; it involves identifying implicit relationships between attributes across semantic domains (Croft, 1993). With strong text comprehension capabilities and large-scale pretraining (Yang et al., 2024b), LLMs have been widely applied to metaphor detection and interpretation. However, it remains unclear whether such performance is accompanied by behavioral evidence consistent with deep metaphor processing.

Linguistic theories of metaphor such as Selection Preference Violation (SPV) and the Metaphor Identification Procedure (MIP) characterize metaphor through violations of conventional selection preferences or literal word meanings in context (Wilks, 1975; Group, 2007). Conceptual Metaphor Theory (CMT), in contrast, views metaphors as cross-domain mappings between a source domain describing tangible objects and a target domain representing abstract ideas (Lakoff and Johnson, 1980). A central difficulty emphasized by these theories is that the core mapping in a metaphor is often implicit rather than explicitly expressed. As a result, models may generate interpretations that focus on salient characteristics while failing to capture the intended mapping attribute. In this work, a *semantic attribute* refers to the salient property selectively projected from the source domain to the target domain in a metaphor. For example, *"The computer is a turtle"* may evoke (low) speed, but also peripheral attributes of turtle (e.g., (long) lifespan), complicating interpretation. This motivates analyzing metaphor processing in terms of whether model-generated interpretations align with the intended semantic attribute (Do Dinh et al., 2018).

Recent studies have applied LLMs to metaphor processing across cultural contexts (Ichien et al., 2024), cross-lingual settings (Shao et al., 2024), and different genres (Toker et al., 2024; Wang et al., 2024b). However, prior work has also identified behavioral patterns that complicate the interpretation of performance on metaphors: Wachowiak and Gromann (2023) identify **trigger word** effects, where interpretations are biased toward highly associated lexical items rather than context. For example, the word *arm* may bias interpretations toward war-related meanings, even when the context does not support such mappings. While prior work

*Corresponding Author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27914–27932
July 2-7, 2026

focused on prediction outcomes such as multiple-choice accuracy (Li et al., 2024; Zhao et al., 2021), we investigate behavioral patterns that shed light on how LLMs process metaphors.

We investigate LLM metaphor processing from a diagnostic perspective, asking whether observed behavioral performance reflects a unified semantic mechanism or heterogeneous underlying signals. To this end, we analyze model behavior along three complementary dimensions. First, we examine whether model-generated interpretations exhibit semantic attribute alignment with reference interpretations. Second, we test whether LLMs exhibit context-invariant lexical associations, which may indicate reliance on stable lexical anchors rather than contextual integration while processing metaphor. Third, we assess how syntactic disruption influences metaphor detection, probing sensitivity to structural cues. Together, these dimensions provide a structured view of metaphor processing, allowing us to distinguish between semantic alignment, lexical bias, and syntactic sensitivity in LLM behavior. Our contributions are as follows:

- We propose a geometric probing framework as a behavioral proxy to measure semantic attribute alignment in metaphor interpretation.
- We probe the persistence of context-invariant lexical associations using Metaphorical Imagination tasks under different context settings.
- We analyze the effects of controlled syntactic perturbations on metaphor detection by selectively disrupting word order, part-of-speech, and positional placement of metaphorical words.
- We identify consistent behavioral patterns across LLMs under these probes, highlighting behavioral regularities and limitations in how models process metaphor-related inputs.

Rather than focusing on performance comparison, we adopt a diagnostic perspective to analyze how LLMs process metaphors under controlled probing conditions. Definitions and terminology used in this paper are provided in Appendix A.

2 Related Work

2.1 Metaphor Detection

Metaphor detection aims to determine whether a given input contains metaphorical expressions. Early approaches relied on rule-based linguistic frameworks (Group, 2007; Dodge et al., 2015), whereas later neural models formulated the task as supervised classification (Rai and Chakraverty, 2020). To better capture metaphorical usage, subsequent work incorporated linguistic and conceptual signals, such as word association statistics, to model semantic relatedness and deviation (Wan et al., 2020; Church and Hanks, 1990). Because metaphorical expressions vary across languages, genres, and socio-cultural contexts, prior studies have also explored domain-specific detection settings (Montefinese et al., 2014; Brysbaert and New, 2009; Cheung et al., 2009; Janschewitz, 2008; Wang et al., 2025). More recently, LLM-based approaches have achieved strong performance on metaphor detection benchmarks (Mao et al., 2024; Ge et al., 2022; Choi et al., 2021).

2.2 Metaphor Interpretation

Metaphor interpretation concerns explaining the meaning of metaphorical expressions, including the conceptual mappings that relate source and target domains. Linguistic theories emphasize that such interpretations arise from systematic cross-domain mappings rather than isolated lexical substitutions (Sullivan, 2013). In computational research, one line of work formulates metaphor interpretation as metaphor component recognition, aiming to identify the source and target domains underlying a metaphor (Sengupta et al., 2024; Ge et al., 2022). Prior studies have also shown that performance in this setting can be affected by trigger word effects (Wachowiak and Gromann, 2023). A separate line guides interpretation with explicit reasoning mechanisms. For example, some frameworks combine Chain-of-Thought (CoT) reasoning with external knowledge resources to generate more structured interpretations (Tian et al., 2024). Other studies treat interpretation as literal paraphrase generation, often including SPV as a supervision signal (Mao et al., 2024).

2.3 Metaphor in LLMs

Recent work has investigated metaphor processing in LLMs at both the task and representation levels. At the task level, prior studies have examined metaphor interpretation, generation, and detection. Wang et al. (2024a) proposed a multi-stage prompt-based framework incorporating conceptual background knowledge for Chinese metaphor interpretation, while other work explored LLMs' associative

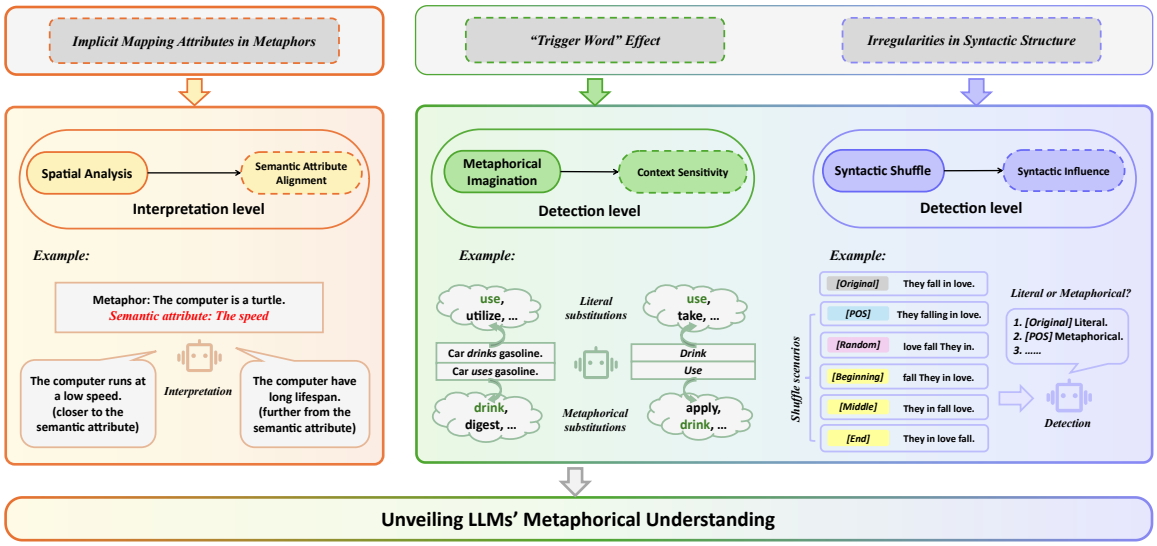

Figure 1: Overview of the experimental framework. Spatial Analysis probes attribute-level semantic alignment in metaphor interpretation. Metaphorical Imagination probes the persistence of stable lexical associations under two contextual settings. Syntactic Shuffle analyzes the influence of syntactic cues on metaphor detection.

capabilities in metaphor generation, particularly with respect to creativity and novelty (DiStefano et al., 2024; Su et al., 2025). In metaphor detection, CoT prompting has been shown to improve performance in multimodal settings (Xu et al., 2024).

Beyond task performance, representation-level analyses suggest that pretrained models encode metaphor-related structure in contextual embeddings (Aghazadeh et al., 2022), and LLMs have demonstrated cross-lingual metaphor detection without explicit fine-tuning (Wachowiak and Gromann, 2023). However, evidence from related non-literal phenomena such as idioms suggests that surface-level distractors continue to pose challenges for cross-lingual semantic alignment (Ye et al., 2026). Taken together, existing work has largely focused on performance outcomes, while representation-level evidence remains indirect, leaving open how LLMs actually process metaphor during inference, and whether task success reflects genuine metaphor understanding or shallow heuristic strategies (Ge et al., 2023).

3 Methodology

We propose a diagnostic framework to examine LLM metaphor processing under controlled probing conditions. Our design is motivated by prior observations, including the limited diagnostic value of answer-based evaluation, trigger word effects, and sensitivity to syntactic irregularities. We construct targeted experiments to characterize LLM behavior at both the interpretation and detection levels. Specifically, there are three complementary dimensions: (1) semantic attribute alignment, (2) context-invariant lexical associations, and (3) syntactic influence. An overview is shown in Figure 1.

3.1 Spatial Analysis

Problem Definition. We propose a geometric probe to characterize semantic attribute alignment in metaphor interpretation, following similarity-based analysis frameworks (Wegmann and Nguyen, 2021). In our setting, each target metaphor sentence m_i is paired with a related metaphor instance m_i' that differs in surface form but shares the same underlying semantic attribute. We denote the LLM-generated interpretation of m_i as M_i.

To construct a local reference region for m_i, we define a **reference plane** γ_i as the affine subspace spanned by $\{R_i, R_i', S_i\}$. Here, R_i and R_i' are human-annotated interpretations that capture the shared semantic attribute, and thus serve as the primary semantic anchors. The third point, S_i, is a model-generated literal paraphrase of the target sentence, providing a complementary literal anchor tied to the same input.

We represent all sentences in a shared embedding space and quantify the alignment of M_i with γ_i by measuring its deviation from the reference plane. Importantly, $\{R_i, R_i', S_i\}$ is used only to construct a local diagnostic reference, rather than a globally calibrated semantic manifold. The resulting geometry is therefore interpreted as a comparative signal of alignment, not as an absolute notion

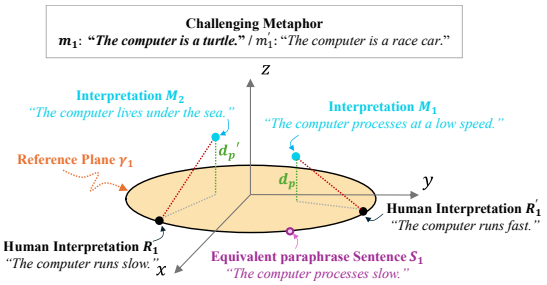

Figure 2: Example of the perpendicular distance d_p in the embedding space. d_p measures the deviation of two LLM-generated interpretations M_1 and M_2 for m_1 from the reference plane γ_1 (defined by $\{R_1, R_1', S_1\}$).

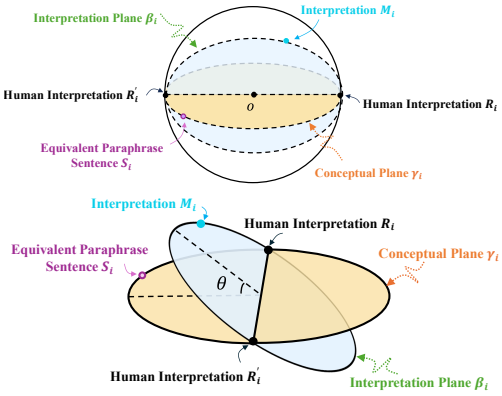

Figure 3: Illustration of the angle θ between the reference plane γ_i (defined by $\{R_i, R_i', S_i\}$) and the interpretation plane β_i (defined by $\{R_i, R_i', M_i\}$) in the embedding space.

of semantic correctness or a recovery of ground-truth semantic structure. All measures are interpreted comparatively across instances and models.

Measures. As illustrated in Figure 2 and Figure 3, we define two complementary measures to quantify alignment:

d_p: the perpendicular distance from M_i to the reference plane γ_i, capturing the magnitude of geometric deviation in the embedding space.

$\cos\theta$: the cosine similarity between the reference plane γ_i and the **interpretation plane** β_i, where β_i is spanned by $\{R_i, R_i', M_i\}$, measuring the orientation difference between the two planes.

Together, d_p and $\cos\theta$ capture complementary aspects of alignment, reflecting both the extent and direction of deviation of model interpretations from the reference semantic region.

3.2 Metaphorical Imagination

Motivated by prior observations of trigger word effects, we examine whether LLMs rely on lexical associations that remain stable across contextual conditions in metaphor processing. We compare two generation settings: *contextualized* generation, in which the target word is produced given its sentence context, and *decontextualized* generation, in which the model is prompted with the target word alone. We operationalize this comparison in two directions. In *Literal-to-Metaphor (LM)*, the model generates a metaphorical counterpart for a literal input, either in isolation or within a literal context. In *Metaphor-to-Literal (ML)*, it generates a literal counterpart for a metaphorical input under the same two settings. Similarity between contextualized and decontextualized outputs is interpreted as a diagnostic signal of context-invariant lexical associations, indicating the persistence of stable lexical anchors in metaphor processing.

3.3 Syntactic Shuffle

Metaphors are often associated with characteristic syntactic patterns (Sullivan, 2013). Because part-of-speech structure and word order encode key syntactic relations, such as argument structure and modifier attachment, perturbing them disrupts compositional structure while largely preserving lexical content. This enables us to test whether metaphor detection relies on integrated sentence structure or on shallower heuristic cues.

We consider three shuffle scenarios. 1) **Random Shuffle**: words are randomly reordered, disrupting both syntactic structure and semantic coherence. 2) **POS Shuffle**: the metaphorical word is replaced with a near-synonymous alternative of a different POS, introducing syntactic irregularity while keeping lexical meaning relatively similar. 3) **Metaphorical Word Reposition**: the metaphorical word is moved to the beginning, a random intermediate position (excluding the original, initial, and final positions), or the end of the sentence, allowing us to assess sensitivity to its positional placement. By comparing detection results across these conditions, we evaluate the extent to which metaphor detection is sensitive to syntactic regularity and positional cues.

4 Experiment

4.1 Datasets

Table 1 summarizes the datasets used in our experiments, each corresponding to a specific probing setting introduced in Section 3. For the spatial analysis experiment, we use Fig-QA, a human-annotated resource designed for Winograd-style

Dataset / Setting	Instances	Example (Metaphor | Literal)
Fig-QA	2.6k	The computer **is a race car**. | The computer runs fast.
		The computer **is a turtle**. | The computer runs slow.
MUNCH (Context)	2.9k	The council **appealed** by case stated ...
		The council **petitioned** by case stated ...
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: Dataset statistics and example instances for Fig-QA (used in Spatial Analysis) and MUNCH (used in Metaphorical Imagination and Syntactic Shuffle). Different MUNCH rows correspond to distinct experimental settings derived from the same set of base instances. Examples are shown as Metaphor | Literal.

metaphorical language understanding (Liu et al., 2022), licensed under the MIT License. Fig-QA organizes instances into sets of four, consisting of two metaphors $\{m_i, m_i'\}$ and their corresponding human-annotated literal interpretations (serving as $\{R_i, R_i'\}$), which reflect the same underlying semantic attribute while differing in surface form. To focus the analysis on metaphor interpretation rather than literal variation, model-generated interpretations are constrained, via span-level annotations, to modify only metaphor-relevant parts of each sentence, with spans identified by GPT-4o (highlighted in bold in Fig-QA examples).

For Metaphorical Imagination and Syntactic Shuffle experiments, we adopt the Metaphor Understanding Challenge Dataset (MUNCH) (Tong et al., 2024), licensed under CC BY 4.0, a linguistically annotated benchmark derived from the VU Amsterdam Metaphor Corpus (Steen et al., 2010). In MUNCH, each sentence contains a metaphor that is challenging for LLMs and is centered around a single annotated metaphorical word. We extract instances from its paraphrase generation task and construct multiple experimental settings from the same set of base sentences. For syntactic shuffle, tokenization and controlled lexical substitutions are performed using WordNet 2020 (McCrae et al., 2020) to introduce systematic perturbations while preserving lexical meanings.

4.2 Models

We evaluate a diverse set of LLMs: 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, DeepSeek-R1-671B (R1-671B) (Guo et al., 2025), and LLaMA-3.1-8B (Grattafiori et al.). For spatial analysis, all model-generated interpretations are encoded using OpenAI's `text-embedding-3-small` so that they can be compared in a shared embedding space. This embedding model is used only for post-hoc geometric analysis and does not influence model generation or task outputs. As a result, geometric comparisons remain consistent across models with different internal representations. Open-source models were run on Google Colab with a T4 GPU. For all generations, we set the temperature to 0 to reduce stochastic variation and improve reproducibility.

4.3 Implementation Details

Spatial Analysis. For each metaphor instance m_i and its LLM-generated interpretation M_i, we construct a reference plane γ_i from $\{R_i, R_i', S_i\}$ and an interpretation plane β_i from $\{R_i, R_i', M_i\}$. We quantify the geometric deviation of M_i in the shared embedding space using two complementary measures: the perpendicular distance d_p from M_i to γ_i, and an angular similarity measure $\cos\theta$ between γ_i and β_i.

To compute these quantities, we derive subspace bases using Singular Value Decomposition (SVD). Because γ_i and β_i are affine planes, we first represent each plane by centered direction vectors with respect to one anchor point, and stack these vectors into a matrix A. We then perform SVD:

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

The top singular vectors define an orthonormal basis for the plane, which is used to compute d_p and $\cos\theta$. Larger d_p or smaller $\cos\theta$ indicates greater deviation from the reference semantic attribute. All

Figure 4: Representative examples of two complementary relationships in Spatial Analysis: left, d_p vs. A_d for V3-671B; right, d_p vs. $\cos\theta$ for Qwen-T.

m_1: **This blanket is as insulating as a wet tissue.**
L'_{11}: The blanket keeps me really cozy.
R_1: The blanket does not keep me warm.
L'_{12}: The blanket makes me feel quite warm.
R'_1: The blanket keeps me very warm.

Accuracy of Multiple-choice Validation			
V3-671B	Qwen-T	GPT-4	GPT-4o
50.89	51.69	50.77	50.92
o3-mini	R1-671B	LLaMA-3.1-8B	
50.85	47.31	46.04	

Table 2: The example and accuracy of multiple-choice validation.

computations are performed in the shared embedding space.

Metaphorical Imagination MUNCH contains metaphors whose meaning is centered around a single metaphorical word, paired with literal substitutions. LLMs are prompted to generate twenty candidate substitutions for the target word under either metaphorical or literal interpretation settings, providing sufficient lexical diversity. To assess the persistence of lexical associations across contextual conditions, we compare contextualized and decontextualized generations using an **Anchor Score**. Specifically, when shared words occur between comparative sets, the Anchor Score is set to 1, indicating a shared lexical choice across contexts (a potential lexical anchor). If no word is shared between the two sets, we define Anchor scores by computing the maximum cosine similarity between words across the two sets using 300-dimensional GloVe embeddings (Pennington et al., 2014). We further analyze Anchor Scores across annotated discourse genres to examine whether such lexical invariance varies across discourse types.

5 Results & Analysis

The complete experimental results and model-specific analyses are reported in the Appendix, including the prompts, detailed per-model results, additional breakdowns across conditions.

5.1 Semantic Attribute Alignment

We contrast geometric probing with two evaluation metrics: a similarity-based signal and a discrete answer-based signal.

Geometric Signal Characterization. As a similarity-based evaluation signal, we define A_d as the sum of cosine similarities between a model-generated interpretation M_i and two reference interpretations $\{R_i, R'_i\}$. Interpretations that better align with the intended semantic attribute are expected to exhibit higher A_d and smaller deviation from the reference plane (d_p). The magnitude of d_p does not have a calibrated absolute interpretation; instead, we treat it as a comparative diagnostic signal. As a reference distribution, d_p values for o3-mini generations have median 0.102, interquartile range [0.029, 0.298], and 95th percentile 0.729. We interpret d_p relative to empirical distribution rather than assigning thresholded semantic meaning.

In practice, as shown in Figure 4, lower A_d is associated with larger d_p, while smaller d_p corresponds to larger $\cos\theta$. These trends are supported by Spearman correlations between d_p and A_d ($\rho = -0.62$) and between $\cos\theta$ and d_p ($\rho = -0.64$), indicating that d_p and $\cos\theta$ capture coherent geometric signals. Distributions across all models are provided in Appendix E. To ensure that these correlations are not artifacts of marginal distributions, we conduct permutation tests that break instance-level pairing; under permutation, correlations collapse to near-zero values, confirming that the observed relationships depend on meaningful alignment rather than spurious structure.

Limitations of Discrete Evaluation. To contrast with discrete evaluation signals, we consider a multiple-choice interpretation setup in which can-

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

Table 3: Average d_p and $\cos\theta$ across models. Mean (M) and standard deviation (SD) are reported. Bold and underlined values indicate the lowest and second-lowest geometric deviation, respectively.

R_1	R_1'	*Metaphor*	S	M_i	d_p	$\cos\theta$
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 4: Example interpretations illustrating attribute-level semantic alignment. Interpretations with lower alignment to the reference semantic attribute region, indicated by higher d_p and lower $\cos\theta$, are shown in *italic*.

didate interpretations are restricted to an attribute-aligned option set. For a metaphor pair $\{m_i, m_i'\}$ in Fig-QA, R_i denotes the correct interpretation of m_i, while R_i' corresponds to m_i', which differs in surface form but shares the same underlying attribute. We generate two paraphrastic variants $\{L_{i1}', L_{i2}'\}$ of R_i', yielding a four-way option set $\{R_i, R_i', L_{i1}', L_{i2}'\}$, and ask models to select the correct interpretation of m_i. The generated paraphrases are used to construct semantically close alternatives around the reference interpretations, so that the candidate options remain attribute-aligned and reduce reliance on superficial lexical cues.

Despite this controlled design, models exhibit near-chance accuracy when alternatives differ only in fine-grained polarity or intensity (Table 2, e.g., "does not keep me warm" vs. "keeps me very warm"). This suggests that even under attribute-aligned candidate sets, discrete evaluation provides limited visibility into model behavior, as it does not reveal how far a generated interpretation deviates from the intended attribute.

Main results. Spatial Analysis provides a complementary geometry-based view. Because the reference plane includes the model-generated literal anchor S_i, we verify it remains semantically close to both human references $\{R_i, R_i'\}$, supporting its role as a local anchor rather than an off-topic artifact (Appendix F). Table 3 reports aggregate results: GPT-4o achieves the lowest mean d_p, while V3-671B shows the highest mean $\cos\theta$. Across models, interpretations still deviate from the intended semantic attribute, suggesting systematic drift. All experiments rely on a shared embedding model, without task-specific fine-tuning.

To further validate that d_p reflects meaningful semantic differences, we conduct a small-scale human evaluation, evaluated by one senior master student with expertise in the relevant languages. Each sample consists of a metaphor and a model-generated interpretation, which annotators rate on a 3-point scale (incorrect / partial / correct) based on whether the intended metaphorical meaning is captured. We sample instances from both low-d_p and high-d_p regions to contrast extreme cases. Low-d_p interpretations receive substantially higher human alignment scores than high-d_p ones (mean 1.96 vs. 0.84; $\Delta = 1.12$), indicating a clear separation between the two groups and supporting the semantic relevance of the geometric signal (see Appendix G).

We also present representative cases in Table 4. For instance, in the monk/lawyer example, the intended attribute is social honor or respectability; the interpretation "the privilege of legal representation" instead shifts toward legal entitlement and does not preserve the intended mapping. Overall, conventional metrics provide only coarse-grained views of alignment, whereas geometric measures reveal finer-grained structure in semantic deviation.

	V3-671B	Qwen-T	GPT-4	GPT-4o	o3-mini	R1-671B	LLaMA-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 5: Anchor Scores for Metaphorical Imagination under LM and ML settings. The four genres include *News*, *Fiction*, *Academic*, and *Conversation*. Best values are shown in bold, and second-best values are underlined.

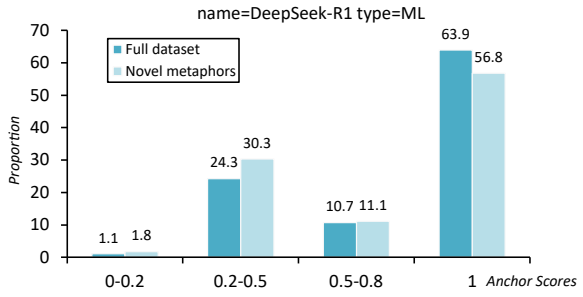

Figure 5: Distribution of ML Anchor Scores under R1-671B for the full dataset and the subset of novel metaphors (novelty score > 0.3).

5.2 Lexical Invariance

Beyond semantic alignment, we next examine whether the lexical content associated with a metaphor remains stable across contextual conditions, or instead shifts more substantially with context. Table 5 summarizes results for Metaphorical Imagination. Anchor Scores between contextualized and decontextualized generations are consistently high (approximately 65%–80%), suggesting that lexical anchors often persist across metaphor–literal substitution settings. Among the evaluated models, GPT-4, o3-mini, and R1-671B achieve relatively higher scores. Meanwhile, Metaphor-to-Literal (ML) consistently yields higher scores than Literal-to-Metaphor (LM), aligning with prior observations that mapping metaphorical expressions to literal paraphrases is generally more constrained than the reverse (Liu et al., 2022). Genre-level results show stable within-model patterns across LM and ML settings, with o3-mini and R1-671B exhibiting more consistent behavior on conversational metaphors, and GPT-4 showing higher Anchor Scores on news metaphors. This variation may reflect differences in conventionalization and contextual dependence across genres; for example, conversational metaphors may be more conventionalized and compatible with stable lexical priors.

To examine whether this pattern also appears for metaphors that may depend more strongly on context, we analyze MUNCH items with novelty scores > 0.3 (Tong et al., 2024). The novelty score ranges from 0 to 1, with higher values indicating greater novelty. Figure 5 presents the distribution of ML Anchor Scores under R1-671B for both the full dataset and the subset of more novel metaphors. Although overall scores decrease relative to the full dataset, more than 50% of cases still reach an Anchor Score of 1, while the remaining instances concentrate around 0.2–0.5. This pattern suggests that persistent lexical anchoring remains common even for novel metaphors, but is less uniformly expressed across items. Complete Anchor Score distributions across models are shown in Appendix H.

High Anchor Scores should therefore be interpreted cautiously: lexical priors and contextual evidence may sometimes point in the same direction, so strong overlap does not by itself imply that context is ignored. We interpret these results more narrowly as evidence that lexical associations can remain stable across contextual conditions, rather than as proof of general context-insensitivity. At the same time, this pattern does not necessarily translate into reliable metaphor detection performance, as discussed in Section 5.3. Stable lexical anchoring may support familiar metaphors by providing readily accessible associations; however, in richer contexts or more novel metaphors, it may also bias interpretations toward highly associated lexical cues, contributing to trigger word effects.

	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
LLaMA-3.1-8B	53.36	50.33	**53.81**	51.75	53.08	53.67

Table 6: Metaphor detection accuracy under Syntactic Shuffle perturbations. Highest values are shown in bold.

5.3 Syntactic Influence

We finally examine how controlled syntactic perturbations affect metaphor detection. Across perturbation settings, detection accuracy varies substantially across models, as shown in Table 6. LLaMA-3.1-8B remains relatively stable around chance level (approximately 50%), showing comparatively little variation across perturbation types. Combined with its low Anchor Scores, this pattern suggests limited responsiveness to the targeted probes on MUNCH. Other models show clearer variation across perturbation settings. In particular, models achieve higher accuracy under POS shuffle than on the original sentences. Because POS shuffle preserves much of the local lexical content while introducing syntactic irregularity, it creates inputs that may amplify cues similar to the kinds of anomalous combinations highlighted in SPV-style accounts of metaphor. By contrast, random shuffle disrupts both syntactic structure and sentence-level coherence, producing more heterogeneous and less interpretable effects across models.

Detection accuracy is relatively stable across positional perturbations (beginning/middle/end), suggesting that the absolute position of a metaphorical word has a weaker effect than the type of structural disruption applied to the sentence. Models therefore appear more responsive to local irregularity, such as POS shuffle, than to changes in word position alone. For example, V3-671B underperforms most other models (except LLaMA-3.1-8B) and even exceeds its original-sentence accuracy under random shuffle, suggesting comparatively unstable behavior under extreme perturbation. We therefore treat random shuffle primarily as a stress-test condition rather than as evidence about natural metaphor processing. Setting random shuffle aside, several syntactic perturbations lead to improved detection accuracy across models, suggesting that syntactic irregularity itself can function as a useful cue for metaphor detection without necessarily implying better metaphor understanding.

Overall, metaphor detection varies across both models and perturbation settings, highlighting the diagnostic value of syntactic manipulation on MUNCH. Taken together with the preceding experiments, these results suggest that metaphor-related behavior in current LLMs is sensitive not only to semantic and lexical factors, but also to surface structural disruption, while providing limited evidence for robust sentence-level syntactic integration under perturbation.

6 Conclusion

This work examines LLM metaphor processing from complementary perspectives: semantic attribute alignment, context-invariant lexical associations, and syntactic influence. Spatial Analysis reveals consistent attribute-level deviation in model-generated interpretations. Metaphorical Imagination shows substantial overlap between contextualized and decontextualized generations, suggesting models may rely on stable metaphor and literal lexical associations that support familiar metaphors while biasing decisions toward highly associated cues. Syntactic Shuffle suggests that models respond to syntactic irregularities as heuristic signals for metaphor detection. Overall, although LLMs achieve strong performance on metaphor processing tasks, this performance may arise from a heterogeneous combination of lexical associations and heuristic cues rather than robust semantic understanding. These signals may support behaviors in conventional metaphors but can constrain context-sensitive integration of semantic attributes and syntactic structure. More broadly, Our results highlight the need for evaluation and modeling approaches that probe attribute-level alignment, contextual reasoning, and robust syntactic integration beyond pattern-based cues.

Limitations

Our analysis is subject to several limitations. First, Spatial Analysis relies on a constructed reference semantic attribute region derived from human interpretations and LLM-generated sentences. While this provides a behavioral proxy for analyzing model outputs, it does not directly capture underlying cognitive representations of semantic attributes. In addition, the two geometric measures d_p and $\cos\theta$ depend on the embedding space used for analysis, different embedding choices may affect absolute distances. Moreover, we introduce a third reference sentence S_i to enable a locally structured geometric comparison. This formulation represents a pragmatic design choice that balances interpretability and representational richness, and we do not claim that the resulting dimensionality is optimal.

Second, the Metaphorical Imagination probe focuses on metaphors instantiated by a single annotated word, and therefore primarily captures word-level metaphor-literal associations. Our findings do not directly generalize to multi-word or discourse-level metaphors. Understanding such phenomena likely requires different experimental designs and remains an important direction for future work.

Furthermore, in Syntactic Shuffle, random shuffling is not intended to model natural language use, but rather to function as an extreme stress test. We do not interpret results under this condition as reflecting natural metaphor processing. Instead, controlled POS and positional perturbations provide interpretable evidence about sensitivity to syntactic irregularity. The observation that metaphor detection accuracy can increase under such conditions suggests that detection behavior may, rely on lexical or heuristic cues independent of sentence-level semantic and syntactic integration.

Finally, our experiments are conducted on English metaphor datasets, and the generality of our findings to other languages or culturally specific metaphors remains to be explored.

Ethics Statement

This work uses two publicly available datasets: FigQA (Liu et al., 2022) and MUNCH (Tong et al., 2024). These datasets are used solely for probing experiments on LLMs. The experiments strictly excluded any materials associated with personal identifiers or sensitive data categories.

Acknowledgments

This work was supported in part by the Science and Technology Development Fund of Macau SAR (Grant Nos. FDCT/0007/2024/AKP, EF2024-00185-FST), the UM and UMDF (Grant Nos. MYRG-GRG2024-00165-FST-UMDF, MYRG-GRG2025-00236-FST), the Tencent AI Lab Rhino-Bird Research Program (Grant No. EF2023-00151-FST), the Stanley Ho Medical Development Foundation (Grant No. SHMDF-AI/2026/001), and the National Natural Science Foundation of China (Grant No. 62266013).

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.

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.

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.

Jie Wang, Jin Wang, and Xuejie Zhang. 2024a. 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.

Shanshan Wang, Derek Wong, Jingming Yao, and Lidia Chao. 2024b. What is the best way for ChatGPT to translate poetry? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14025–14043.

Shanshan Wang, Junchao Wu, Fengying Ye, Derek F. Wong, Jingming Yao, and Lidia S. Chao. 2025. Benchmarking the detection of LLMs-generated Modern Chinese poetry. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9533–9552. 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.

Fengying Ye, Yanming Sun, Runzhe Zhan, Lidia S. Chao, Zheqi Zhang, and Derek F. Wong. 2026. G-idiomalign: A gloss-pivoted benchmark for cross-lingual idiom alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, San Diego, California. Association for Computational Linguistics.

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.

A Definitions and Terminology

To support the motivation and experimental design of this study, we introduce several key definitions and terms. Table 7 provides a consolidated list of the terms used throughout the paper and their definitions.

B Prompts for Spatial Analysis

Figure 6 shows the two prompts used in the Spatial Analysis experiment. Prompt 1 instructs the model to generate a literal sentence S that is a semantic paraphrase of the human-annotated interpretations. Prompt 2 instructs the model to generate a literal interpretation M_i of a given metaphor m_i by replacing the metaphorical expression with a literal alternative.

C Prompts for Syntactic Shuffle

Figure 8 presents the prompt used for metaphor detection under different syntactic perturbations in the Syntactic Shuffle experiment.

D Prompts for Metaphorical Imagination

Figure 7 presents the prompts used for the Metaphorical Imagination experiment. The prompts require LLMs to perform two complementary tasks: (1) *Metaphor-to-Literal* (ML), in which literal words are generated from metaphorical inputs, and (2) *Literal-to-Metaphor* (LM), in which metaphorical words are generated from literal inputs. Both tasks are conducted under contextualized and decontextualized settings. Accordingly, the prompts are divided into two types depending on whether sentence context is provided.

E Distributions of $(d_p, \cos\theta)$ and (d_p, A_d)

Figures 9–10 present the distributions of the distance d_p (between the model-generated interpretation M_i and the reference plane γ_i), plotted against the cosine similarity $\cos\theta$ (between the interpretation plane β_i and the reference plane γ_i), as well as against the auxiliary similarity measure A_d.

F Sanity Check for the Literal Anchor S_i

To assess whether the model-generated literal anchor S_i introduces unintended semantic bias, we analyze its alignment with the human reference interpretations $\{R_i, R_i'\}$ in the shared embedding space.

Alignment with human anchors. We compute cosine similarities between S_i and each of the human references:

$$\cos(S_i, R_i), \quad \cos(S_i, R_i').$$

Across all instances (reported here for o3-mini), we observe that S_i remains consistently aligned with both anchors. Specifically, $\cos(S_i, R_i)$ has a median of 0.753 (IQR: [0.662, 0.829]; 5–95% range: [0.513, 0.922]), while $\cos(S_i, R_i')$ has a median of 0.686 (IQR: [0.600, 0.767]; 5–95% range: [0.467, 0.873]). For reference, the similarity between the two human interpretations $\cos(R_i, R_i')$ has a median of 0.712 (IQR: [0.626, 0.779]).

These results indicate that S_i is typically located within the same semantic neighborhood as the human anchors, rather than drifting toward unrelated meanings.

Balance between the two anchors. To evaluate whether S_i is disproportionately closer to one reference than the other, we measure the absolute difference:

$$|\Delta| = |\cos(S_i, R_i) - \cos(S_i, R_i')|.$$

We find that this imbalance is generally moderate, with a median of 0.140 (IQR: [0.086, 0.192]) and a 95th percentile of 0.300. This suggests that S_i does not strongly bias the reference plane toward either anchor in most cases.

Interpretation. Overall, these diagnostics support the use of S_i as a stable literal anchor for constructing the local reference plane. While S_i is model-generated, it typically remains semantically aligned with the intended attribute encoded by $\{R_i, R_i'\}$, and does not introduce systematic off-topic artifacts. Accordingly, the resulting geometric measures are best interpreted as capturing relative semantic deviation within a locally consistent reference region.

G Human Evaluation of Semantic Alignment

To assess whether the geometric deviation measure d_p reflects meaningful semantic differences in

metaphor interpretation, we conduct a small-scale human evaluation.

Setup. Each evaluation instance consists of (i) an English metaphorical sentence and (ii) a corresponding model-generated interpretation. The annotator is asked to judge whether the interpretation captures the intended metaphorical meaning based solely on the given text, without external context.

We use a 3-point rubric:

- **2 (Correct):** The interpretation captures the intended abstract meaning or mapping.
- **1 (Partial):** The interpretation is related but underspecified, ambiguous, or only partially aligned.
- **0 (Incorrect):** The interpretation is off-topic, incorrect, or reduces the metaphor to a literal or unrelated description.

Sampling. To test whether d_p meaningfully separates aligned and misaligned interpretations, we sort instances by d_p and sample from both extremes. Specifically, we select 25 instances from the lowest-d_p region (bottom-200) and 25 instances from the highest-d_p region (top-200), based on o3-mini generations.

Results. We observe a clear separation between the two groups. For high-d_p instances, the label distribution is $11/7/7$ for scores $0/1/2$, with a mean score of 0.84. In contrast, low-d_p instances yield $0/1/24$, with a mean score of 1.96. The difference between the two groups is substantial ($\Delta = 1.12$).

To assess statistical reliability, we perform bootstrap resampling and obtain a 95% confidence interval of $[0.76, 1.44]$ for the mean difference. A Mann–Whitney U test further indicates that the difference is statistically significant ($p \approx 1.05 \times 10^{-6}$).

Interpretation. These results provide independent evidence that lower geometric deviation (d_p) corresponds to better semantic alignment as judged by humans. While the evaluation is limited in scale, it supports the interpretation of d_p as a meaningful proxy for relative semantic alignment in metaphor interpretation.

H Anchor Score Distributions in Metaphorical Imagination

In the Metaphorical Imagination experiment, we analyze the distribution of *Anchor Scores*, which quantify the degree of lexical overlap between contextualized and decontextualized generation sets for the same target word. Figures 11–20 report Anchor Score distributions across different tasks (Metaphor-to-Literal and Literal-to-Metaphor) and discourse genres. Higher Anchor Scores indicate the presence of stable lexical associations that persist across contextual conditions, providing empirical evidence for lexical invariance in LLM behavior. Such associations may support certain types of metaphor processing, although their interaction with context-sensitive inference remains an open question. Future work may further investigate how these lexical associations can be more explicitly characterized.

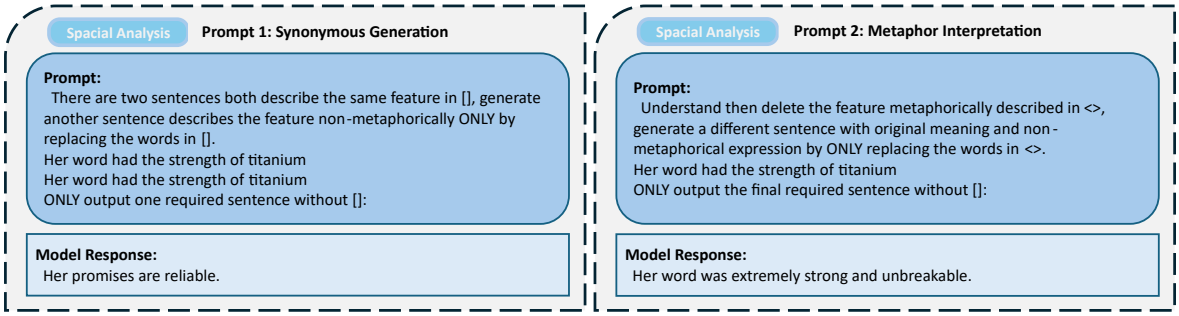

Figure 6: The prompts of Spatial Analysis.

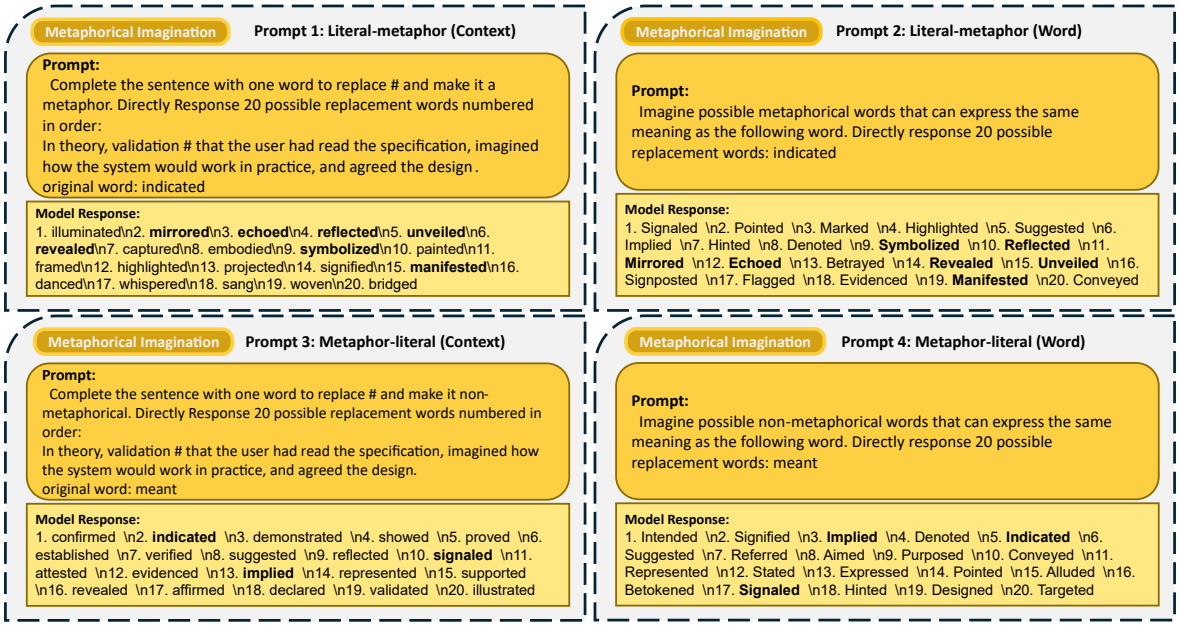

Figure 7: The prompts of Metaphorical Imagination.

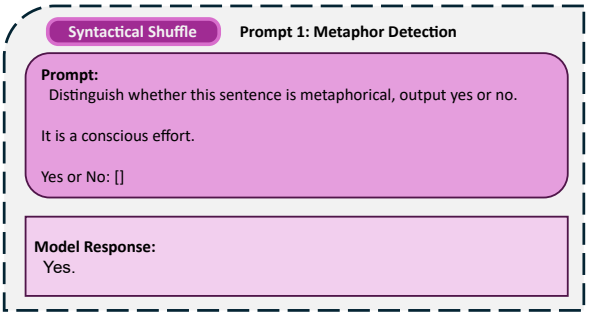

Figure 8: The prompt of Syntactic Shuffle.

Terminology	Definition	Notes/Examples
Selection Preference Violation (SPV)	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)	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)	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
Semantic Attribute	The salient property or relational feature that is selectively mapped from the source domain to the target domain in a metaphor.	Metaphor: The computer is a turtle. Semantic attribute: The speed (Slowness)
Reference Plane γ_i	Constructed by the embeddings of three sentences R_1, R_1' and S_i, implying target semantic attribute. Representing the ideal attribute the metaphor intended to convey.	R_1: The computer runs fast. R_1': The computer runs slow. S_1: The computer processes fast.
Interpretation Plane β_i	The plane defined by LLM-generated interpretation M_i and the two human-annotated interpretations R_i and R_i' to evalutate the diviation of interpertations.	R_1: The computer runs fast. R_1': The computer runs slow. M_1: The computer runs at a high speed.
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.	Metaphors with the word *arm* may falsely activate war-related interpretations due to frequent lexical co-occurrence, even when the context does not support such mappings.
Lexical Invariance	The tendency of a model to produce the same or highly similar lexical realizations for a given word or concept regardless of whether it is presented in isolation or embedded within a sentential context	Even when the surrounding context does not support such a mapping, models tend to consistently associate metaphors containing the word *arm* with *war*.
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 unrelated words without meaningful 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 metaphorical word is rearranged to the beginning, a random intermediate location, or the end of the sentence.	**appealed** The council by case stated. The council by case **appealed** stated. The council by case stated **appealed**.

Table 7: Terminology and definitions in this study.

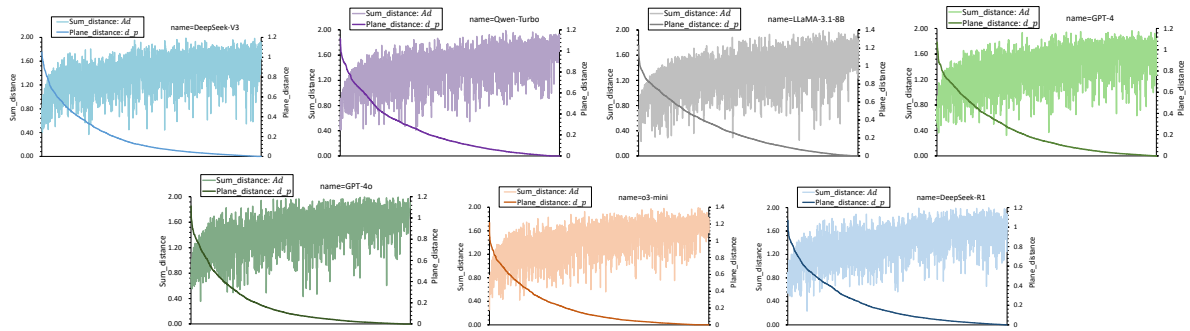

Figure 9: The (d_p, Ad) distribution of every model. Sort d_p in decreasing order. Significant fluctuation can be observed due to the variance in the non-metaphorical part of sentences.

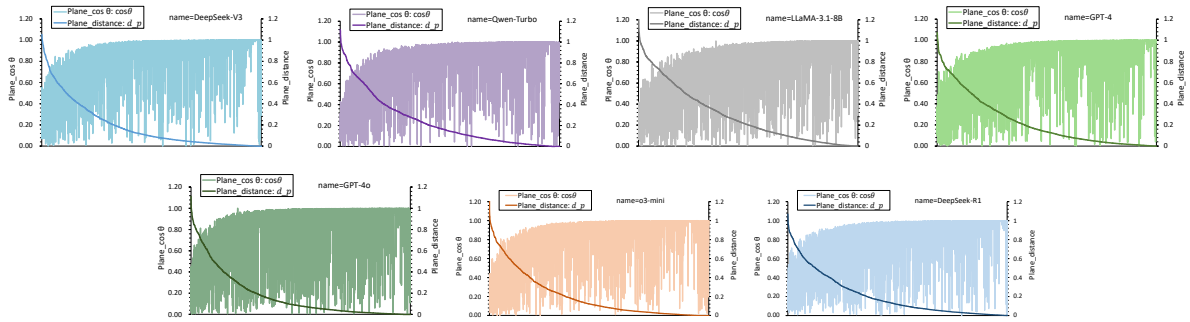

Figure 10: The (d_p, $\cos\theta$) distribution of every model. Sort d_p in decreasing order. Significant fluctuation can be observed due to the variance in the non-metaphorical part of sentences.

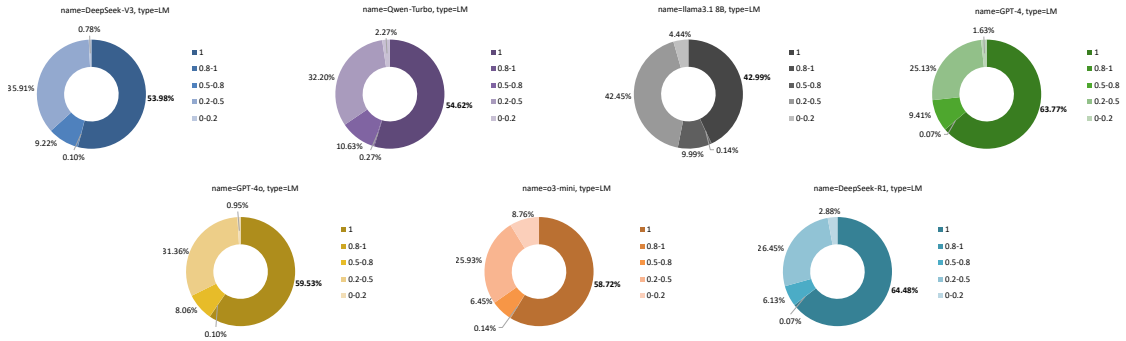

Figure 11: The Anchor Scores distributions of literal-metaphor (LM) word imagination task on every model (the largest portion is in bold and the second largest is underlined).

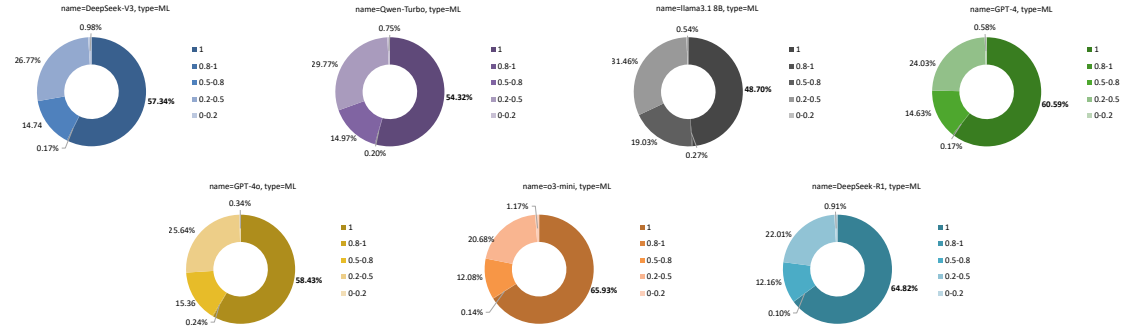

Figure 12: The Anchor Scores distributions of metaphor-literal (ML) word imagination task on every model (the largest portion is in bold and the second largest is underlined).

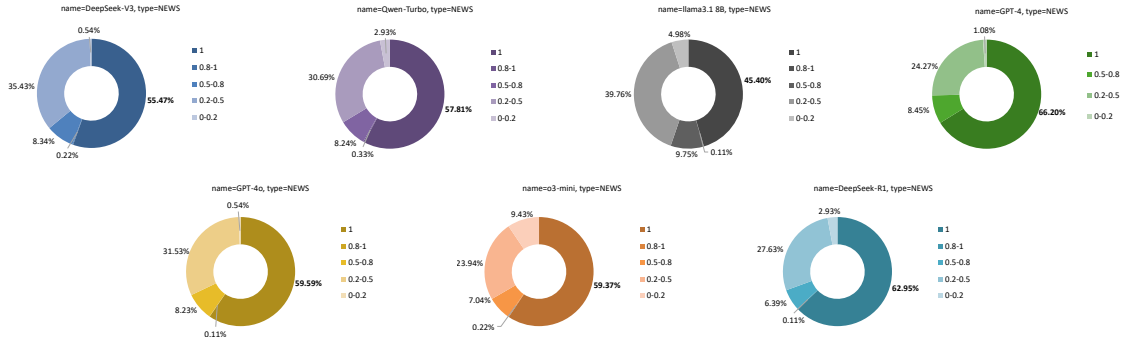

Figure 13: The Anchor Scores distributions of literal-metaphor (LM) word imagination task with sentences in NEWS on every model (the largest portion is in bold and the second largest is underlined).

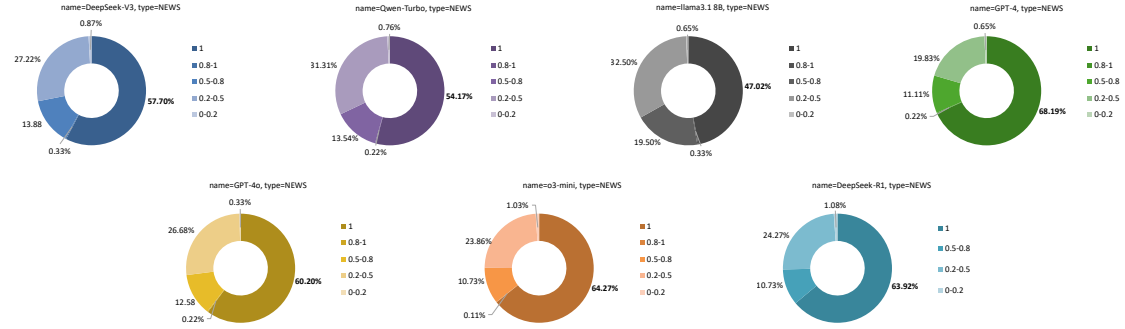

Figure 14: The Anchor Scores distributions of metaphor-literal (ML) word imagination task with sentences in NEWS on every model (the largest portion is in bold and the second largest is underlined).

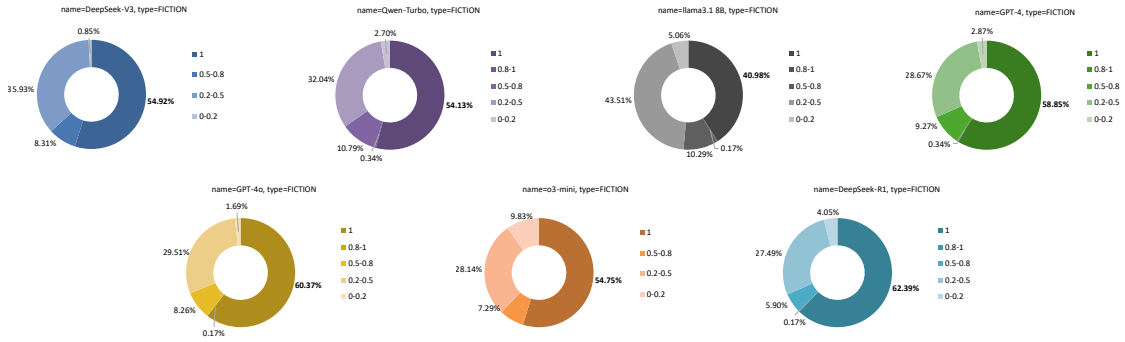

Figure 15: The Anchor Scores distributions of literal-metaphor (LM) word imagination task with sentences in FICTION on every model (the largest portion is in bold and the second largest is underlined).

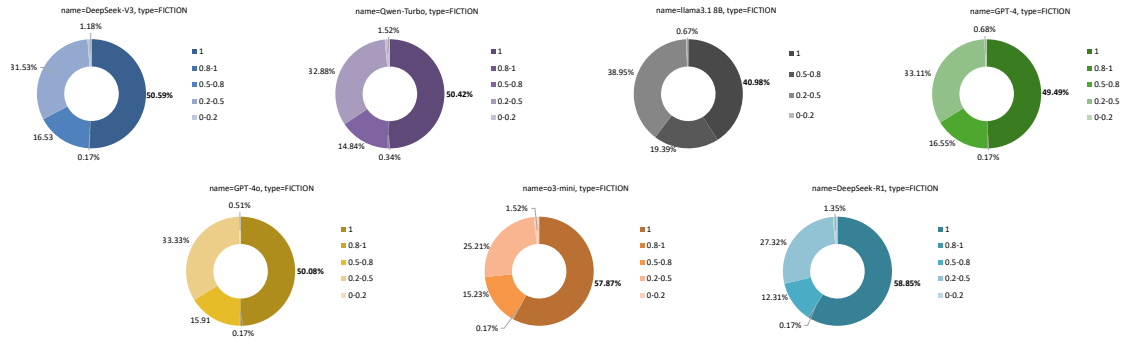

Figure 16: The Anchor Scores distributions of metaphor-literal (ML) word imagination task with sentences in FICTION on every model (the largest portion is in bold and the second largest is underlined).

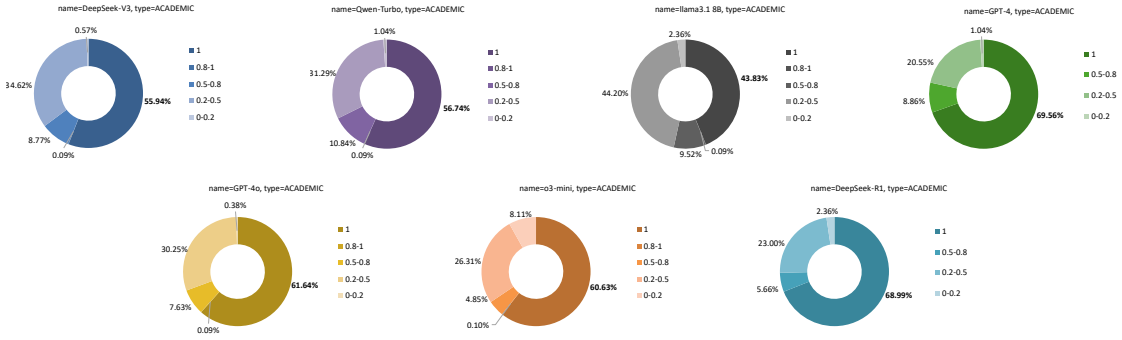

Figure 17: The Anchor Scores distributions of literal-metaphor (LM) word imagination task with sentences in ACADEMIC on every model (the largest portion is in bold and the second largest is underlined).

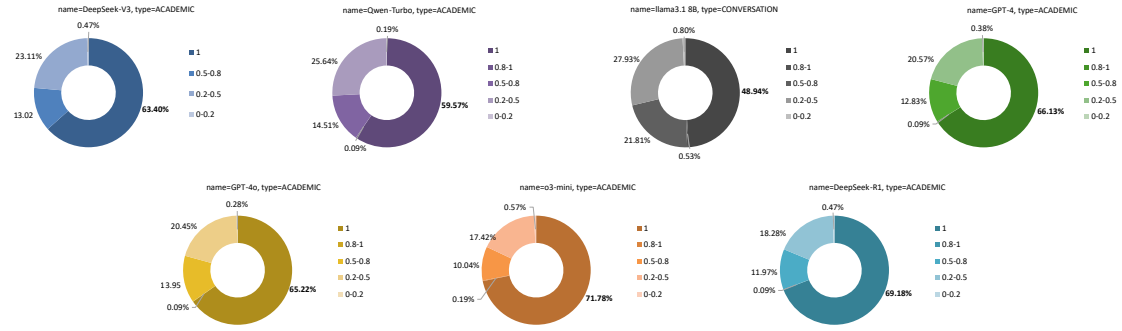

Figure 18: The Anchor Scores distributions of metaphor-literal (ML) word imagination task with sentences in ACADEMIC on every model (the largest portion is in bold and the second largest is underlined).

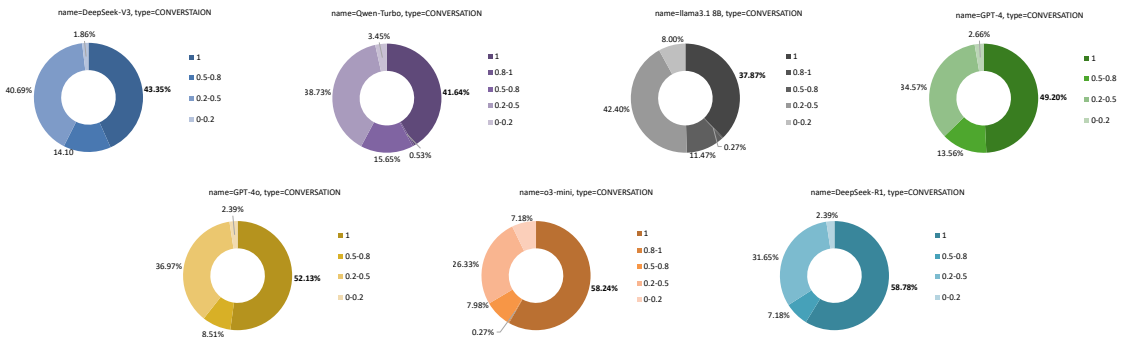

Figure 19: The Anchor Scores distributions of literal-metaphor (LM) word imagination task with sentences in CONVERSATION on every model (the largest portion is in bold and the second largest is underlined).

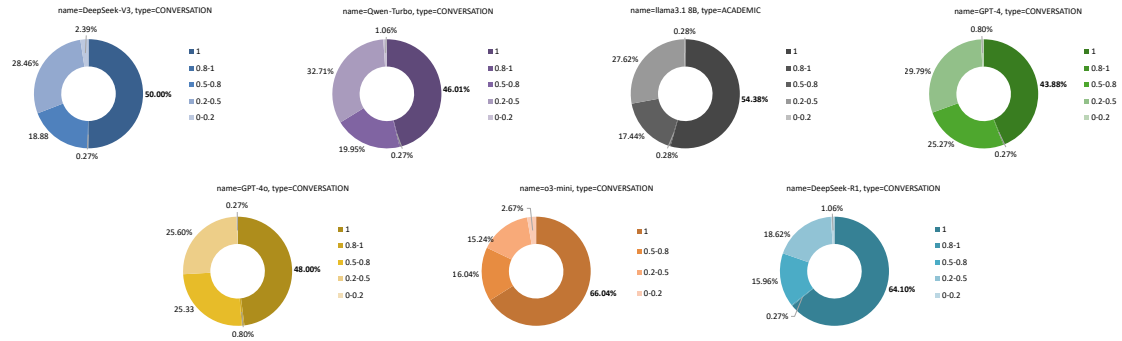

Figure 20: The Anchor Scores distributions of metaphor-literal (ML) word imagination task with sentences in CONVERSATION on every model (the largest portion is in bold and the second largest is underlined).