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

Anonymous ACL submission

## Abstract

Large language models (LLMs) achieve strong performance on metaphor detection and interpretation tasks, yet it remains unclear what such behavioral success actually 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 from 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). However, prior work has also identified behavioral patterns that complicate the 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 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.

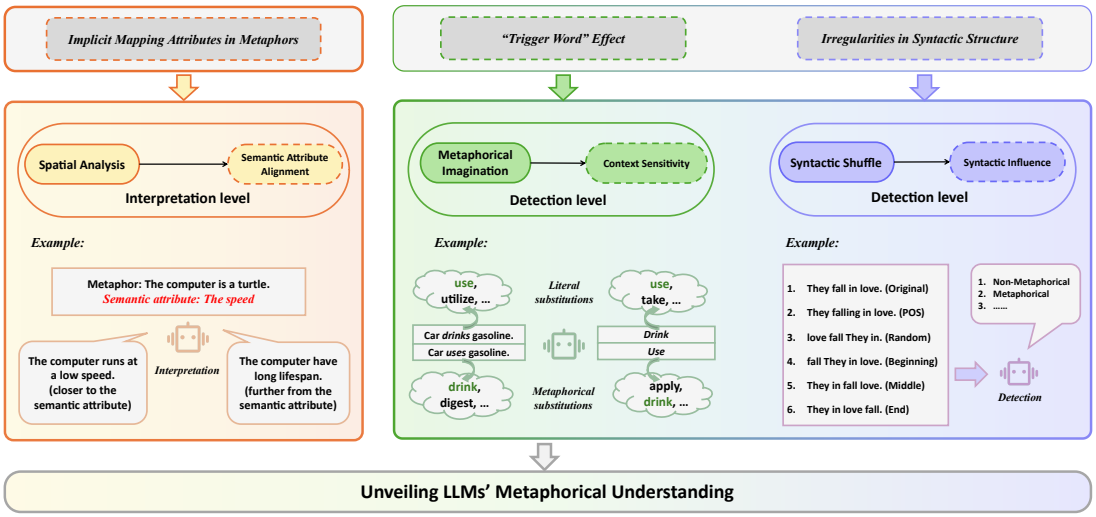


Figure 1: Overview of the experimental framework. Spatial Analysis probes attribute-level semantic alignment in metaphor interpretation. Metaphorical Imagination probes the persistence of stable lexical associations under two contextual settings. Syntactic Shuffle analyzes the influence of syntactic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

We investigate LLM metaphor processing from a diagnostic perspective through three complementary questions: (1) whether generated metaphor interpretations exhibit semantic attribute alignment with reference interpretations; (2) whether LLMs exhibit context-invariant lexical associations across different context conditions; and (3) how syntactic disruption influences metaphor detection. 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 examine the 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.

Terminologies and definitions used in this paper are provided in Appendix 8.

## 2 Related Work

### 2.1 Metaphor Detection

Metaphor detection aims to identify whether a given input contains metaphorical expressions. Early work relied on rule-based linguistic frameworks (Group, 2007; Dodge et al., 2015), while subsequent neural models reformulated the task as supervised classification (Rai and Chakraverty, 2020). To better capture metaphorical usage, researchers incorporated linguistic and conceptual frameworks, including word association statistics (e.g., *nurse* and *doctor*) (Wan et al., 2020; Church and Hanks, 1990). Metaphorical expressions vary across languages, genre, and socio-cultural context, motivating studies on domain-specific metaphor detection (Montefinese et al., 2014; Brysbaert and New, 2009; Cheung et al., 2009; Janschewitz, 2008). Recent work has also detected metaphors in the context of LLMs, reporting strong performance (Mao et al., 2024; Ge et al., 2022; Choi et al., 2021). However, it remains unclear what such detection performance reveals about underlying metaphor processing behavior.

### 2.2 Metaphor Interpretation

Metaphor interpretation focuses on uncovering metaphorical mappings and their associated conceptual domains. Linguistic analyses highlight that metaphors rely on systematic cross-domain mappings rather than isolated lexical substitutions (Sullivan, 2013). In computational settings, a common approach to metaphor interpretation is metaphor component recognition, which aims to identify the source and target domains underlying metaphorical expressions (Sengupta et al., 2024; Ge et al., 2022). Prior work has also reported trigger word effects in this setting (Wachowiak and Gromann,

2023). Other approaches incorporate explicit reasoning mechanisms to guide metaphor interpretation. Frameworks combining Chain-of-Thought reasoning with external knowledge resources have been proposed to guide models toward more structured interpretations (Tian et al., 2024). Additionally, some studies focus on explaining metaphors through literal paraphrases, often drawing on SPV as a learning signal (Mao et al., 2024). While prior work primarily evaluates metaphor interpretation through task-level performance, our work adopts a diagnostic perspective to analyze what such performance reveals about LLM behavior.

### 2.3 Metaphor in LLMs

Recent work has increasingly explored metaphor processing with LLMs. Wang et al. (2024) proposed a multi-stage prompt-based framework to incorporate conceptual background for Chinese metaphor interpretation, while other work examined LLMs' associative capabilities in metaphor generation, particularly with respect to creativity and novelty (DiStefano et al., 2024; Su et al., 2025). In metaphor detection, Chain-of-Thought prompting has been applied to improve LLM performance in multimodal scenarios (Xu et al., 2024). At the representation level, Aghazadeh et al. (2022) showed that pretrained models exhibit metaphor-related structure in contextual embeddings. Moreover, LLMs have demonstrated the ability for cross-lingual metaphor detection without explicit fine-tuning (Wachowiak and Gromann, 2023). However, as Ge et al. (2023) noted, existing studies predominantly report task-level accuracy, motivating complementary diagnostic analyses that examine metaphor processing behavior.

## 3 Methodology

We adopt a diagnostic perspective to examine how LLMs process metaphors under targeted probing conditions. Motivated by recurring observations in prior works: the limited diagnostic value of answer-based evaluation, trigger word effects, and sensitivity to syntactic irregularities, we design probing experiments to analyze LLM behavior at both interpretation and detection levels. Specifically, we examine three complementary dimensions: (1) semantic attribute alignment, (2) context-invariant lexical associations, and (3) syntactic influence, each corresponding to an experimental setting. An overview of the framework is shown in Figure 1.

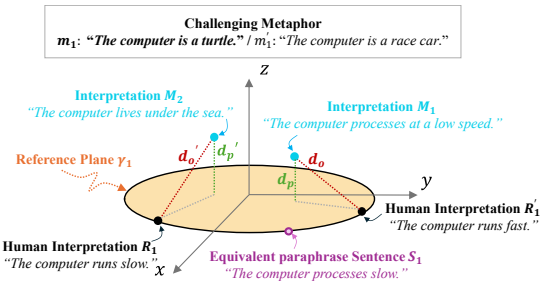


Figure 2: Example of the perpendicular distance $d_p$ in the embedding space. $d_p$ measures the deviation of two LLM-generated interpretations $M_1$ and $M_2$ for $m_1$ from the reference plane $\gamma_1$ (defined by $\{R_1, R_1', S_1\}$).

### 3.1 Spatial Analysis

**Problem Definition.** Following similarity-based analysis frameworks (Wegmann and Nguyen, 2021), we propose a geometric probe to approximate semantic attribute alignment in metaphor interpretation. In our setting, each target metaphor sentence $m_i$ is paired with another metaphor $m_i'$ that differs in surface meaning but shares the same semantic attribute. We denote an LLM-generated interpretation of $m_i$ as $M_i$. To approximate a reference semantic attribute region for $m_i$, we construct a **reference plane** $\gamma_i$ as an affine subspace (hereafter referred to as a plane) defined by $\{R_i, R_i', S_i\}$. $\{R_i, R_i'\}$ are the human-annotated interpretations of $\{m_i, m_i'\}$, and $S_i$ is an LLM-generated literal paraphrase used as a comparable semantic baseline anchor, providing a comparable semantic baseline. We represent all sentences in a shared embedding space and quantify how closely $M_i$ aligns with $\gamma_i$ by measuring its deviation from this plane. All measures are interpreted comparatively across instances and models.

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

$d_p$: the perpendicular distance from $M_i$ to the reference plane $\gamma_i$, used as an indicator of geometric deviation in the embedding space.

$\cos\theta$: the cosine similarity between $\gamma_i$ and an **interpretation plane** $\beta_i$ defined by $\{R_i, R_i', M_i\}$ (an affine subspace just like $\gamma_i$), capturing how the geometric orientation defined by $M_i$ diverges.

Concretely, $d_p$ and $\cos\theta$ provide geometric signals that reflect how model interpretations relate to the constructed reference semantic attribute region.

### 3.2 Metaphorical Imagination

Motivated by prior observations of trigger word effects in LLM metaphor processing, we investi-

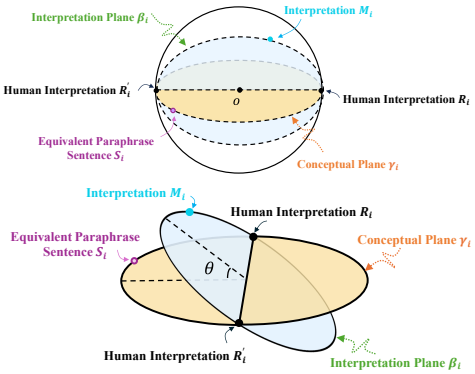


Figure 3: Illustration of the angle $\theta$ between the reference plane $\gamma_i$ (defined by $\{R_i, R_i', S_i\}$) and the interpretation plane $\beta_i$ (defined by $\{R_i, R_i', M_i\}$) in the embedding space.

gate whether LLMs encode stable lexical associations between words and their metaphorical or literal uses that persist across contextual conditions. We compare two settings: *contextualized* generation, where the target word is produced given its sentence context, and *decontextualized* (word-only) generation, where the model is prompted with the target word in isolation. We instantiate this comparison in two complementary directions: In *Literal-to-Metaphor (LM)*, LLMs generate meaning-equivalent metaphorical words from literal inputs, either in isolation or embedded in a literal context. In *Metaphor-to-Literal (ML)*, LLMs follow the same input settings but generate meaning-equivalent literal words from metaphorical inputs. Similarities between contextualized and decontextualized generations are interpreted as an indicator of context-invariant lexical associations, reflecting the persistence of stable lexical anchors in metaphor processing.

### 3.3 Syntactic Shuffle

Metaphors are often associated with characteristic syntactic patterns (Sullivan, 2013). Since word POS and order carries key syntactic relations like argument structure and modifier attachment, disrupting them breaks compositional structure while largely preserving lexical content, allowing us to probe whether detection rely on integrated sentence structure versus heuristic cues. We design three shuffle scenarios: 1) **Random Shuffle**: Words are randomly reordered, disrupting both semantic coherence and syntactic structure, creating sequences that substantially disrupt both syntactic structure and semantic coherence. 2) **POS Shuffle**: Metaphorical words are replaced with near-synonymous alternatives with different POS, introducing syntactic irregularities while minimally altering lexical semantics. 3) **Metaphorical Word Reposition**: The position of the metaphorical word is rearranged to the beginning, a random intermediate location (excluding the original, initial, and final positions), or the end of the sentence, allowing us to examine the detection behavior to the positional placement of the metaphorical word. By comparing those detection results, we assess sensitivity 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 non-literal language understanding (Liu et al., 2022). 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 their surface meanings. 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, as spans identified by GPT-4o (bold in Fig-QA examples).

For the Metaphor-Literal Imagination and syntactic shuffle experiments, we adopt Metaphor Understanding Challenge Dataset (MUNCH) (Tong et al., 2024), a linguistically annotated benchmark derived from the VU Amsterdam Metaphor Corpus (Steen et al., 2010). In MUNCH, each sentence is a LLM-challenging metaphor and primarily instantiated by a single annotated metaphorical word. We extract instances from its paraphrase generation task and derive multiple experimental settings from the same set of base sentences. For syntactic shuffle, tokenization and controlled lexical substitutions are implemented using WordNet 2020 (McCrae et al., 2020) to introduce systematic perturbations while minimally altering lexical meaning.

### 4.2 Models

We evaluate a diverse set of LLMs, including DeepSeek-V3-671B (V3-671B) (Liu et al., 2024),

| Dataset / Setting | Instances | Example (Metaphor ǀ Literal) |
|---|---|---|
| Fig-QA | 2.6k | The computer **is a race car**. ǀ The computer runs fast.<br>The computer **is a turtle**. ǀ The computer runs slow. |
| MUNCH (Context) | 2.9k | The council **appealed** by case stated ... ǀ<br>The council **petitioned** by case stated ... |
| MUNCH (Word) | 2.9k | **appealed** ǀ **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 ǀ Literal.

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 (LLaMA-3.1-8B) (Grattafiori et al.). To ensure comparability in the spatial analysis, all model-generated interpretations are encoded using `text-embedding-3-small` from OpenAI. This embedding model is used only for post-hoc geometric analysis and does not affect generation or decision-making. This allows us to perform consistent geometric comparisons across models with different internal representations. Open-source models were run on a T4 GPU of Google Colab. For all generations, we set the temperature parameter to 0 to minimize 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 $\gamma_i$ (defined by $\{R_i, R_i', S_i\}$) and an interpretation plane $\beta_i$ (defined by $\{R_i, R_i', M_i\}$). To quantify the geometric deviation of $M_i$ in the embedding space, we compute two measures: the perpendicular distance $d_p$ from $M_i$ to $\gamma_i$, and the cosine similarity $\cos\theta$ between $\gamma_i$ and $\beta_i$.

Both measures rely on Singular Value Decomposition (SVD) to obtain subspace bases in the shared embedding space. For each plane, we stack the corresponding sentence embeddings into a matrix $A$ and perform SVD:

$$A = U\Sigma V^\top. \quad (1)$$

We retain the leading singular components to derive an orthonormal subspace basis, which is then used to compute $d_p$ and $\cos\theta$. Larger $d_p$ or smaller $\cos\theta$ indicates greater geometric deviation from the reference semantic attribute. All distances and angles are computed in the original embedding space with respect to the constructed affine subspaces.

**Metaphorical Imagination** MUNCH contains metaphors whose meaning is instantiated by 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 approximate 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 Appendix 7. This includes the full prompts, detailed per-model metrics, additional breakdowns across conditions.

| $m_1$: **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.

| $R_1$ | $R_1{}'$ | *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 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*.

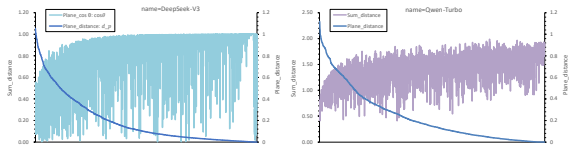

Figure 4: Distributions of ($d_p$, $Ad$) and ($d_p$, $\cos\theta$) for V3-671B and Qwen-T.

### 5.1 Semantic Attribute Alignment

To contextualize the geometric measures, we introduce an auxiliary similarity-based metric $Ad$, defined as the sum of cosine similarities between a model-generated interpretation $M_i$ and the two reference interpretations $\{R_i, R'_i\}$. Interpretations better aligned with the reference semantic attribute are expected to show higher $Ad$ and smaller deviation from the reference plane ($d_p$). In practice, as shown in Figure 4, lower $Ad$ is generally associated with larger $d_p$, and smaller $d_p$ corresponds to larger $\cos\theta$. These trends are supported by Spearman correlations between $d_p$ and $Ad$ ($\rho = -0.62$) and between $\cos\theta$ and $d_p$ ($\rho = -0.64$), suggesting that $d_p$ and $\cos\theta$ capture coherent geometric signals that are consistent with relative deviation from the reference semantic attribute proxy.

As a diagnostic contrast, we evaluate a multiple-choice interpretation setup in which candidate interpretations are restricted to an attribute-aligned option set. Specifically, for a given metaphor pair $\{m_i, m'_i\}$ in Fig-QA, $R_i$ denotes the correct human-annotated interpretation of $m_i$, while $R'_i$ is the interpretation of $m'_i$, which differs in surface meaning but shares the same underlying semantic attribute. We prompt GPT-4o to generate two meaning-preserving literal paraphrases $\{L'_{i1}, L'_{i2}\}$ of $R'_i$, yielding a four-way option set $\{R_i, R'_i, L'_{i1}, L'_{i2}\}$. Given a metaphor $m_i$, models are tasked with selecting the correct interpretation ($R_i$) from this set. Despite this controlled design, in which candidate interpretations are restricted to an attribute-aligned option set to avoid the confounding errors discussed above, models still 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 this restricted option design, discrete multiple-choice evaluation remains insufficient for characterizing model behavior. In particular, such a setup cannot reveal how far a model-generated interpretation deviates from the intended semantic attribute, motivating the need for a continuous analysis of deviation beyond forced-choice accuracy.

Spatial Analysis offers a complementary view of deviation from a reference semantic attribute region. Table 4 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 substantially from the reference semantic attribute proxy, suggesting that LLM-generated interpretations exhibit systematic drift

| | 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 4: 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.

| | **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 | <u>78.11</u> | 65.09 |
| **ML** | 76.96 | 75.17 | 79.22 | 78.01 | **81.55** | <u>80.87</u> | 72.86 |
| **News (LM)** | 74.43 | 74.91 | **80.96** | 76.90 | 72.76 | <u>78.00</u> | 66.28 |
| **News (ML)** | 77.82 | 75.23 | **83.77** | 79.32 | <u>81.27</u> | 80.68 | 72.11 |
| **Fiction (LM)** | 73.95 | 73.55 | 76.06 | **79.30** | 68.20 | <u>76.22</u> | 64.32 |
| **Fiction (ML)** | 75.21 | 73.19 | 74.73 | 75.70 | **78.73** | <u>77.97</u> | 67.79 |
| **Academic (LM)** | 75.02 | 76.11 | **82.71** | 77.15 | 71.09 | <u>80.37</u> | 65.46 |
| **Academic (ML)** | 80.69 | 79.69 | <u>84.93</u> | 81.84 | **85.25** | 83.62 | 75.09 |
| **Conversation (LM)** | 66.80 | 66.17 | 69.74 | 71.20 | <u>73.22</u> | **75.00** | 61.27 |
| **Conversation (ML)** | 73.35 | 71.13 | 71.16 | 73.86 | **81.87** | <u>81.11</u> | 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.

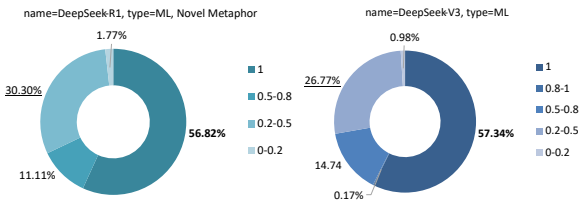


Figure 5: Anchor Scores distributions for metaphor-to-literal (ML) generation on novel metaphors by R1-671B and literal-to-metaphor (LM) generation by V3-671B.

relative to the attribute implied by the metaphor. We present the representative aligned and misaligned interpretations in Table 3. All experiments are conducted without task-specific fine-tuning and rely on a shared embedding model.

### 5.2 Lexical Invariance

Table 5 summarizes results for Metaphorical Imagination. Anchor Scores between contextualized and decontextualized generations are consistently high (approximately 65%–80%), indicating frequent context-invariant lexical anchors in metaphor–literal substitution. 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 further 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.

To examine lexical invariance for novel metaphors that may rely more on context, we analyze MUNCH metaphors with novelty scores > 0.3 (Tong et al., 2024). The score ranges from 0 to 1 (higher scores indicate greater novelty). Figure 5 presents the distribution of ML Anchor Scores for novel metaphors under R1-671B and V3-671B. Although overall scores decrease relative to the full dataset, more than 50% of cases still reach an Anchor Score of 1. The remaining instances concentrate around 0.2–0.5, suggesting that strong lexical invariance persists for certain metaphorical words, whereas models may not consistently converge on the same lexical anchor for other novel metaphors. Importantly, high Anchor Scores across generation settings do not necessarily translate into reliable metaphor detection performance. As shown in Section 5.3, models achieve substantially lower accuracy (around 30%) on direct metaphor detection in MUNCH. Therefore, lexical invariance may

| | 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.

support familiar metaphors by providing readily accessible associations; however, in richer contexts or more novel metaphors, reliance on stable anchors may bias interpretations toward highly associated lexical cues, contributing to trigger word effects.

### 5.3 Syntactic Influence

Across perturbation settings, detection accuracy varies substantially across models 6. LLaMA-3.1-8B exhibits relatively stable performance around chance level (approximately 50%), suggesting limited sensitivity to syntactic variation in metaphor detection. Combined with its low Anchor Scores in the Metaphorical Imagination, this pattern suggests limited sensitivity to the targeted probes on MUNCH. For other models, detection accuracy varies across perturbation types. Notably, LLMs achieve higher accuracy under POS shuffle than on the original sentences. POS shuffle preserves core lexical semantics while introducing syntactic irregularities, producing patterns that resemble selection preference violations described in SPV-based accounts of metaphor. In contrast, random shuffle disrupts both syntactic structure and sentence-level coherence, resulting in more heterogeneous effects.

Detection accuracy is relatively stable across positional perturbations (beginning/middle/end), suggesting that the absolute position of a metaphorical word has limited influence. Models appear more responsive to the presence of irregular linguistic patterns like POS shuffle (yielding the highest accuracy across models) than to global syntactic structure. Notably, V3-671B underperforms most other models (except LLaMA-3.1-8B) and even exceeds its original-sentence accuracy under random shuffle, suggesting less robust detection behavior; extreme perturbations may amplify irregular surface cues that it opportunistically uses for detection. More broadly, excluding the random shuffle condition (setting aside LLaMA-3.1-8B and o3-mini), most syntactic perturbations lead to improved detection accuracy across models, suggesting that syntactic irregularity itself can inflate detection performance without better metaphor understanding.

Overall, metaphor detection performance varies across models and perturbation settings, highlighting both the difficulty of MUNCH and the heterogeneous strategies employed by LLMs. These results suggest that current models can treat syntactic irregularity as a heuristic cue for metaphor detection, exhibiting limited evidence of robust integration of sentence-level syntactic structure.

## 6 Conclusion

This work investigates LLM behavior in metaphor processing from three complementary perspectives: semantic attribute alignment, context-invariant lexical associations, and syntactic influence. Spatial Analysis reveals consistent patterns of attribute-level deviation in model-generated interpretations. Using Metaphorical Imagination as a behavioral probe, we observe substantial overlap between contextualized and decontextualized generations, suggesting that models may encode metaphor–literal lexical associations that can support common metaphors or bias decisions toward internally associated cues. Syntactic Shuffle further suggest models frequently responding to syntactic irregularities as heuristic cues for metaphor detection.

Overall, while LLMs exhibit strong performance on metaphor processing tasks, our results indicate that models may reflect a combination of stored lexical associations and heuristic cues, which can support interpretation in familiar cases but may constrain context-sensitive integration of semantic attributes and syntactic structure. This underscores the importance of evaluation and modeling approaches that explicitly target attribute-level alignment, contextual reasoning, and robust syntactic integration beyond pattern-based cues.

## Limitations

Our analysis is subject to several limitations. First, the 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: although we use a shared embedding model to ensure comparability across LLMs, different embedding choices may affect absolute distances. Moreover, since each instance provides only two annotated sentences $R_i$,$R'_i$, we construct a third reference sentence $S_i$ (generated by the model) to avoid degenerating to a one-dimensional line and to enable more informative geometric comparison. We adopt this formulation as a pragmatic trade-off between interpretability and representational richness, and do not claim that the resulting dimensionality is optimal.

Second, the Metaphorical Imagination experiment focuses on single word substitutions, thus primarily explores word-level metaphor–literal associations. However, reasoning mechanisms under multi-word metaphor at discourse-level require further exploration. Furthermore, we note that some syntactic perturbations in Syntactic Shuffle, particularly random shuffling, intentionally produce inputs that are no longer linguistically interpretable. This is not intended to model natural language use, but rather to serve as diagnostic stress tests. The fact that metaphor detection accuracy is higher under such conditions highlights the extent to which detection behavior may reflect reliance 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.

## 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. 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.

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

This appendix provides definitions of key terminologies used in the paper, details of the prompt settings, and visualization results for each model and experiment.

# 8 Definitions and Terminology

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

## 8.1 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 semantically paraphrase to the predefined 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.

## 8.2 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), where literal words are generated from given metaphorical words, and (2) *Literal-to-Metaphor* (LM), where metaphorical words are generated from given literal words. Both tasks are conducted under contextualized and decontextualized settings. Accordingly, the prompts are divided into two types based on whether sentence context is provided.

## 8.3 Prompts for Syntactic Shuffle

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

## 8.4 Distributions of $(d_p, \cos\theta)$ and $(d_p, Ad)$

Figures 9–10 present the distributions of the distance $d_p$ between the model-generated interpretation $M_i$ and the reference plane $\gamma_i$, plotted against the cosine similarity $\cos\theta$ between the interpretation plane $\beta_i$ and the reference plane $\gamma_i$, as well as against the auxiliary similarity measure $Ad$, defined as the sum of cosine similarities between the interpretation and the standard references.

## 8.5 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 invariant associations may serve as a form of internal metaphor–literal knowledge that can potentially facilitate certain types of metaphor reasoning, while their interaction with context-sensitive inference remains an open question. Future work may explore how these lexical associations can be more explicitly characterized and controllably leveraged.

| 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.<br>Cars **drink** gasoline.<br>(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.<br>Metaphorical: obsessed in feelings<br>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.<br>Source domain: COMPETITION<br>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.<br>Semantic attribute: The speed (Slowness) |
| Conceptual Plane $\gamma_i$ | Constructed by the embeddings of three sentences $R_1$, $R_1'$ and $S_i$, implying correct concept mapping. Representing the ideal concept the metaphor intended to convey. | $R_1$: The computer runs fast.<br>$R_1'$: The computer runs slow.<br>$S_1$: The computer processes fast.<br>The conceptual plane include the concept of the speed of computers. |
| Interpretation Plane $\beta_i$ | 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.<br>$R_1'$: The computer runs slow.<br>$M_1$: 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.<br>The council by case **appealed** stated.<br>The council by case stated **appealed**. |

Table 7: Terminology and definitions in this study.

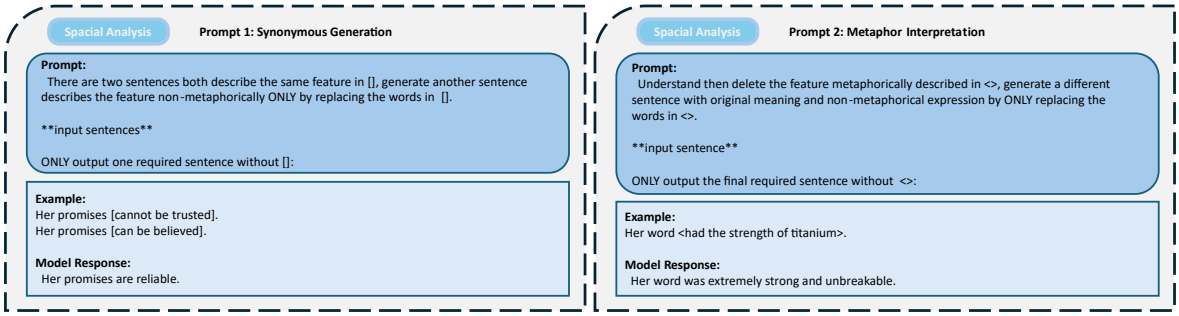


Figure 6: The prompts of Spatial Analysis.

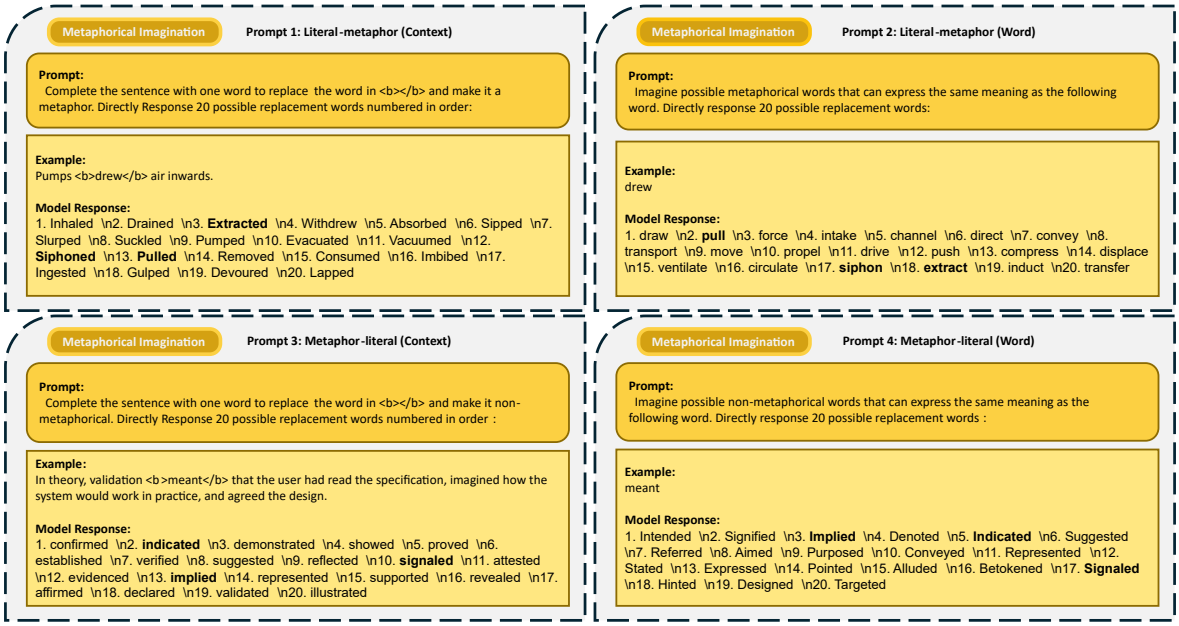


Figure 7: The prompts of Metaphorical Imagination.

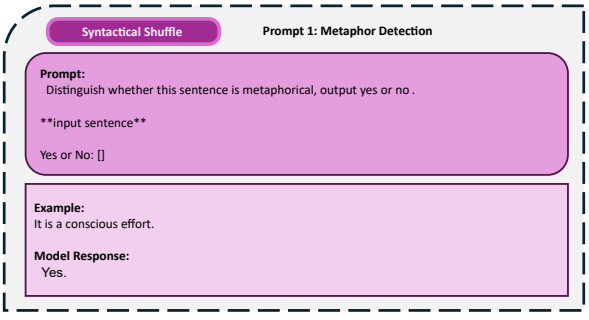


Figure 8: The prompt of Syntactic Shuffle.

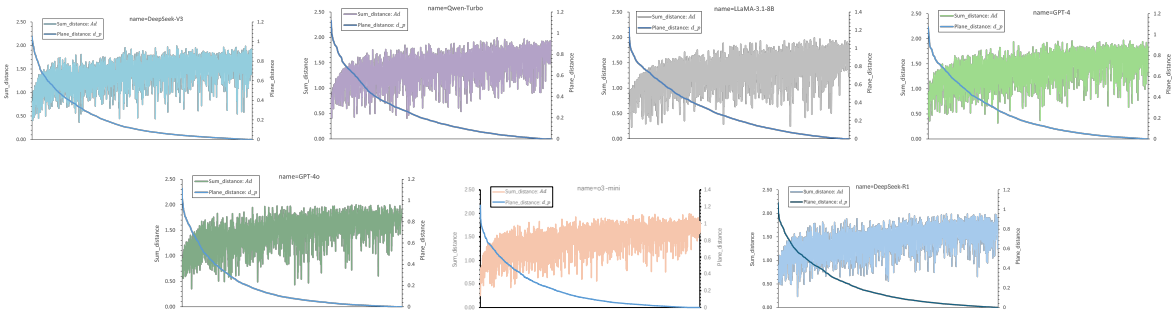

Figure 9: The ($d_p$, $Ad$) distribution of every model. Sort $d_p$ in decreasing order. Significant fluctuation can be observed due to the variance in the non-metaphorical part of sentences.

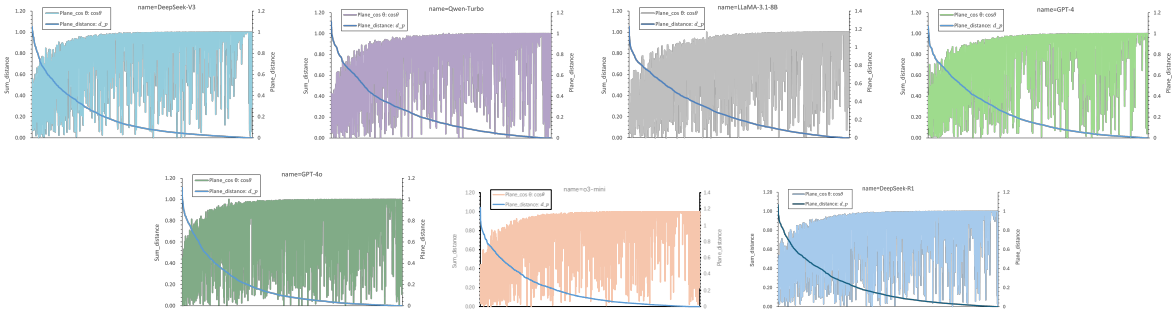

Figure 10: The ($d_p$, $\cos\theta$) distribution of every model. Sort $d_p$ in decreasing order. Significant fluctuation can be observed due to the variance in the non-metaphorical part of sentences.

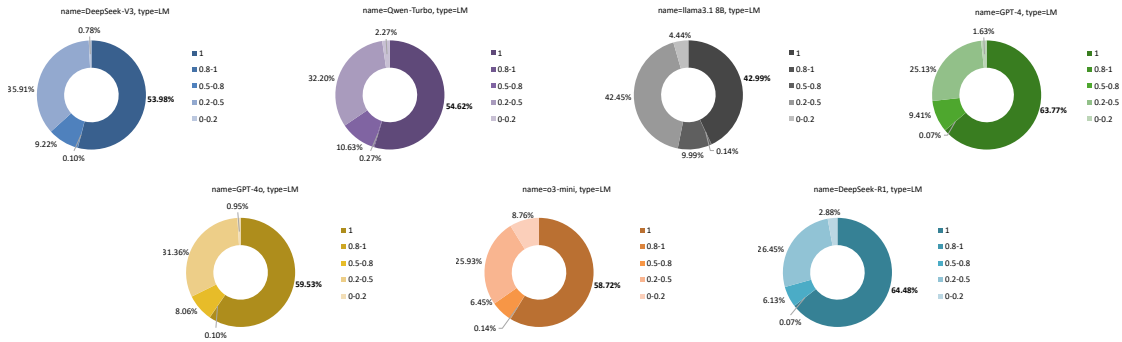


Figure 11: The overlap ratio distributions of literal-metaphor (LM) word imagination task on every model (the largest portion is in bold and the second largest is underlined).

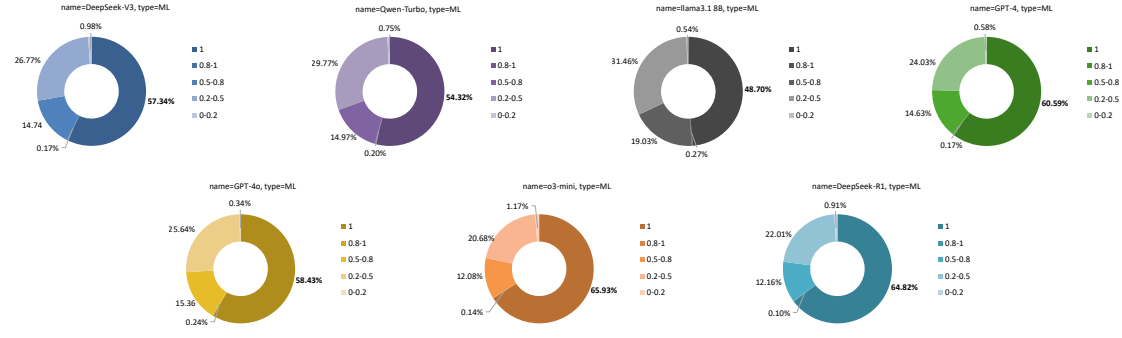


Figure 12: The overlap ratio distributions of metaphor-literal (ML) word imagination task on every model (the largest portion is in bold and the second largest is underlined).

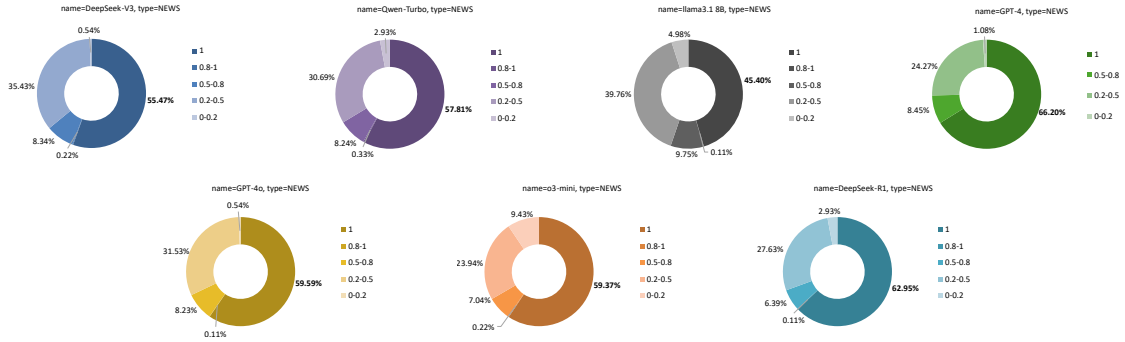


Figure 13: The overlap ratio distributions of literal-metaphor (LM) word imagination task with sentences in NEWS on every model (the largest portion is in bold and the second largest is underlined).

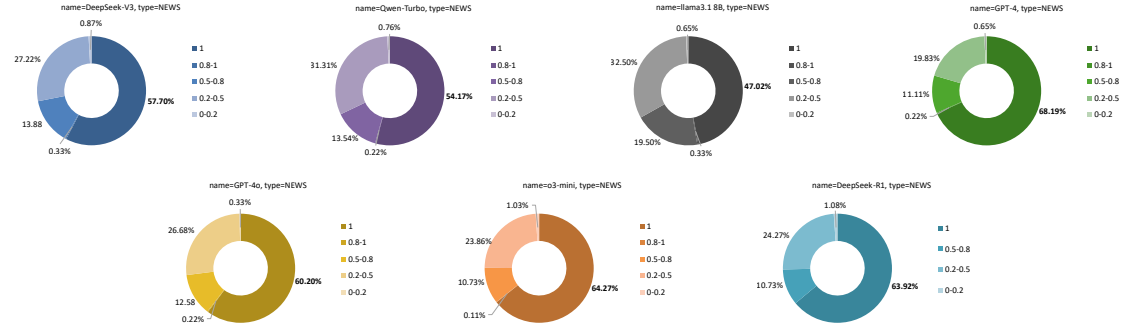


Figure 14: The overlap ratio distributions of metaphor-literal (ML) word imagination task with sentences in NEWS on every model (the largest portion is in bold and the second largest is underlined).

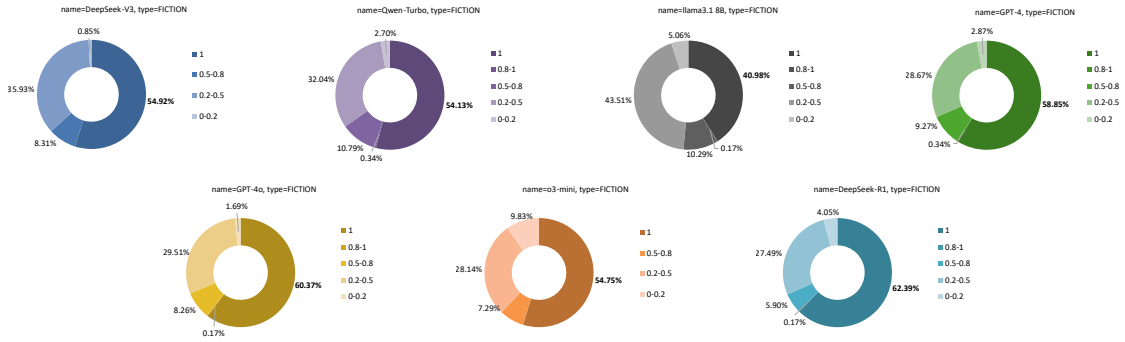


Figure 15: The overlap ratio distributions of literal-metaphor (LM) word imagination task with sentences in FICTION on every model (the largest portion is in bold and the second largest is underlined).

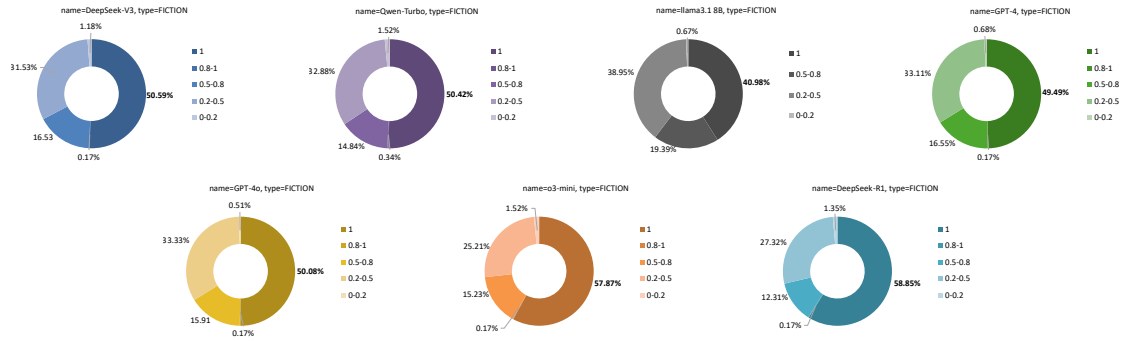


Figure 16: The overlap ratio distributions of metaphor-literal (ML) word imagination task with sentences in FICTION on every model (the largest portion is in bold and the second largest is underlined).

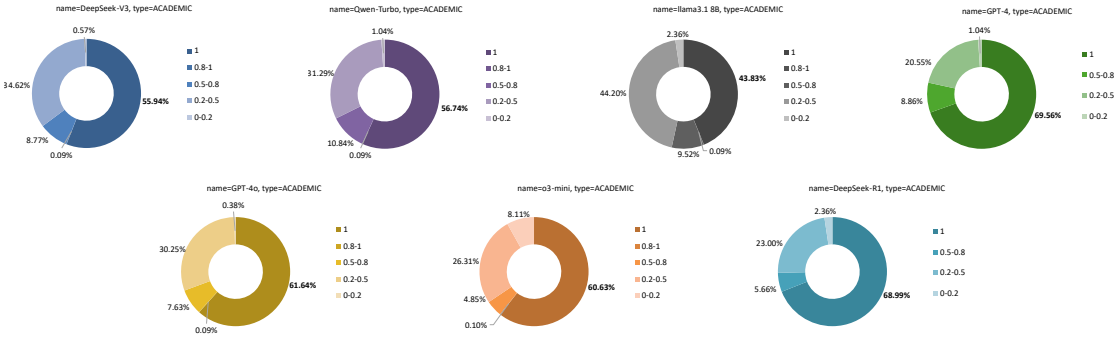


Figure 17: The overlap ratio distributions of literal-metaphor (LM) word imagination task with sentences in ACADEMIC on every model (the largest portion is in bold and the second largest is underlined).

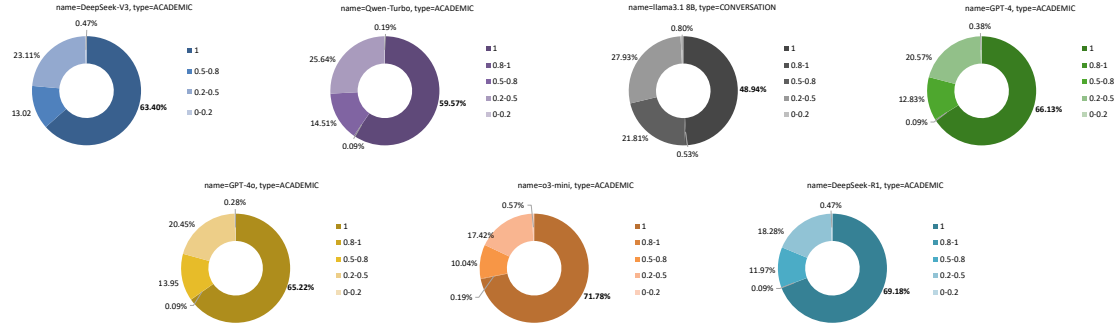


Figure 18: The overlap ratio distributions of metaphor-literal (ML) word imagination task with sentences in ACADEMIC on every model (the largest portion is in bold and the second largest is underlined).

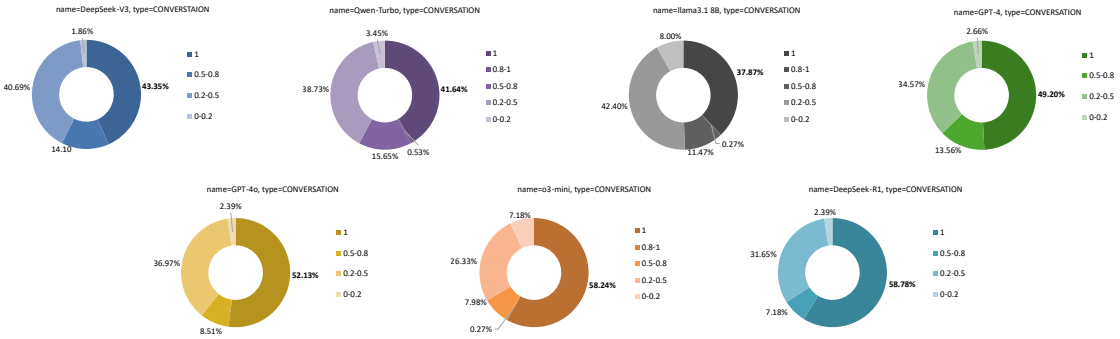


Figure 19: The overlap ratio distributions of literal-metaphor (LM) word imagination task with sentences in CONVERSATION on every model (the largest portion is in bold and the second largest is underlined).

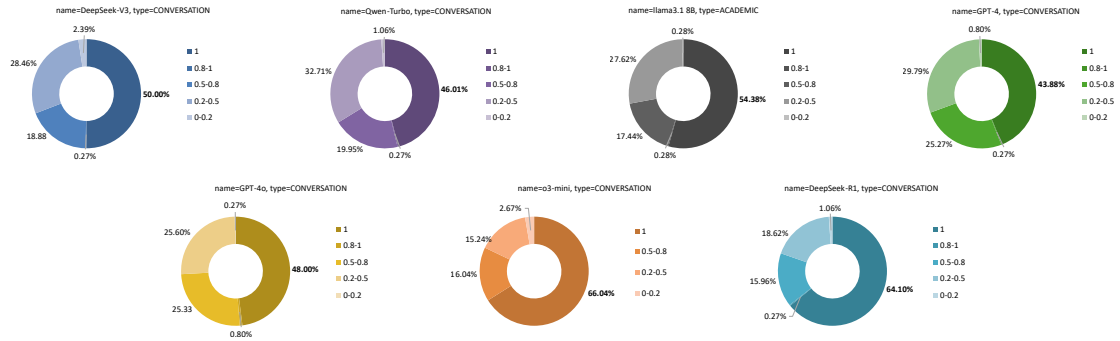


Figure 20: The overlap ratio distributions of metaphor-literal (ML) word imagination task with sentences in CONVERSATION on every model (the largest portion is in bold and the second largest is underlined).