Cognitive Parallels in Metaphor Processing: Human Acquisition vs. Large Models

Anonymous ACL submission

Abstract

Metaphor comprehension is a complex cognitive task in language acquisition that requires reasoning between surface structures and deeper semantic representations. Prior research has predominantly treated metaphor acquisition and automatic metaphor detection as separate topics, lacking a direct comparative analysis. This paper systematically reviews studies on metaphor acquisition in linguistics and identifies four cognitive aspects that align with the capabilities of large language models: aptness, language proficiency, transferable comprehension, and literal salience hypothesis. Experimental results reveal significant parallels between large model performance and human metaphor learning. Specifically, large models achieve higher accuracy on highly aptness metaphor samples. Language proficiency is reflected in their capacity for metaphor comprehension, which benefits from richer corpora, larger parameter scales, and more efficient architectures. Furthermore, large models exhibit sensitivity to transferable comprehension, as demonstrated by the substantial influence of cross-linguistic knowledge on metaphor processing. Similarly, they align with the literal salience hypothesis, prioritizing literal meanings over metaphorical ones, a pattern evident in their higher accuracy in metaphor detection.

1 Introduction

Metaphor is not merely a linguistic device or an intrinsic reflection of an individual's cognitive structure but also an adaptive behavior shaped by perceptual and cultural influences (Gibbs Jr, 1999). Traditional research on metaphor acquisition primarily focuses on cognitive modeling, the developmental process of metaphor comprehension, and cross-linguistic metaphor acquisition.

Cognitive models seek to elucidate the cognitive mechanisms underlying metaphor understanding. For instance, conceptual metaphor theory posits that the mapping between source and target domains conveys meanings beyond the literal level through structured associations (Gibbs Jr, 1999; Lakoff and Johnson, 2008). Additionally, metaphorical aptness refers to the extent to which a metaphor encapsulates the core attributes of a given concept (Chiappe and Kennedy, 1999; Gibbs Jr, 1993). For example, the metaphorical expressions "Time is money" and "The clouds are old newspapers" differ significantly in aptness. The former draws on a well-established conceptual analogy, where "time" is commonly perceived as valuable, akin to "money," reinforcing the notion that wasting time equates to financial loss. Consequently, its metaphorical relevance is high. In contrast, the latter lacks an intuitive conceptual bridge between "clouds" and "old newspapers." While clouds may appear fragmented or darkened, "old newspapers" is not a conventional metaphorical mapping for this phenomenon, resulting in lower metaphorical relevance. The absence of conceptual connection further diminishes its interpretability. Investigating aptness effects can provide insights into improving model performance on low-aptness metaphors.

The literal salience hypothesis posits that literal interpretations are typically activated preferentially during semantic processing, as evidenced by faster response times to literally meaningful phrases (López et al., 2017; Citron et al., 2020) and the preferential activation of literal meanings (Yang et al., 2023; Giora, 1999). Large-scale models generally prioritize learning literal semantics when processing metaphors, resulting in a cognitive bias in metaphor comprehension. Understanding this hypothesis is crucial for optimizing model performance in handling metaphorical expressions.

Research on metaphor comprehension acquisition investigates how language learners process and use metaphors. Prior studies indicate that metaphor acquisition improves progressively with age (Willinger et al., 2019). Language proficiency

refers to an individual's ability to comprehend, express, utilize, and adapt to a language, as reflected in their linguistic performance. Learners with higher language proficiency typically exhibit stronger metaphor comprehension (Aleshtar and Dowlatabadi, 2014; Fabry, 2021) and lower cognitive processing costs when interpreting metaphors (Carrol et al., 2016; Jankowiak et al., 2021). In large models, language proficiency emerges from enriched training data, increased parameters, and optimized architectures. Enhancing metaphor processing through corpus expansion, parameter scaling, or architectural refinement is of significant theoretical and practical importance.

Cross-linguistic metaphor acquisition studies examine how cultural differences impact metaphor comprehension, particularly in second language (L2) learners. Research on transferable comprehension highlights the influence of a speaker's native language (L1) on L2 metaphor understanding, demonstrating that L1 metaphorical competence is a strong predictor of L2 comprehension (Wang and Sun, 2020). Furthermore, L1 knowledge is often automatically applied in L2 metaphor learning (Carrol et al., 2016; Cieślicka, 2015). Investigating transferable comprehension enhances large models' multilingual metaphor processing capabilities.

The above studies are essential for enhancing large models in metaphor detection and cross-linguistic metaphor comprehension. This paper conducts a comprehensive comparison between language technology and human language learning, investigating whether aptness, language proficiency, transferable comprehension, and literal salience influence the metaphor detection capability of large models. For aptness estimation, we first employ MetaPro2.0 (Mao et al., 2024) to extract source and target domain information from metaphorical texts. We then utilize WordNet's (Miller, 1995; Christiane, 1998) superordinate word relations to compute the semantic similarity between source and target domains, thereby quantifying metaphor aptness. Regarding language proficiency, we define it in large models based on the richness of their training corpus and the scale of model parameters. For transferable comprehension, we evaluate the model's ability to detect metaphors across different linguistic contexts. Lastly, to examine the literal salience hypothesis, we compare classification accuracies between literal and metaphorical samples across multiple linguistic metaphor datasets.

In summary, this paper makes the following key contributions:

1. We systematically reviews studies on metaphor acquisition in linguistics, synthesizing prior work and summarizing key advancements.

2. We conduct the first systematic investigation of the similarities and differences between language technology and human language learning, offering theoretical insights and practical guidance for metaphor understanding.

3. We design and implement four experimental frameworks to examine metaphor aptness, language proficiency, transferable comprehension, and literal salience, analyzing their impact on metaphor processing in large models.

2 Related Work

2.1 Aptness

Metaphor comprehension involves both inter-domain and intra-domain similarity, which together determine aptness. Studies have shown that aptness is positively correlated with inter-domain distance and negatively correlated with intra-domain distance (Tourangeau and Sternberg, 1981). Aptness is a more decisive factor than conventionality, as highly aptness metaphors are easier to understand and accept, whereas conventionality has a weaker influence. This finding supports the categorization-based model of metaphor comprehension, which posits that metaphor understanding depends on its relevance to an ontological framework (Jones and Estes, 2006). Although conventionality and aptness are closely related—both influenced by metaphor frequency in corpora—aptness scores may be affected by processing fluency, making independent measurement challenging (Thibodeau and Durgin, 2011). Individual cognitive abilities also play a role in metaphor comprehension, with crystallized intelligence being more influential in processing high-aptness metaphors, while fluid intelligence serves a compensatory function in understanding low-aptness metaphors (Stamenković et al., 2023).

The aptness of a metaphor is crucial for comprehension, choice of expression, and the distinction between metaphorical and explicit comparisons. Research indicates that comparisons with high aptness are more likely to be expressed as

metaphors, while those with low aptness tend to appear as explicit comparisons. Additionally, aptness affects memory bias—high-aptness metaphors are more likely to be recalled as metaphors, whereas low-aptness ones are often remembered as explicit comparisons (Chiappe and Kennedy, 1999; Chiappe et al., 2003). Cultural context also influences metaphor aptness. Cross-cultural studies reveal significant differences in aptness ratings of the same metaphors across linguistic groups. For example, certain metaphors receive different ratings from native English and Persian speakers, suggesting that cultural experience and linguistic conventions shape metaphor acceptance (Eskandari and Khoshsima, 2021).

2.2 Language Proficiency

Research on language proficiency and metaphorical competence suggests that higher proficiency correlates with improved metaphor comprehension and usage (Aleshtar and Dowlatabadi, 2014; Willinger et al., 2019; Fabry, 2021). L2 learners generally require greater cognitive effort for metaphor collocation processing, whereas increased proficiency reduces processing costs (Willinger et al., 2019). Additionally, bilinguals demonstrate a higher initial cognitive load when processing novel metaphors in L2, but their cognitive processing aligns with L1 during the late-stage meaning integration (Carrol et al., 2016). Moreover, metaphor production ability improves with language development, highlighting the role of linguistic resources in metaphor acquisition (Jankowiak et al., 2021). However, in L2 metaphor comprehension, executive control exerts less influence on familiar metaphors, while the processing of unfamiliar metaphors is constrained by the conceptual similarity between languages (Lü et al., 2019). These findings underscore the pivotal role of language proficiency in metaphor processing.

2.3 Transferable Comprehension

Studies highlight the critical role of L1 knowledge in L2 metaphor and idiom acquisition. Multilinguals outperform monolinguals in novel metaphor comprehension, attributed to their greater cognitive flexibility (Horvat et al., 2022). Furthermore, advanced non-native speakers exhibit native-like formulaic processing in L2 idiom comprehension, suggesting automatic activation of L1 knowledge (Carrol et al., 2016). L1 metaphorical competence not only surpasses L2 competence but also serves as a strong predictor of L2 metaphor processing ability, supporting the cross-linguistic transfer hypothesis (Wang and Sun, 2020). During the early stages of L2 idiom acquisition, learners rely on L1 vocabulary and conceptual structures, with direct L2 connections forming as proficiency increases (Cieślicka, 2015). Transparency, context, and L1-L2 similarity all influence L2 idiom comprehension; context, in particular, facilitates L2 metaphor processing while mitigating L1 interference (Wang et al., 2021).

2.4 Literal Salience Hypothesis

The Literal Salience Hypothesis posits that literal meanings are generally more readily activated and processed than non-literal meanings, particularly in the early stages of cognitive processing. Empirical research supports this claim, demonstrating that bilinguals make faster and more accurate judgments on literal phrases, regardless of their experience with language mediation (López et al., 2017). Additionally, bilinguals struggle to suppress literal meanings when interpreting metaphorical expressions in L2, indicating that highly salient meanings are preferentially activated even in metaphor-biased contexts (Yang et al., 2023). Further evidence for literal salience comes from studies showing that familiar expressions activate both literal and metaphorical meanings in idiomatic and metaphor comprehension, whereas less familiar expressions primarily activate literal meanings (Giora, 1999). L2 learners also exhibit distinct processing patterns compared to L1 speakers when interpreting conventional metaphors, with metaphorical meanings being less semantically integrated in the L2 mental lexicon—further reinforcing the dominance of literal meanings (Werkmann Horvat et al., 2021). Collectively, these findings suggest that literal meanings typically hold greater salience than non-literal meanings in bilingual language processing.

3 Method

This paper aims to comprehensively compare the similarities and differences between language technology and human language learning. We extensively review the literature related to metaphors in the field of linguistics and summarize four aspects that are representative and similar to the capabilities of the larger model, i.e., aptness, language proficiency, transferable comprehension, and literal salience hypothesis.

Aptness. Metaphorical Aptness quantifies the extent to which a metaphor encapsulates the core attributes of the target concept (Chiappe and Kennedy, 1999; Gibbs Jr, 1993).. This study investigates the categorization performance of large models on metaphors with varying levels of aptness, which is formalized based on the theory of double similarity (Tourangeau and Sternberg, 1981). According to this theory, metaphor comprehension is governed by both intra-domain and inter-domain similarity: aptness positively correlates with inter-domain distance and negatively correlates with intra-domain distance. Inter-domain distance reflects the conceptual disparity between domains (e.g., "animal" and "political leader" belong to biology and social organization, respectively, exhibiting a large inter-domain distance), while intra-domain distance captures similarity within a domain (e.g., "lion" and "eagle" share high similarity in "power" and "aggressiveness," resulting in a small intra-domain distance). The study of aptness aims to: (1) enhance large models' comprehension of low-aptness metaphors and (2) integrate aptness into metaphor evaluation frameworks to refine and extend the assessment of metaphor understanding in large models.

Language Proficiency. Linguistic proficiency, defined as an individual's ability to comprehend, express, and adapt to language, is positively correlated with metaphor comprehension (Aleshtar and Dowlatabadi, 2014; Fabry, 2021). Analogously, a large model's performance in a given language reflects its language proficiency, which is primarily influenced by corpus richness, parameter scale, and structural optimization. Corpus richness, akin to human language exposure, enables models to capture diverse linguistic patterns, improving comprehension and generation. Parameter scale, reflecting cognitive capacity, allows models to learn complex patterns, boosting metaphor understanding and production. Structural optimization, resembling cognitive strategy refinement, enhances model efficiency and accuracy in metaphor processing. Investigating the role of large model language proficiency in metaphor comprehension serves two purposes: (1) assessing whether corpus size, parameter scale, and structural optimization jointly enhance metaphor understanding, thereby informing model development, and (2) identifying key factors influencing metaphor processing to refine baseline settings and improve evaluation frameworks.

Transferable Comprehension. This concept examines the extent to which native language (L1) semantic and conceptual frameworks facilitate second language (L2) metaphor processing. Prior research has demonstrated the significant influence of L1 knowledge on L2 acquisition (Horvat et al., 2022; Wang and Sun, 2020). Analogously, in natural language processing, cross-linguistic transfer emerges as a key phenomenon, where L1 knowledge may enhance a model's ability to detect metaphors in L2, potentially conferring advantages over monolingual models. Investigating whether large models exhibit transferability in metaphor comprehension serves two objectives: (1) improving metaphor recognition in low-resource languages by leveraging shared cognitive foundations across languages and (2) enhancing cross-linguistic metaphor detection by incorporating linguistic resources with similar cognitive and cultural structures.

Literal Salience Hypothesis. The literal salience hypothesis posits that literal meanings are cognitively prioritized over metaphorical meanings during language processing, with literal interpretations being activated first in semantic comprehension (Citron et al., 2020; Yang et al., 2023; Giora, 1999). In natural language processing, large language models similarly exhibit a bias toward literal semantics, often achieving higher accuracy on non-metaphorical samples than on metaphorical ones. This pattern parallels human language acquisition, where learners typically grasp literal meanings and conventional metaphorical expressions before acquiring novel metaphors in both L1 and L2. Investigating whether large models align with the literal salience hypothesis has two primary goals: (1) analyzing cross-linguistic commonalities in metaphor construction through the lens of literal salience, thereby providing a unified framework for metaphor comprehension, and (2) establishing a theoretical foundation for improving contextual modeling approaches in metaphor processing.

4 Aptness Experiment

4.1 Experimental Design

Aptness experiment investigates the impact of metaphor aptness on the metaphor detection performance of large language models. Due to resource constraints, the study is conducted exclusively on English data. We employ the MetaPro2.0 (Mao et al., 2024) to preprocess the VUA20 cor-

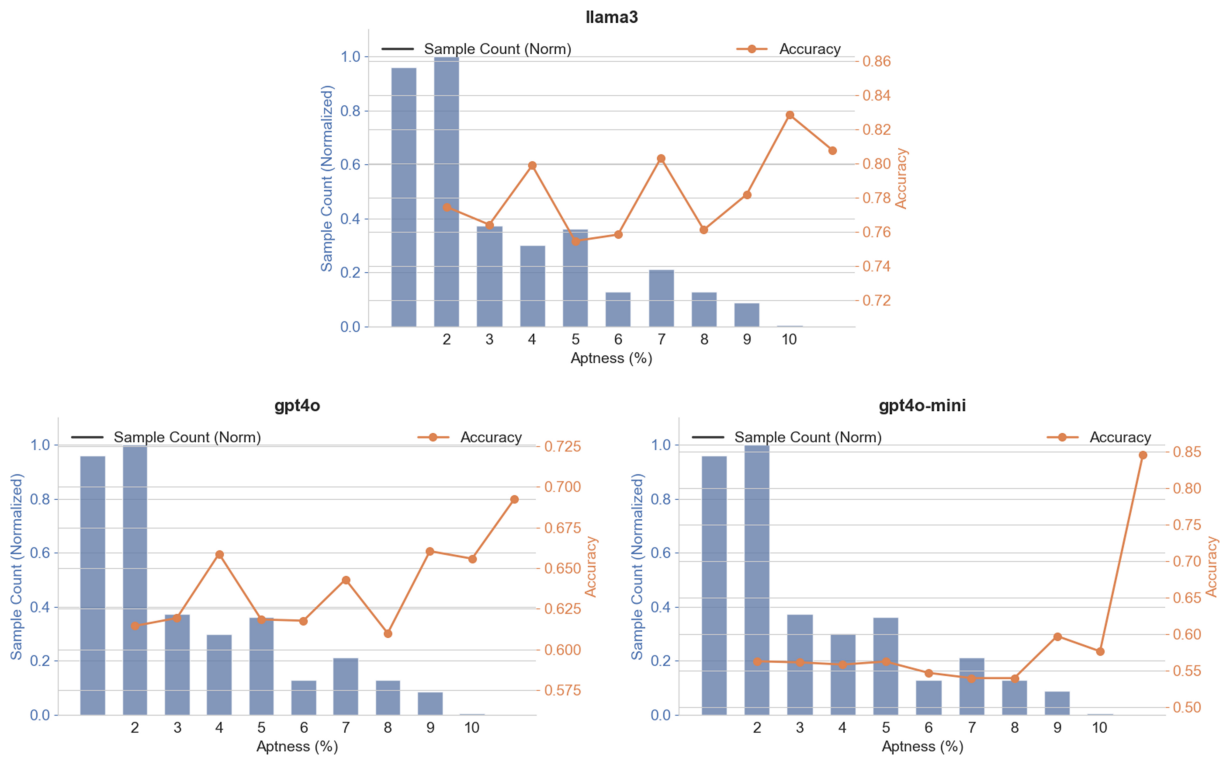

Figure 1: Aptness experiment results. The Sample Count Ratio represents the proportion of samples in each suitability interval relative to the maximum interval. Accuracy denotes the classification accuracy of metaphor samples.

pus (Leong et al., 2020), extracting the source and target domains of metaphorical expressions. The aptness of metaphor samples are computed based on the hypernym structure of the WordNet (Miller, 1995; Christiane, 1998) semantic hierarchy.

To quantify inter-domain and intra-domain distances more efficiently, we introduce the *concept similarity ratio* as a metric for metaphor aptness, formulated as follows:

$$S(c_1, c_2) = \frac{D_L}{D_L + d_1 + d_2} \quad (1)$$

where c_1 and c_2 represent the source and target domain concepts, respectively, and L denotes Lowest Common Subsumer (LCS). D_L is the depth of the LCS, reflecting the semantic hierarchical distance between the two domains; a larger D_L corresponds to greater inter-domain distance and, consequently, a higher aptness. d_1 and d_2 represent the shortest path lengths from c_1 and c_2 to L, respectively; a smaller $d_1 + d_2$ indicates lower intra-domain distance, implying higher aptness.

After computing the aptness for each metaphor sample in VUA20 using Eq. (1), we partitioned the samples into *10 intervals* within the range $[0, 1]$, each with a step size of 0.1. Lower indices correspond to lower aptness. To examine differences in classification performance between high-aptness and low-aptness metaphors, we evaluate both closed-source and open-source models (See Appendix 11.1 for details).

4.2 Experimental Analysis

Figure 1 presents the association between sample distribution and model performance using a dual-coordinate statistical system. The left vertical axis (blue bars) represents the *sample size ratio*, defined as the ratio of the number of samples in each metaphor aptness interval to the number in the largest interval. The right vertical axis (red line) denotes the *metaphor detection accuracy* of the models across intervals.

Experimental results indicate an overall upward trend in detection accuracy as metaphor aptness increases; however, model-specific variations exist. The GPT-4o-mini accuracy curve remains relatively stable, fluctuating within 1.5% across intervals 1 to 7, followed by a slight increase in the higher aptness intervals (>7). Spearman correlation analysis suggests that its positive correlation is not statistically significant ($\rho_s = 0.345, p = 0.328$).

In contrast, GPT-4o and LLaMA3 exhibit greater

sensitivity to interval variation, showing a fluctuating yet generally increasing trend as aptness increases ($\rho_s = 0.515, p = 0.328$). Notably, both models display a consistent rise in accuracy in high-aptness intervals (>7). Although LLaMA3 experiences a slight decline at the highest interval, this drop carries limited statistical weight due to the small sample size in that range.

Despite the overall positive correlation observed in all models, statistical significance remains insufficient. This may be attributed to two key factors: (1) the *long-tail effect* in sample distribution, where high-aptness samples constitute less than 20% of the dataset; and (2) *labeling noise*, arising from the inherent ambiguity in defining metaphor aptness boundaries.

5 Language Proficiency Experiment

5.1 Experimental Design

Language proficiency experiment investigates the impact of language proficiency on metaphor detection by evaluating three multilingual pre-trained models: mBERT, XLM-RoBERTa, and mDeBERTa (See Appendix 11.1 for details). To systematically evaluate model performance on metaphor detection, we analyze three key dimensions:

1. **Corpus Coverage**: We compare models trained on Wikipedia and CC100, noting that while both cover 100 languages, CC100 provides a larger and more balanced dataset.

2. **Model Size**: We assess the impact of model capacity by comparing the base (mRoB-r) and large (mRoB-l) versions of XLM-RoBERTa.

3. **Architectural Enhancements**: We examine improvements in DeBERTa and its multilingual variant (mDeBERTa) over mBERT and XLM-RoBERTa in metaphor detection.

For dataset selection, we integrate multiple publicly available metaphor corpora to ensure broad applicability (See Appendix 11.2 for details). Table 4 provides data statistics.

5.2 Experimental Analysis

The experimental results, presented in Table 1, offer the following key insights derived from comparative analysis of different models across multi-source datasets:

Impact of Pre-trained Corpus Diversity on Metaphor Comprehension. Models trained on the CC100 corpus outperform mBERT, which was trained on the Wikipedia monolingual corpus, in the metaphor detection task. For instance, on the VUA20 dataset, mRob-b, mRob-l, and mDeb-b achieved F1 score improvements of 0.5%, 1.9%, and 3.1%, respectively, over the benchmark mBERT (F1 = 0.756). These results indicate that a more diverse pre-training corpus fosters broader conceptual mappings, thus enhancing the model's capacity to capture cross-domain relations in metaphorical expressions. This aligns with the human language acquisition process, where the richness of linguistic input directly influences metaphor comprehension.

Effect of Model Parameter Scale Expansion. The experiments further demonstrate the significant positive impact of model capacity expansion on metaphor detection. For example, expanding the parameter scale from mRob-b (L = 12, H = 768, A = 12, 270M params) to mRob-l (L = 24, H = 1024, A = 16, 550M params) resulted in F1 score improvements of 1.4%, 0.7%, 3.9%, and 0.6% across four benchmark datasets. Notably, on the CoMeta dataset, which has a sparse distribution of metaphors, the model's parameter expansion caused a significant F1 improvement, from 0.505 to 0.544. This highlights the advantages of large models in addressing long-distance dependencies and metaphorical inference, further enhancing cross-linguistic generalization.

Gains from DeBERTa Structural Optimization. DeBERTa outperforms RoBERTa in metaphor detection through its decoupled attention mechanism, enhanced mask decoder, and virtual adversarial training. For instance, compared to mRob-b, mDeb-b achieved an 11.8% increase in F1 score on the CoMeta dataset (from 0.505 to 0.623), and surpassed the larger RoBERTa model on other datasets. These results suggest that architectural optimization not only improves the model's capacity to handle long texts and complex syntactic structures, but also enhances its ability to perform metaphorical reasoning tasks.

From a cognitive perspective, corpus richness parallels linguistic input in human learning, model scale reflects cognitive resources, and architectural optimization mirrors cognitive strategy adaptation. This aligns with findings in cognitive science, where metaphor ability is strongly related to cognitive resources and linguistic input (Gask-

Table 1: Language proficiency experiment results. Metrics include Precision (Pre), Recall (Rec), and F1 score (F1).

Model	VUA20			PSUCMC			CoMeta			KOMET		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
mbert	0.785	0.730	0.756	0.753	0.720	0.736	0.430	0.560	0.487	0.710	0.703	0.706
mRob-b	0.754	0.767	0.761	0.803	0.692	0.744	0.609	0.431	0.505	0.741	0.682	0.710
mRob-l	0.798	0.754	0.775	0.767	0.735	0.751	0.562	0.526	0.544	0.752	0.684	0.716
mDeb-b	0.789	0.784	**0.787**	0.791	0.735	**0.762**	0.635	0.612	**0.623**	0.736	0.730	**0.733**

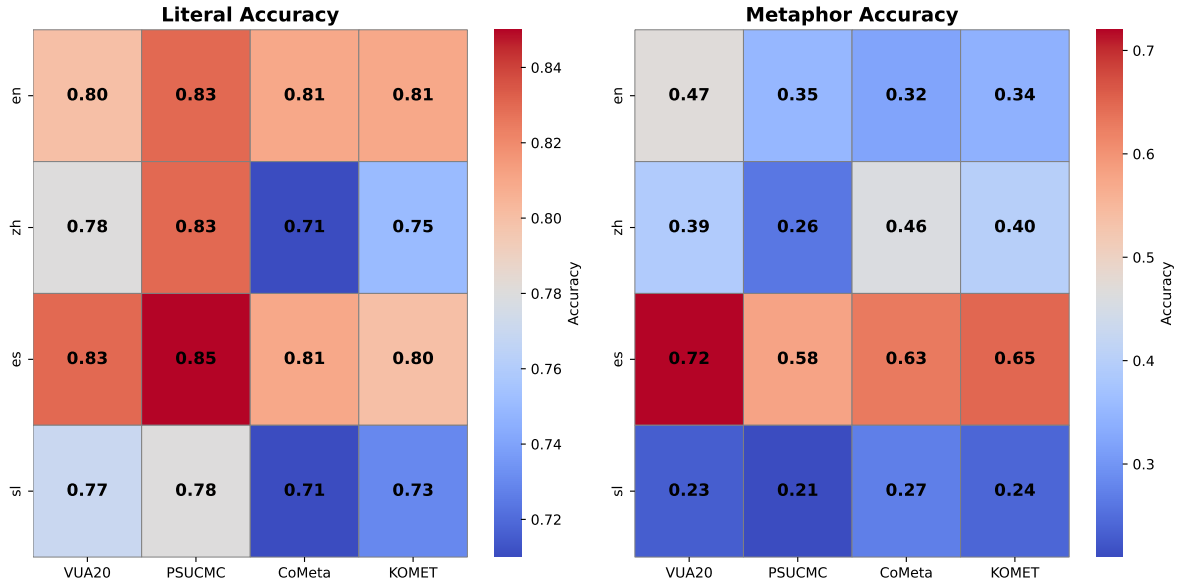

Figure 2: Transferability experiment results. The left and right figures show detection accuracy for literal and metaphorical target words, respectively. The vertical axis represents the inquiry language (L1), and the horizontal axis denotes the target language.

ins and Rundblad, 2023), with most studies highlighting language proficiency as a key factor influencing metaphor comprehension (Aleshtar and Dowlatabadi, 2014; Carrol et al., 2016; Fabry, 2021; Jankowiak et al., 2021).

6 Transferability Experiment

6.1 Experimental Design

Transferability experiment was designed as a series of 4×4 cross-language metaphor recognition tasks to investigate the impact of the questioning language (L1) on metaphor comprehension in the target language (L2). The study involved four languages—English, Chinese, Spanish, and Slovenian—resulting in 16 distinct test sets. Each set utilized GPT-4o-mini for metaphor detection. Specifically, the model processed a text in the target language and identified metaphorical expressions using a fixed-format prompt. Below is an example of the prompt (in the case of an English question):

The experimental results are shown in Figure 2. In the literal condition (Fig. 3(a)), the model demonstrates high stability across the multilingual questioning conditions. For example, in the VUA20 dataset, the accuracy difference between the best-performing model (0.85) and the lowest (0.78) is only 6 percentage points, while the difference in the PSUCMC dataset is similarly small (0.83 vs. 0.76, a 7% variation). This stability suggests that literal semantic comprehension remains largely unaffected by cross-linguistic conditions, likely due to the task's reliance on lexical matching and basic grammatical structures rather than higher-order cognitive processing.

In contrast, the metaphorical part (Fig. 3(b)) reveals substantial variations in the model's cross-linguistic performance. In the VUA20 dataset, the

Table 2: Literal saliency experiment results. Lit. and Met. denote detection accuracy for literal and metaphorical samples, respectively, while ALL represents overall accuracy.

Model	VUA20			PSUCMC			CoMeta			KOMET		
	Lit.	Met.	All	Lit.	Met.	All	Lit.	Met.	All	Lit.	Met.	All
mbert	0.964	0.767	0.939	0.978	0.721	0.957	0.995	0.431	0.985	0.981	0.703	0.964
mRob-b	0.971	0.730	0.941	0.984	0.692	0.960	0.993	0.526	0.984	0.985	0.684	0.967
mRob-l	0.973	0.754	0.945	0.980	0.735	0.959	0.987	0.560	0.979	0.984	0.682	0.966
mDeb-b	0.970	0.784	0.946	0.982	0.735	0.961	0.994	0.612	0.987	0.983	0.730	0.967

accuracy fluctuated by as much as 49% (0.72 vs. 0.23) across different linguistic conditions, while the PSUCMC dataset showed a 37% variation (0.58 vs. 0.21). These significant fluctuations highlight the strong dependence of metaphor comprehension on L1 knowledge, a phenomenon consistent with the cognitive characteristics of metaphors in human L2 learning. For instance, L2 learners' understanding of metaphorical expressions is influenced by the lexical and conceptual similarities between L1 and L2 (Wang et al., 2021). Additionally, L1 knowledge tends to be automatically activated when non-native speakers process L2 idioms (Carrol et al., 2016). Therefore, future research could explore how to enhance the model's metaphor comprehension for specific low-resource languages by targeting language pairs with similar semantic structures.

7 Salience Experiment

The results, summarized in Table 2, show that the model's classification accuracies were significantly higher for literal samples (Lit.) compared to metaphorical samples (Met.), a trend that was consistent across models and datasets. For instance, on the VUA20 dataset, the mBERT model achieved an accuracy of 0.964 for literal samples, far surpassing its performance on metaphorical samples (0.767). Similarly, the mRob-b, mRob-l, and mDeb-b models all demonstrated significantly better performance in detecting literal semantics than metaphorical ones. In addition, Transferability experiment further tests the literal salience hypothesis. Figures 2 reveal that the model's ability to capture literal semantics is significantly better than its ability to detect metaphorical semantics ($p < 0.01$). For example, on the VUA20 dataset, the model's detection accuracy for literal samples is 80%, whereas the accuracy for metaphorical samples is only 47%, a 33% difference. This trend is also evident in cross-linguistic scenarios. In the Chinese context, for example, the model's accuracy on the VUA20 English dataset is 78% for literal samples, but only 39% for metaphorical samples, further supporting the literal salience hypothesis.

This phenomenon aligns with the cognitive characteristics of human learners. Research has demonstrated that bilinguals are quicker and more accurate at judging phrases with literal meanings (López et al., 2017). Additionally, studies (Citron et al., 2020) have found that the activation of literal meanings occurs more frequently during reading, as indicated by shorter reading times. This suggests that processing literal semantics is not only more intuitive for neural network models but also represents a more cognitively efficient mode for humans.

8 Conclusion

This study reviews theories on metaphor acquisition in linguistics and conducts experiments to explore metaphor comprehension in large models. Results show that as the aptness interval increases, the model's accuracy in metaphor detection improves, demonstrating its ability to capture metaphor usage across contexts. Factors such as a richer training corpus, larger model size, and optimized architecture enhance metaphor comprehension, emphasizing the importance of large data and efficient modeling. The models also show strong cross-linguistic adaptation, leveraging shared semantic features for cross-cultural metaphor reasoning. However, the model recognizes literal meanings better than metaphorical ones, supporting the literal salience hypothesis and highlighting the need for further advancements in metaphor detection.

9 Limitations

This paper explores the similarities and differences in metaphor acquisition between humans and large language models (LLMs). However, limitations exist due to the long-tailed data distribution and varying metaphor similarity across languages, leading to some non-significant results. In particular, the multilingual transferable comprehension task revealed instability in the model's cross-linguistic generalization. Future research will use high-quality metaphor resources like parallel corpora and expand language types to enhance robustness. Additionally, integrating cognitive science approaches for metaphor representation learning is expected to improve the model's reasoning and alignment with human cognition.

10 Ethics Statement

This study adheres to academic ethical standards, ensuring fairness and transparency in data collection, processing, and experimental design. All metaphor corpora were sourced from publicly available resources, without involving sensitive or personal data. In multilingual experiments, we accounted for linguistic and cultural differences to prevent bias. Given the potential impact of model bias on metaphor parsing, we carefully analyzed the limitations of the models, particularly in cross-cultural understanding, to avoid inappropriate generalizations. Future research will focus on creating a more equitable and culturally adaptive metaphor comprehension system to minimize bias and enhance fairness and interpretability in multilingual settings.

References

Maryam Teymouri Aleshtar and Hamidreza Dowlatabadi. 2014. Metaphoric competence and language proficiency in the same boat. *Procedia-Social and Behavioral Sciences*, 98:1895–1904.

Gareth Carrol, Kathy Conklin, and Henrik Gyllstad. 2016. Found in translation: The influence of the l1 on the reading of idioms in a l2. *Studies in Second Language Acquisition*, 38(3):403–443.

Dan L Chiappe and John M Kennedy. 1999. Aptness predicts preference for metaphors or similes, as well as recall bias. *Psychonomic Bulletin & Review*, 6(4):668–676.

Dan L Chiappe, John M Kennedy, and Penny Chiappe. 2003. Aptness is more important than comprehensibility in preference for metaphors and similes. *Poetics*, 31(1):51–68.

Fellbaum Christiane. 1998. Wordnet: an electronic lexical database. *Computational Linguistics*, pages 292–296.

Anna B Cieślicka. 2015. Idiom acquisition and processing by second/foreign language learners. *Bilingual figurative language processing*, pages 208–244.

Francesca MM Citron, Nora Michaelis, and Adele E Goldberg. 2020. Metaphorical language processing and amygdala activation in l1 and l2. *Neuropsychologia*, 140:107381.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

B Edition, BNC Baby, and BNC Sampler. 2007. British national corpus.

Zahra Eskandari and Hooshang Khoshsima. 2021. A study of cross-cultural variations of metaphor aptness and their implications in foreign language teaching. *International Journal of Knowledge and Learning*, 14(3):193–215.

Regina E Fabry. 2021. Getting it: A predictive processing approach to irony comprehension. *Synthese*, 198(7):6455–6489.

Dorota Gaskins and Gabriella Rundblad. 2023. Metaphor production in the bilingual acquisition of english and polish. *Frontiers in Psychology*, 14:1162486.

Raymond W Gibbs Jr. 1993. Process and products in making sense of tropes.

Raymond W Gibbs Jr. 1999. Taking metaphor out of our heads and putting it into the cultural world. In *Metaphor in cognitive linguistics*, page 145. John Benjamins.

Rachel Giora. 1999. On the priority of salient meanings: Studies of literal and figurative language. *Journal of pragmatics*, 31(7):919–929.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Ana Werkmann Horvat, Marianna Bolognesi, Jeannette Littlemore, and John Barnden. 2022. Comprehension of different types of novel metaphors in monolinguals and multilinguals. *Language and Cognition*, 14(3):401–436.

Katarzyna Jankowiak, Marcin Naranowicz, and Karolina Rataj. 2021. Metaphors are like lenses: Electrophysiological correlates of novel meaning processing in bilingualism. *International Journal of Bilingualism*, 25(3):668–686.

Lara L Jones and Zachary Estes. 2006. Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, 55(1):18–32.

Matej Klemen and Marko Robnik-Šikonja. 2023. Neural metaphor detection for slovene. In *CLARIN Annual Conference*, pages 77–89.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 vua and toefl metaphor detection shared task. In *Proceedings of the second workshop on figurative language processing*, pages 18–29.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Belem G López, Jyotsna Vaid, Sümeyra Tosun, and Chaitra Rao. 2017. Bilinguals' plausibility judgments for phrases with a literal vs. non-literal meaning: the influence of language brokering experience. *Frontiers in psychology*, 8:1661.

Junmei Lü, Lijuan Liang, and Baoguo Chen. 2019. The effect of executive control ability on the comprehension of second language metaphor. *International Journal of Bilingualism*, 23(1):87–101.

Xiaofei Lu and Ben Pin-Yun Wang. 2017. Towards a metaphor-annotated corpus of mandarin chinese. *Language Resources and Evaluation*, 51:663–694.

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.

Anthony McEnery and Zhonghua Xiao. 2004. The lancaster corpus of mandarin chinese: A corpus for monolingual and contrastive language study. *Religion*, 17:3–4.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new spanish corpus for multilingual and crosslingual metaphor detection. *arXiv preprint arXiv:2210.10358*.

Dušan Stamenković, Katarina Milenković, Nicholas Ichien, and Keith J Holyoak. 2023. An individual-differences approach to poetic metaphor: Impact of aptness and familiarity. *Metaphor and Symbol*, 38(2):149–161.

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.

Paul H Thibodeau and Frank H Durgin. 2011. Metaphor aptness and conventionality: A processing fluency account. *Metaphor and Symbol*, 26(3):206–226.

Roger Tourangeau and Robert J Sternberg. 1981. Aptness in metaphor. *Cognitive psychology*, 13(1):27–55.

Juanjuan Wang and Yi Sun. 2020. How is chinese english learners' l2 metaphoric competence related to that of l1? an e-prime-based multi-dimensional study. *International Journal of English Linguistics*, 10(4).

Xiaolu Wang, Yizhen Wang, Wanning Tian, Wei Zheng, and Xiaoli Chen. 2021. The roles of familiarity and context in processing chinese xiehouyu: An erp study. *Journal of Psycholinguistic Research*, pages 1–21.

Ana Werkmann Horvat, Marianna Bolognesi, and Katrin Kohl. 2021. The status of conventional metaphorical meaning in the l2 lexicon. *Intercultural Pragmatics*, 18(4):447–467.

Ulrike Willinger, Matthias Deckert, Michaela Schmöger, Ines Schaunig-Busch, Anton K Formann, and Eduard Auff. 2019. Developmental steps in metaphorical language abilities: The influence of age, gender, cognitive flexibility, information processing speed, and analogical reasoning. *Language and Speech*, 62(2):207–228.

Huilan Yang, J Nick Reid, and Yuru Mei. 2023. Conceptual metaphor activation in chinese–english bilinguals. *Bilingualism: Language and Cognition*, 26(2):345–355.

11 Appendix

11.1 Model Introduction

1. **LLaMA3**[1]: Released by Meta AI on April 18, 2024, LLaMA3 is available in 8B and 70B parameter versions. This study utilizes the *LLaMA3-70B-Instruct* model, accessible via official request.

2. **ChatGPT**[2]: Developed by OpenAI, ChatGPT is a closed-source model accessible through API-based subscription. Two versions are employed in this study: *GPT-4o-mini-2024-07-18* and *GPT-4o-2024-08-0*.

3. **mBERT** (Multilingual BERT) (Conneau and Lample, 2019) is a multilingual extension of BERT (Devlin, 2018), pre-trained on Wikipedia[3] across 100 languages using masked language modeling (MLM) and next sentence prediction (NSP).

4. **XLM-RoBERTa** (Conneau, 2019) extends RoBERTa (Liu, 2019), pre-trained solely with MLM on the CC100 dataset[4]. Both its base (mRoB-r) and large (mRoB-l) versions are included in this study.

5. **mDeBERTa-V3** (He et al., 2021) is a multilingual variant of DeBERTa (He et al., 2020), structurally aligned with its monolingual counterpart and trained on CC100. It incorporates ELECTRA-style contrastive pretraining and a gradient-decoupled embedding-sharing mechanism to enhance generalization.

11.2 Dataset Introduction

1. **VUA** (VU Amsterdam Metaphor Corpus) (Steen et al., 2010), based on the British National Corpus (BNC) (Edition et al., 2007), contains 187,570 word-level metaphor annotations labeled using MIPVU. It covers four text genres: academic, dialogue, fiction, and news. This study employs the VUA20 version (Leong et al., 2020).

2. **CoMeta** (Sanchez-Bayona and Agerri, 2022) comprises two subsets:

   (a) *Universal Dependencies (UD)*: Processed and deduplicated news, blogs, and Wikipedia texts, totaling 2,862 sentences.

   (b) *Political Discourse (PD)*: Parliamentary records from Spanish and Basque governments, containing 771 sentences.

   Both subsets are annotated using MIPVU.

3. **PSUCMC** (Lu and Wang, 2017), derived from the Lancaster Corpus of Mandarin Chinese (LCMC) (McEnery and Xiao, 2004), includes 1M words spanning academic, fiction, and news texts, with MIPVU-based metaphor annotations.

4. **KOMET** (Klemen and Robnik-Šikonja, 2023), sourced from the Slovenian Corpus of Young People's Literature (MAKS), contains 13,963 annotated sentences from news reports, literary works (e.g., novels, essays), and online texts, following MIPVU annotation.

11.3 Prompt Design

Table 3: Prompt for Metaphor Word Identification

LLM Prompt
Determine whether `sent` contains metaphorically used words. If so, output only those words separated by semicolons; otherwise, return `none`.

[1] https://llama.meta.com/llama-downloads
[2] https://platform.openai.com/
[3] https://meta.wikimedia.org/wiki/List_of_Wikipedias
[4] https://huggingface.co/datasets/statmt/cc100

Table 4: Dataset Statistics (VUA20, CoMeta, PSUCMC, KOMET)

Dataset	**Total**	**Metaphor Samples**	**Metaphor (%)**	**Sentences**	**Avg. Sentence Length**
VUA20	182,263	23,146	12.70%	14,482	12.59
CoMeta	117,038	2,144	1.83%	3,595	32.56
PSUCMC	35,753	2,918	8.16%	1,718	20.81
KOMET	258,099	16,009	6.20%	13,696	18.84