MetFuse: Figurative Fusion between Metonymy and Metaphor

Saptarshi Ghosh and Tianyu Jiang
University of Cincinnati
ghosh2si@mail.uc.edu, tianyu.jiang@uc.edu

Abstract

Metonymy and metaphor often co-occur in natural language, yet computational work has studied them largely in isolation. We introduce a framework that transforms a literal sentence into three figurative variants: metonymic, metaphoric, and hybrid. Using this framework, we construct **MetFuse**,[1] the first dedicated dataset of figurative fusion between metonymy and metaphor, containing 1,000 human-verified meaning-aligned quadruplets totaling 4,000 sentences. Extrinsic experiments on eight existing benchmarks show that augmenting training data with MetFuse consistently improves both metonymy and metaphor classification, with hybrid examples yielding the largest gains on metonymy tasks. Using this dataset, we also analyze how the presence of one figurative type impacts another. Our findings show that both human annotators and large language models better identify metonymy in hybrid sentences than in metonymy-only sentences, demonstrating that the presence of a metaphor makes a metonymic noun more explicit.

1 Introduction

Metonymy and metaphor are two fundamental linguistic phenomena in figurative language that involve concept mapping (Radden and Kövecses, 1999). While both entail a shift in meaning, they operate through distinct mechanisms. Metonymy primarily occurs through a change in meaning of the noun. Metaphors, by contrast, are more varied in form. Verbal metaphors are a prominent subtype, where the figurative shift arises through the verb. The crucial difference between metonymy and metaphor is that in metaphoric mapping, two discrete domains are involved, whereas mapping in metonymy occurs within a single domain (Goossens, 1990; Lakoff and Johnson, 1980).

For instance, in the metonymic sentence "*The **stadium** celebrated joyfully*," the noun *stadium* (a location) refers to the fans in the stadium. The mapping stays within the same domain because both the location and the people associated with it belong to the broader conceptual domain of real-world entities tied to a place. Now looking at the metaphor "*The fans **erupted** with joy*," the verb *erupted* highlights the intense emotional reaction of the fans through the domain of physical explosion, making it a cross-domain mapping. Hence, metonymy preserves domain continuity, while metaphor requires a conceptual leap across domains—making it structurally and cognitively more distinct (Lakoff and Johnson, 1980). Interestingly, a metonymic noun and metaphoric verb can co-exist in a sentence particularly in creative writing, resulting in figurative fusion like "*The **stadium erupted** with joy*." Such constructions are powerful because they capture multiple layers of meaning into a single expression, enriching the expressive potential of the language, enabling writers to create vivid, memorable scenes that resonate with readers at both conceptual and emotional level (Lakoff and Johnson, 1980; Turner and Fauconnier, 2002). By modeling these combinations, we can push LLMs toward generating more nuanced, stylistically rich text, which is crucial for applications in narrative generation and other creative domains.

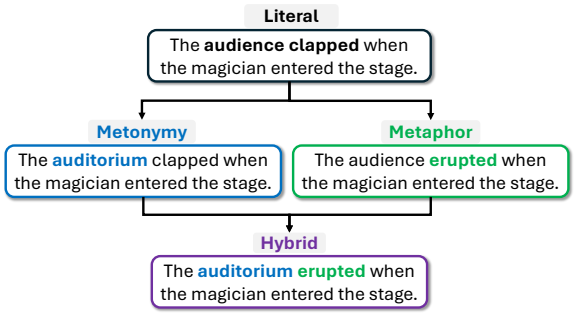


Figure 1: Illustration of generating metonymy, metaphor and hybrid sentences from literal expressions.

Prior linguistic works have studied metonymy

[1] https://github.com/cincynlp/MetFuse

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15337–15350

and metaphor separately, primarily focusing on their differences in terms of their linguistic and cognitive nature (Lakoff and Johnson, 1980; Radden and Kövecses, 1999). There are also some linguistic works, namely that of Goossens (1990) and Barcelona (2003), that study metonymy and metaphor together, theorizing their interactions in context. However, these theoretical works were constrained by the lack of resources (Goossens, 1990). To our knowledge, very few works have explored metonymy and metaphor jointly in the context of modern NLP and LLMs.

Filling in these research gaps, we explore the fusion of metonymy and metaphor. We introduce a literal-to-figurative transformation framework that takes a literal sentence as input, and produces three outputs: a metonymic, a metaphoric and a hybrid (containing both metonymy and metaphor) variant of the sentence, as depicted in Figure 1. Our approach significantly outperforms baselines. Human annotation statistics show that 78% of the generated sentences using our framework were judged as valid realizations of the intended figurative category, compared to only 53% when using a straightforward prompt.

Leveraging this framework, we introduce **MetFuse**, the first dataset of figurative fusion between metonymy and metaphor, containing 1,000 quadruplets, each comprising a literal sentence paired with its metonymic, metaphoric and hybrid counterparts, yielding 4,000 sentences in total. On eight existing metonymy and metaphor classification benchmarks, augmenting training data with MetFuse consistently improves performance, with hybrid examples yielding the largest gains on metonymy tasks. Using this dataset, we also find that the presence of a metaphor makes a metonymic noun more explicit. Human annotators rate the metonymy in hybrid sentences as more explicit than in metonymy-only sentences (3.65 vs 3.47 out of 5). Models show the same pattern: augmenting BERT's training data with hybrid examples improves metonymy classification more than augmenting with metonymy-only examples across four benchmarks, and four LLMs more accurately identify metonymy in hybrid sentences than in metonymy-only sentences in a zero-shot setting. In summary, our contributions are:

1. **Framework:** We propose a framework to generate metonymy, metaphor, and hybrid variations of a literal expression.
2. **MetFuse dataset:** We introduce MetFuse, the first dataset of figurative fusion between metonymy and metaphor, with 1,000 meaning-aligned quadruplets totaling 4,000 sentences.
3. **Findings:** Our experimental results demonstrate that augmenting training data with MetFuse consistently improves metonymy and metaphor classification across eight existing benchmarks, and both humans and LLMs identify metonymy more easily when a metaphoric verb is present.

2 Related Work

Metonymy & Metaphor. Lakoff and Johnson (1980) introduced the Conceptual Metaphor Theory (CMT), the now-standard view that metaphor involves cross-domain mappings, which are relatively *unconstrained and highly generative* (Kövecses, 2010). In contrast, metonymy operates within a single domain (Nunberg, 1995), playing a crucial structural role in meaning construction (Herrero-Ruiz, 2004), making it cognitively constrained (Ruiz de Mendoza, 2002; Ruiz de Mendoza and Baicchi, 2005; Herrero-Ruiz, 2004), and therefore limited by pre-existing contiguity relations such as part–whole and container–content (Radden and Kövecses, 1999; Brigitte Nerlich, 2001). Unlike metaphor, metonymy is often approached as a cognitive and pragmatic phenomenon, rather than purely linguistic terms (Radden and Kövecses, 1999; Jodlowiec and Piskorska, 2015; Papafragou, 1996).

Computational approaches to metonymy have primarily focused on resolution tasks (Markert and Nissim, 2007; Gritta et al., 2017; Pedinotti and Lenci, 2020; Ghosh and Jiang, 2025), treating the metonymy resolution as a classification task. Recently, metonymy has also been explored in a multimodal setting (Ghosh et al., 2026). Metaphor identification has been an extensive field of research (Steen, 2010; Mohammad et al., 2016; Birke and Sarkar, 2007). With the advent of LLMs, recent studies have focused on metaphor interpretation (Chakrabarty et al., 2022a,b), probing LLMs' figurative language understanding capabilities.

Existing literature has also studied metonymy and metaphor together. Goossens (1990) defined "*metaphtonymy*," a phenomenon where metaphor and metonymy co-occur in figurative expressions. Barcelona (2003) shows how metaphor often builds on a metonymic base, or how metonymy becomes

foregrounded under metaphoric pressure. Maudslay et al. (2024) introduces a structured resource that integrates metaphor and metonymy into WordNet, providing relational chains that capture figurative meaning extensions.

Generation. In the NLP space, early work in metaphor generation was heuristic in approach (Abe et al., 2006; Terai and Nakagawa, 2010). Recently, a lot of work has explored metaphor generation, including using literal-metaphor verb pairs (Yu and Wan, 2019), transforming literal expressions to metaphors (Stowe et al., 2021b) and similes (Chakrabarty et al., 2020) using fine-tuned models. Stowe et al. (2021b) proposed a supervised and unsupervised method to generate verbal metaphors, with the former yielding better results. The work of Stowe et al. (2021a) analyzed verbal metaphor generation, finding that controlled generation improves metaphoricity, while free generation tends to generate more fluent paraphrases. Our metaphor generation strategy is inspired by Chakrabarty et al. (2021), which proposed a supervised method to generate verbal metaphors by replacing the relevant verb in the literal expression, and Tian et al. (2021) that uses a partly-supervised framework to generate hyperbole without exploring the figurative expression it creates such as metaphor and sarcasm. Metonymy generation has not received much attention (Hernández, 2017), and our research aims to combine metonymy and metaphor generation.

3 Methodology

In this section, we discuss our methodology of generating metonymic, metaphoric and hybrid sentences from literal expressions. Our model takes a literal text as input, and transforms it into three figurative expressions: metonymy, metaphor and a hybrid sentence combining metaphor and metonymy.

3.1 Preprocessing Steps

Not all literal expressions can be transformed to its respective metonymy or metaphor counterpart. Metonymy occurs through a reference shift in the meaning of the noun, therefore, it is crucial that the literal sentence must contain a noun. Prior linguistic studies have shown that metonymy is more constrained than metaphor, since the mapping must remain within a single conceptual domain (Goossens, 1990; Kövecses, 2010). In a canonical *subject–verb–object* structure, both the *subject* and the *object* can in principle serve as sites of metonymic substitution. However, we restrict our focus to the subject noun for three reasons: (i) the *subject* more often denotes animate, human entities, which are particularly productive for metonymic shifts, (ii) the *subject* position offers more flexibility in shifting reference without disrupting grammatical structure, whereas *object* substitutions can more easily lead to semantic or syntactic anomalies, and (iii) focusing on *subjects* provides a consistent and controlled framework for generation, while still capturing the core phenomenon. Accordingly, we impose three conditions: the noun must denote a human entity, it must function as the *subject*, and it must be in a dependency relation with a verb (enabling us to study its interaction with verbal metaphors). Under these constraints, the resulting instances mainly correspond to location metonymy. We then use the SpaCy dependency parser to extract the literal sentences from Wikipedia dumps[2] based on the conditions set above.

3.2 Generating Metonymy

Generating metonymy is a challenging task. The noun undergoes a reference shift, but the mapping must be intra-domain. Our framework contains three different stages—i) using an LLM to generate candidate replacement nouns, ii) identifying the best candidate noun, iii) replacing the noun and asking an LLM to refine the sentence.

i. Generating candidate metonymic nouns. We first provide the LLM with the literal sentence and the target noun. Naively asking it to "make a metonymic replacement" yields incorrect responses. Instead, we draw out the contiguity relations that evoke metonymy. We pose short, targeted questions about the noun's location, occupants/constituents, or salient parts. Each answer is treated as a candidate replacement for the noun phrase. We set a temperature of 0.7 to encourage diversity.

ii. Selecting the metonymic noun. To select the most suitable metonymic candidate from those generated, we used a masked language modeling approach with BERT. Specifically, we replace the target noun in the literal sentence with a [MASK] token and present the sentence to BERT. For each candidate c, we compute its log probability of filling the masked position. The candidate with the highest probability is chosen as the replacement,

[2]https://huggingface.co/datasets/wikimedia/wikipedia

Noun	Literal Input	Candidates	Scores	Metonymic Sentence
judge	The [MASK] in Lincoln saw issues	**law office** briefcase Springfield	**-4.25** -12.28 -11.96	The **law office** in Lincoln saw issues
Queen	The [MASK] officially pronounced Turing pardoned in August 2014.	crown monarchy **Buckingham Palace**	-1.58 -11.58 **-0.27**	**Buckingham Palace** officially pronounced Turing pardoned in August 2014.

Table 1: Illustrations of transforming literal expressions into metonymic ones using an LLM to generate candidate substitutions and masked token probabilities to select the best replacement.

as it is deemed the most contextually appropriate. Table 1 shows an example.

iii. Replacing the noun. We select candidate $c*$ with the highest probability and substitute it for the original noun, yielding the metonymic variant of the sentence. To address any syntactic or grammatical inconsistencies introduced by the substitution, we then prompt the LLM to refine the sentence while preserving the metonymic noun. We use a temperature of 0.4 to discourage creative rewrites, ensuring that the refinement maintains high semantic similarity with the original literal sentence.

3.3 Generating Metaphor

In this work, we focus on generating verbal metaphors by mapping the verb to another domain. While metonymy typically arises through a reference shift in the noun, the verb also plays a crucial role in meaning construction (Radden and Kövecses, 1999). Mapping the verb to another domain makes the metaphor and metonymy directly dependent, enabling us to study the relation between them. Importantly, linguistic studies highlight that metaphor mapping allows for greater structural freedom than metonymy, which is typically constrained by contiguity relations (Kövecses, 2010). Building on this observation, Stowe et al. (2021a) showed that in verbal metaphor generation, controlled settings tend to increase metaphoricity, whereas free generation produces more fluent paraphrases. Motivated by these findings, and given the expressive power of recent LLMs, we allow the model absolute freedom in selecting the metaphoric domain—unlike in the case of metonymy—thereby exploiting this structural flexibility. To further intensify the domain shift, we also incorporate hyperbolic verbs in the spirit of Tian et al. (2021), since hyperboles often make the metaphoric mapping more dramatic and salient.

Similar to metonymy generation, we use a 3-step pipeline—i) using an LLM to generate candidate verbal hyperboles, ii) identifying the best candidate verb, iii) using the verb and asking an LLM to refine the sentence.

i. Generate metaphor candidates. We provide the LLM with the literal sentence and the target verb. To generate verbal hyperboles, we instruct the model to exaggerate the verb to the extent it maps to another domain (Lakoff and Johnson, 1980). However, we observed that these hyperboles often clashed with the overall tone of the sentence. Since Lakoff and Johnson (1980); Mohammad et al. (2016) noted that tone plays a crucial role in metaphor interpretation, we incorporate this insight by asking the LLM to generate verbal hyperboles under three distinct tones—*positive, negative* and *neutral*, with a temperature of 0.7 and top-p 0.9 to encourage diversity. This additional context provides the LLM more flexibility and guidance, leading to coherent candidate generations.

ii. Selecting the hyperbole verb. To select the most suitable metaphor candidate verb that aligns with the overall tone of the literal sentence, we use a lightweight sentiment analysis model (Camacho-collados et al., 2022). Based on the predicted sentiment, we select the hyperbole candidate that best matches with the tone.

iii. Refining the Metaphor. We then replace the target word in the literal expression with the selected hyperbole candidate. We provide the sentence to the LLM and ask it to refine the sentence. We prompt it to maintain the metaphoric meaning while making any syntactical or structural adjustments to improve the sentence quality with a temperature of 0.6 and top-p 0.9.

3.4 Constructing Hybrid Expressions

To make the hybrid sentences, we take the metonymic noun phrase from the refined metonymic sentence, and replace it with the noun phrase in the refined metaphor sentence. While the metaphor generation step often alters the structure

of the literal sentence—reflecting the greater freedom of metaphoric mapping—metonymy is more constrained in nature (Radden and Kövecses, 1999). As a result, metonymic generation typically preserves the original syntactic structure, modifying only the noun phrase. Consequently, inserting the metonymic noun phrase into the metaphor sentence produces a hybrid expression without the need for additional post-processing.

We extracted literal sentences from Wikipedia as mentioned earlier in this section, and fed them into our framework to generate the metonymic, metaphoric and hybrid variants of the sentence. For our main experiment, we use Llama-3.1-8B.

4 Framework Evaluation & Dataset

In this section, we first evaluate our framework using both human and automatic evaluation. Then, we use this framework to create the MetFuse dataset, the first dataset to contain instances of metonymy and metaphor combined.

4.1 Evaluation of Our Framework

To compare against our proposed method, we employ a general-purpose prompting baseline. For each literal sentence, the LLM is queried in three independent passes to produce: (i) a metonymic variant: given the sentence and target noun, the LLM is asked to replace the noun with a metonymic paraphrase, (ii) a metaphoric variant: given the sentence and target verb, the LLM is asked to transform the verb into a verbal metaphor, and (iii) a hybrid variant combining both: given the sentence, noun, and verb, the LLM is asked to introduce a metonymic paraphrase complemented by a verbal metaphor. Each pass uses a carefully designed chain-of-thought prompt for that particular variant with few-shot exemplars, ensuring fair comparison.

Human Evaluation. We pick 250 literal sentences from our pool and use our framework to generate the three figurative variants. We also use the general prompting method to generate the three variants for the same sample of 250 literal sentences. Human annotators are then asked to classify the sentences. A sentence is classified as positive if: (i) it carries the intended figurative expression, and (ii) original meaning of the sentence remains intact.

Table 2 shows the result of the human evaluation. Our framework consistently outperforms the general method in generating the figurative expressions from a literal text. LLMs particularly struggle with generating metonymic variations, as only 38.8% of the sentences generated using a general method were judged as metonymy, compared to 75.2% of the sentences generated by our framework. Overall, our framework has a consistently better performance, with 84.0% of the sentences labeled as metaphors, and 74.0% containing both metonymy and metaphor.

	Metonymy	Metaphor	Hybrid
General	38.8%	70.8%	49.2%
Ours	**75.2%**	**84.0%**	**74.0%**

Table 2: Percentage of sentences evaluated by humans to have the intended figurative expression in a sample of 250. General row are sentences from basic prompting. The sentences generated by our framework are significant better with figurative expressions.

	Metonymy	Metaphor	Hybrid
General	0.70	0.60	0.44
Ours	**0.84**	**0.82**	**0.70**

Table 3: Cosine similarity score between the original literal sentence and the generated variant using a sentence transformer. Our framework better preserves the semantic meaning of the sentence.

Automatic Evaluation. While the human evaluation showed the percentage of sentences judged to have the intended expression, an important aspect of transforming literal expressions to their figurative variations is that the generated sentences must be faithful to the input (Chakrabarty et al., 2021). To evaluate this criteria, we use sentence transformer (Reimers and Gurevych, 2019) to calculate the semantic similarity between the generated sentences and the original input sentence. Table 3 shows the result of this evaluation. Metonymic sentences have higher semantic similarity with the literal expression, followed by metaphor, while hybrid expressions have the least. The sentences generated by our framework show significantly higher semantic similarity than the general method. The sentences generated by general prompting (without any structural guidance) are semantically different from the literal text, with the LLM altering its structure, which often leads to loss or change in meaning.

Overall, the human and automatic evaluation on 250 samples shows that naive prompting struggles to generate metonymic and hybrid sentences from literal texts. The semantic structure of the gen-

Literal Input	Metonymy	Metaphor	Hybrid
The researchers formed a rational statement of his question.	The **laboratory** formed a rational statement of his question.	The researchers **sculpted** a rational statement of his question.	The **laboratory** **sculpted** a rational statement of his question.
The reporter couldn't have done too good a job on you.	The **newsroom** couldn't have done too good a job on you.	The reporter **butchered** you in that interview.	The **newsroom** **butchered** you in that interview.
The Queen officially pronounced Turing pardoned in august 2014.	**Buckingham Palace** officially pronounced Turing pardoned in august 2014.	The Queen's pardon for Turing **thundered** through history in August 2014.	**Buckingham Palace's** pardon for Turing **thundered** through history in August 2014.
The police exposed the crime ring in 1956.	**NYPD** exposed the crime ring in 1956.	The police **unearthed** the crime ring in 1956.	**NYPD** **unearthed** the crime ring in 1956.

Table 4: Example of a literal expression and its respective figurative variations from MetFuse dataset.

	Metonymy	Metaphor	Hybrid
Fluency	3.30	**3.64**	3.61
Meaning	**3.51**	3.10	2.74
Creativity	2.95	4.01	**4.25**
Metonymicity	3.47	-	**3.65**
Metaphoricity	-	**3.95**	3.82

Table 5: Human score on five criteria from 250 samples from the MetFuse dataset.

erated text also shifts to the degree of a loss or alteration of meaning. Our framework significantly outperforms the baseline, generating intended figurative variations of the literal text that are also semantically similar to its source.

4.2 MetFuse Dataset

Leveraging our figurative-to-literal framework, we construct the MetFuse dataset. To build this dataset, human annotators reviewed 1,500 samples generated by our framework. Of these, 1,104 were judged to contain all three intended figurative variants, with around 74% accuracy, as shown in Table 2. From this pool, we randomly sampled 1,000 instances to form the final MetFuse dataset. The final dataset contains 1,000 literal sentences, each paired with its metonymic, metaphoric and hybrid variants, resulting in a total of 4,000 sentences. The inter-annotators score between the annotators, measured as the raw agreement was measured to be 96.3% for metonymy, 91.2% for metaphor and 91.1% for hybrid sentences. Table 4 shows some illustrative examples from the dataset.

Human Score. The MetFuse dataset provides researchers the unique opportunity to study interaction and entanglement of metonymy and metaphor within a single expression. To this end, we recruit human annotators to rate the figurative expressions from the MetFuse dataset. Each annotator was provided with a sample and asked to rate the figurative texts on a scale of 1 to 5, with the literal sentences as references. We used four criteria from Chakrabarty et al. (2021), in addition to one of ours: (1) *Fluency* ("How fluent, grammatical, well formed and easy to understand are the generated utterances?"), (2) *Meaning* ("Are the input and the output referring or meaning the same thing?") (3) *Creativity* ("How creative are the generated utterances?"), (4) *Metonymicity* ("How explicit is the metonymy?") and (5) *Metaphoricity* ("How explicit is the metaphor?").

Table 5 shows the human scores. Metaphor sentences are rated to have the highest fluency. Metonymy is judged as less creative and fluent, but they best conserve the meaning. The fluency, meaning, and creativity ratings corroborate with previous linguistic works that states that metonymy is structurally limited and play a crucial role in meaning preservation (Radden and Kövecses, 1999; Ruiz de Mendoza and Baicchi, 2005), while metaphor is more nonrestrictive (Kövecses, 2010), leading to higher creativity and lower meaning preservation. Hybrids have the highest creativity score due to both the noun and verb being altered in a figurative way. But this tends to lose the implicit meaning, having the lowest meaning score. Table 5 also shows that metaphor sentences were judged to have higher metaphoricity (how explicit is the metaphor), slightly edging out hybrid sentences. Interestingly, hybrid sentences were rated as having noticeably higher metonymicity (how explicit is the metonymy) over metonymic sentences. The results indicates that the human judges found the metonymy to be more explicit in the sentences when it was accompanied by a verbal metaphor.

5 Analysis

In this section, we analyze the MetFuse dataset through an extrinsic evaluation, examining how metaphoric verbs influence metonymic noun.

	Train	Train +MTF$_{mty}$	Train +MTF$_{hyb}$
ConMeC	75.49	76.71 (+1.22)	**79.33** (+3.84)
Pedinotti	68.42	66.92 (-1.50)	**70.44** (+2.02)
RelocaR	67.33	69.99 (+2.66)	**70.67** (+3.34)
WiMCor	81.67	82.33 (+0.66)	**82.67** (+1.00)

Table 6: Test accuracy for downstream tasks on metonymy datasets. Each dataset is split in a 70-30 train-test ratio. Train = original training sample, Train+MTF$_{mty}$ = original training sample augmented with metonymic examples from MetFuse, Train+MTF$_{hyb}$ = original training sample augmented with hybrid examples from MetFuse.

	Train	Train +MTF$_{mtr}$	Train +MTF$_{hyb}$
VUA Verb	64.38	64.53 (+0.15)	**66.55** (+2.17)
Flute	76.26	**79.73** (+3.47)	78.47 (+2.21)
MOH-X	76.26	76.92 (+0.16)	**77.43** (+0.51)
TroFi	60.42	**63.30** (+2.88)	61.76 (+1.34)

Table 7: Test accuracy for downstream tasks on metaphor datasets. Each dataset is split in a 70-30 train-test ratio. Train = original training sample, Train+MTF$_{mty}$ = original training sample augmented with metaphor examples from MetFuse, Train+MTF$_{hyb}$ = original training sample augmented with hybrid examples from MetFuse.

5.1 Extrinsic Evaluation

While the MetFuse dataset is helpful for analyzing how metonymy and metaphor interact together, it can also be used for other tasks. To showcase this, we perform an extrinsic evaluation. Specifically, we examine if our dataset can improve metaphor and metonymy classification using existing datasets via data augmentation.

Datasets. We use four existing metonymy datasets: two common noun metonymy datasets—ConMeC (Ghosh and Jiang, 2025) and Pedinotti and Lenci (2020), and two named entity datasets—RelocaR (Gritta et al., 2017) and WiMCor (Alex Mathews and Strube, 2020). For metaphor, we use four verbal datasets as well: VUA Verb (Steen, 2010), Flute (Chakrabarty et al., 2022b), MOH-X (Mohammad et al., 2016), and TroFi (Birke and Sarkar, 2007). All eight datasets are binary classification tasks—a system should determine whether the given sentence contains a figure of speech or not.

Experimental Setup. For a given dataset, we use a 70–30 train–test split and fine-tune BERT-base (Devlin et al., 2019) under three settings: (i) the original training samples (Train), (ii) the training set augmented with MetFuse figurative examples (Train + MTF$_{mty}$, or Train + MTF$_{mtr}$), and (iii) the training set augmented with MetFuse hybrid examples. For the metonymy classification task, the figurative examples in (ii) are metonymic variants, whereas for the metaphor downstream task, the same setup is followed except that the added samples are metaphoric variants. In all cases, the MetFuse augmentation size is fixed at 50% of the original training set. We fine-tune BERT for 3 epochs with a learning rate 1e-5 and a batch size 8.

Metonymy results. Table 6 presents the results of the downstream experiment on metonymy datasets. Test accuracy improves when the training set is augmented with metonymic samples from MetFuse in three out of the four datasets, with only Pedinotti and Lenci (2020) being the exception. Notably, across all datasets, the highest test accuracy is achieved when the training set is augmented with hybrid examples. This suggests that the model learns the metonymic usage of a noun more effectively when exposed to a co-occurring metaphoric verb. This pattern aligns with the findings in Table 2, where human annotators also found metonymic nouns easier to identify when paired with metaphoric verbs.

Metaphor. Table 7 presents the results of the metaphor downstream experiment. Test accuracy improves consistently across all four datasets when the original training data is augmented with the MetFuse samples, underscoring the usefulness and generalizability of MetFuse. Hybrid examples yield the best performance on VUA Verb and MOH-X, whereas metaphoric examples perform better on Flute and TroFi. This suggests that the influence of metonymy on metaphor is not uniform—unlike the more consistent effect of metaphor on metonymy. In some cases, hybrid constructions dominate, while in others, purely metaphoric examples are more effective.

5.2 Metaphor Improves Metonymy

Our previous results highlighted that both humans (Table 5) and supervised BERT (Table 6) found it easier to identify the metonymic usage of a noun in hybrid sentences, i.e., when the metonymic noun was paired with a metaphoric verb. We investigate

Model	Metonymy			Hybrid		
	Pre	Rec	F1	Pre	Rec	F1
GPT-OSS-20B	95.8	51.9	67.3	**96.1**	**57.0**	**71.6**
Qwen3-30B	79.3	92.5	85.4	**79.9**	**96.3**	**87.3**
Llama-3.1-70B	85.6	95.8	90.4	**85.9**	**97.5**	**91.3**
Gemini-2.5	92.4	95.4	93.9	**92.5**	**97.0**	**94.7**

Table 8: F1 score of positive metonymic sentences under two conditions: using metonymy only sentences as positive sentences, and using hybrid sentences as positive sentences. Literal sentences are the negative sentences.

	$sim(\mathrm{N}_{lit}, \mathrm{N}_{mty})$	$sim(\mathrm{N}_{lit}, \mathrm{N}_{hyb})$
GPT-OSS-20B	71.00	**72.38** (+1.38)
Qwen3-30B	90.88	**91.75** (+0.87)
Llama-3.1-70B	57.78	**59.64** (+1.86)
BERT	65.42	**65.62** (+0.20)

Table 9: Similarity score between the contextual embeddings of the noun tokens. $sim(\mathrm{N}_{lit}, \mathrm{N}_{mty})$ = similarity between the noun in the literal sentence and metonymic sentence. $sim(\mathrm{N}_{lit}, \mathrm{N}_{hyb})$ = similarity between the noun in the literal sentence and hybrid sentence.

this further by performing metonymy resolution task using LLMs. We use two setups designed to test how the presence of metaphor influences metonymy identification: (i) treating metonymic sentences as positive cases and their literal counterparts as negative, and (ii) treating hybrid sentences as positive cases and their literal counterparts as negative. For each input, we provide the sentence and the noun to the LLM and ask if the noun is used metonymically (Ghosh and Jiang, 2025). We use four state-of-the-art models: GPT-OSS-20B (OpenAI et al., 2025), Qwen3-30B (Yang et al., 2025), Llama-3.1-70B (Grattafiori et al., 2024), and Gemini-2.5-Flash (Comanici et al., 2025).

Table 8 shows the results. Across all four models, the precision, recall and F1 scores are consistently higher for hybrid samples than for purely metonymic ones. This suggests that LLMs can easily identify the metonymic usage of the noun when it co-occurs with a metaphoric verb.

We investigate further with the token embeddings to check if similar patterns persist. We take the contextual embeddings of the nouns in the literal sentence (N_{lit}) and embeddings of the metonymic nouns in the metonymic sentences (N_{mty}) and calculate the cosine similarity between them. This effectively tells us how similarly the nouns are used in the literal and metonymic sentences. Similarly, we find the embeddings of the metonymic nouns in the hybrid sentences (N_{hyb}) and calculate the similarity with (N_{lit}). This tells us how similarly the nouns are used in the literal and hybrid sentences. We then compare $sim(\mathrm{N}_{lit}, \mathrm{N}_{mty})$ with $sim(\mathrm{N}_{lit}, \mathrm{N}_{hyb})$.

Table 9 shows the results. $sim(\mathrm{N}_{lit}, \mathrm{N}_{hyb})$ consistently has a higher value than $sim(\mathrm{N}_{lit}, \mathrm{N}_{mty})$. This means according to the LLM contextual embeddings, the same noun in the hybrid sentence is more similar to its literal sentence counterpart than in metonymic sentence. This suggests the metonymy in the hybrid sentence is more explicit as its embeddings are more similar to that of a non-metonymic usage. This corroborates with the human judgement in Table 5.

Qualitative Discussion. Our empirical results show that metonymy's strength tends to increase when paired with a metaphor. We look at this from a purely linguistic and cognitive point of view. In a purely metonymic sentence such as "The *newsroom* was harsh on the actor,' the noun *newsroom* can remain cognitively unresolved due to metonymy's single domain mapping nature. Readers may or may not consciously resolve it to *journalists*, since both the literal (place) and metonymic (people inside) readings remain available. When a metaphoric verb is introduced, the interpretive dynamics change. In "*The newsroom butchered the actor,*" a verb like "*butchered*" belongs to a semantic domain of physical violence, carrying strong selectional preferences for an animate, agentive subject. Because *newsroom* is not literally animate, the metaphor exerts pressure on the reader to resolve the metonymy.

In this way, the out-of-domain mapping introduced by the metaphor forces explicit metonymy resolution, making the metonymy more salient than it would be in isolation. Thus, metaphor functions as a forcing device: its cross-domain mapping imposes constraints that push the metonymic noun into an explicit, agentive reading. This explains why in hybrid cases, metonymy is often perceived as more prominent and harder to ignore than in purely metonymic sentences. As a comparison, further analysis in Appendix B shows that the co-occurrence of metonymy does not impact metaphor's performance consistently.

Takeaway. Our analysis shows that humans and LLMs agree on one thing: *a metaphoric verb can strengthen the metonymic nouns strength* if they co-occur in a sentence with a dependency relation.

Type	Original Sentence	Generated Sentence
Metonymy	(1) His **guitarist** realizes what he did and knocks him out.	His **chord** realizes what he did and knocks him out.
	(2) The reporters questioned, "why would a **cricketer** do this?"	The reporters questioned, "why would a **bat** do this?"
	(3) His **father** guided him in his early year.	His **wisdom** guided him in his early years.
	(4) The **player** was furious at the referee.	The **athlete** was furious at the referee.
	(5) The **teacher** encouraged her to apply for the position.	The **staff at the school** encouraged her to apply for the position.
Metaphor	(6) A dancer should **watch** her diet carefully.	A dancer should **obsess** over her diet carefully.
	(7) A great menacing student **warns** her not to trust her family.	A great menacing student **poisoned her mind**.
	(8) The economist **described** the "battle for the net".	The economist **waged a war**.
Hybrid	(9) The **painters worked** on this masterpiece for 11 years.	The **studio forged** this masterpiece for 11 years.

Table 10: Error analysis of the generated sentences using our framework. Words highlighted in **blue** in the original sentence are the target words that are being altered. The **red** words are the cause of the errors in the generated sentence.

5.3 Error Analysis

Table 10 shows some qualitative examples of the error cases during generating metonymic, metaphoric and hybrid expressions using our framework. Sentence (1) and (2) are examples of the major error cases in generating metonymy, caused by the semantic structure of the original sentence. These instances of nouns renders itself difficult to be transformed to a metonymic variant, often yielding no natural metonymic replacements. In sentence (3), the metonymic substitution (*father* $\rightarrow$ *wisdom*) alters the meaning of the original sentence. Sentence (4) is an instance of a literal substitution, an error made by Llama when generating metonymic substitutions. Sentence (5) is an example of the error occurring due to LLM paraphrasing after substituting the noun. In this case, Llama correctly generated a metonymic noun substitution (*teacher* $\rightarrow$ *school*). However, when the LLM was asked to refine the already metonymic sentence "*The school encouraged her to apply for the position*," the paraphrasing added the clause *staff at the school*, rendering the sentence non-metonymic.

In sentence (6), the annotators agreed the verb is non-metaphoric. Sentence (7) and (8) are the major error cases in metaphor generation, with the meaning of the paraphrased sentence being inherently different from the literal source. In sentence (9), replacing the noun and the verb makes the hybrid expression creative, but the annotators agree that the sentence loses some of the intended meaning as the literal one. The literal expression refers to the art of painting, but one cannot understand the context just by looking at the hybrid expression (it can refer to any type of art, such as painting, singing, music, or sculpture).

6 Conclusion

We introduced a framework that transforms a literal sentence into metonymic, metaphoric, and hybrid variants while preserving the meaning, and used it to construct MetFuse, the first dedicated dataset of figurative fusion between metonymy and metaphor. MetFuse contains 1,000 meaning-aligned quadruplets totaling 4,000 sentences. Across eight existing benchmarks, augmenting training data with MetFuse consistently improves both metonymy and metaphor classification. Our analysis further shows that both humans and large language models identify metonymy more easily when a metaphoric verb is present, suggesting that the metaphor's cross-domain mapping forces a more explicit reading of the co-occurring metonymic noun. We hope MetFuse will enable further study of how metonymy and metaphor interact in context.

Limitations

While our study makes important progress in addressing metonymy and metaphor interaction within NLP, certain limitations remain.

Our current work does not encompass the full range of metonymy that occurs in natural language. We limit ourselves to the animate subject nouns leading to location-for-people or institution-for-people metonymy, focusing on instances that are common and frequent. This choice allows us to build a strong foundation while avoiding excessive fragmentation of the problem space. We leave other

metonymy types for future exploration.

In our experiments, we do not explicitly compute or annotate the domain mappings for both metonymy and metaphor. Our goal is to lay the groundwork for metonymy-metaphor interaction study without constraining them to a fixed inventory of conceptual domain. We hope our work inspires future research to computationally study the domain mapping of these linguistic phenomena.

Ethical Considerations

In this work, we employed large language models (LLMs) to generate candidate nouns and verbs for constructing metonymic, metaphoric, and hybrid expressions. For named-entity metonymy, in particular, the LLM was prompted to suggest location-based entities (e.g., cities, institutions) as replacements for target nouns. While this procedure leverages the generative capacity of LLMs, we recognize that such models may reproduce unintended biases or stereotypes, especially when dealing with named entities (e.g., associating certain locations with actions). We did not observe such issues in our generated samples; however, we acknowledge that these risks are possible. Importantly, our dataset does not include any private or personally identifiable information: all named entities are in the public domain. By acknowledging these limitations and safeguards, we emphasize that our use of LLMs is confined to controlled generation for research purposes, with human oversight applied to ensure that the final dataset avoids harmful content.

Acknowledgments

We sincerely thank the CincyNLP group for their valuable feedback and help in annotation. We also thank the anonymous ARR reviewers for their insightful suggestions and discussions.

References

Keiga Abe, Kayo Sakamoto, and Masanori Nakagawa. 2006. A computational model of metaphor generation process. In *Annual Meeting of the Cognitive Science Society*.

Kevin Alex Mathews and Michael Strube. 2020. A large harvested corpus of location metonymy. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*.

Antonio Barcelona. 2003. *Metaphor and Metonymy at the Crossroads: A Cognitive Perspective*. Mouton de Gruyter.

Julia Birke and Anoop Sarkar. 2007. Active learning for the identification of nonliteral language. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 21–28, Rochester, New York. Association for Computational Linguistics.

David D. Clarke Brigitte Nerlich. 2001. Ambiguities we live by: towards a pragmatics of polysemy. *Journal of Pragmatics*, 33:1–20.

Jose Camacho-collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez Cámara. 2022. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2022)*.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2021)*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, and others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*.

Saptarshi Ghosh and Tianyu Jiang. 2025. ConMeC: A dataset for metonymy resolution with common

nouns. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2025)*.

Saptarshi Ghosh, Linfeng Liu, and Tianyu Jiang. 2026. A computational approach to visual metonymy. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*.

Louis Goossens. 1990. Metaphtonymy: The interaction of metaphor and metonymy in expressions for linguistic action. *Cognitive Linguistics*, 1(3):323–342.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, and others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2017. Vancouver welcomes you! minimalist location metonymy resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*.

Alberto Morón Hernández. 2017. Paraphrasing verbal metonymy through computational methods. *Preprint*, arXiv:1709.06162.

Javier Herrero-Ruiz. 2004. Panther, k-u., and thornburg. eds. 2003. "metonymy and pragmatic inferencing". amsterdam/philadelphia: John benjamins. 280 pp. *Journal of English Studies*, 4:237.

Maria Jodlowiec and Agnieszka Piskorska. 2015. Metonymy revisited: Towards a new relevance-theoretic account. *Intercultural Pragmatics*, 12.

Zoltán Kövecses. 2010. Metaphor and culture. 2:197–220.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Katja Markert and Malvina Nissim. 2007. SemEval-2007 task 08: Metonymy resolution at SemEval-2007. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval 2007)*.

Rowan Hall Maudslay, Simone Teufel, Francis Bond, and James Pustejovsky. 2024. ChainNet: Structured metaphor and metonymy in WordNet. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Vivi Nastase and Michael Strube. 2009. Combining collocations, lexical and encyclopedic knowledge for metonymy resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*.

Geoffrey Nunberg. 1995. Transfers of meaning. *Journal of Semantics - J SEMANT*, 12:109–132.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, and others. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Anna Papafragou. 1996. On metonymy. *Lingua*, 99(4):169–195.

Paolo Pedinotti and Alessandro Lenci. 2020. Don't invite BERT to drink a bottle: Modeling the interpretation of metonymies using BERT and distributional representations. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*.

Günter Radden and Zoltán Kövecses. 1999. Towards a theory of metonymy. *Metonymy in Language and Thought*, pages 17–59.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*.

Francisco Ruiz de Mendoza. 2002. *Metonymy, Grammar, and Communication*.

Francisco Ruiz de Mendoza and Annalisa Baicchi. 2005. *Cognitive Linguistics: Internal dynamics and interdisciplinary interaction*.

G. Steen. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging evidence in language and communication research. John Benjamins Publishing Company.

Kevin Stowe, Nils Beck, and Iryna Gurevych. 2021a. Exploring metaphoric paraphrase generation. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL 2021)*.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021b. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th*

International Joint Conference on Natural Language Processing (ACL 2021).

Asuka Terai and Masanori Nakagawa. 2010. A computational system of metaphor generation with evaluation mechanism. In *Proceedings of the 20th International Conference on Artificial Neural Networks*, Berlin, Heidelberg. Springer-Verlag.

Yufei Tian, Arvind krishna Sridhar, and Nanyun Peng. 2021. HypoGen: Hyperbole generation with commonsense and counterfactual knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021 (Findings of EMNLP 2021)*.

Mark Turner and Gilles Fauconnier. 2002. *The Way We Think: Conceptual Blending And The Mind's Hidden Complexities*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, and others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zhiwei Yu and Xiaojun Wan. 2019. How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL 2019)*.

A Generalization of our Framework

To assess how well our framework generalizes across different LLMs, we use human to annotate 100 quadruplets generated by different models. Table 11 shows the results, which highlights that our framework generalizes well across models of all sizes. The success rate of the framework depends primarily on the semantic structure of the input sentence, and not on the LLMs capability. Due to this, we chose Llama 3.1-8B to as the base model in our framework, as it keeps our model lightweight without compromising performance.

	Metonymy	**Metaphor**	**Hybrid**
Llama-3.1-8B	74	86	73
GPT-oss-20B	74	87	74
Qwen3-30B	**75**	86	**75**
Llama-3.1-70B	**75**	87	74
GPT-5	72	**90**	72

Table 11: Number of sentences evaluated by humans to have the intended figurative expression in a sample of 100 when ran on our framework with different LLMs.

A.1 Diversity

An important factor in generating the metonymic replacements of the noun is diversity. Open-source LLMs like Llama and Qwen are more suited to our framework as it gives the user flexibility in terms of diversity by increasing temperature and top-p value. While being the latest model in the bunch, GPT-5 suffered from repetitive replacements. We ran experiments with different "thinking effort" conditions, but the repetitiveness did not subside.

B Metonymy's Impact on Metaphor

To analyze the impact of metonymy on metaphors, we repeat the token embedding similarity experiment conducted in Section 5.2, but this time, we compare the embeddings of the verb. Specifically, we find how similar the verbs are in the literal and metaphoric sentence. We calculate the similarity between the verb in the literal sentence and the metaphoric sentence $sim(\mathrm{V}_{lit}, \mathrm{V}_{mtr})$, and compare it with the verb in the literal sentence and hybrid sentence $sim(\mathrm{V}_{lit}, \mathrm{V}_{hyb})$.

Table 12 shows the results. The results are less consistent than Table 9. GPT-oss-20B and Qwen3-30B shows higher similarity between the verb in the literal-hybrid pair, while Llama 3.1-70B and BERT shows higher similarity in literal-metaphor

	$sim(\mathrm{V}_{lit}, \mathrm{V}_{mtr})$	$sim(\mathrm{V}_{lit}, \mathrm{V}_{hyb})$
GPT-oss-20B	70.19	**70.28** (+0.09)
Qwen3-30B	92.83	**93.07** (+0.24)
Llama-3.1-70B	**57.32**	52.88 (-4.44)
BERT	**65.24**	65.02 (-0.22)

Table 12: Similarity score between the contextual embeddings of the verb tokens. $sim(\mathrm{V}_{lit}, \mathrm{V}_{mtr})$ = similarity between the verb and in the literal sentence and metaphoric sentence. $sim(\mathrm{N}_{lit}, \mathrm{N}_{hyb})$ = similarity between the verb and in the literal sentence and hybrid sentence.

pair. This observation aligns with the metaphor downstream experiment results in Table 7. The impact of metonymic noun on the metaphor verb is not consistent. We believe there are deeper semantic complexities at play, and we leave this for future work.

Model	**Similarity**		**Entailment**	
	Mtr	Hyb	Mtr	Hyb
GPT-OSS-20B	**86.45**	84.82	**96.03**	93.84
Qwen3-30B	**84.79**	83.34	**94.54**	94.31
Llama-3.1-70B	**82.02**	78.59	**85.94**	79.40
Gemini-2.5	**86.31**	82.13	**93.24**	90.96

Table 13: Semantic similarity and entailment scores of metaphor (Mtr) and hybrid (Hyb) sentences with respect to their literal counterparts.

We also investigate how the presence of a metonymic noun impacts the interpretability of the metaphor, i.e., if LLMs can recover the literal meaning from the metaphor expression (Chakrabarty et al., 2022a,b). For this task, we provide the metaphor and hybrid sentence to the LLM and ask it to paraphrase them to their literal meaning. We then evaluate the paraphrased outputs using entailment (Liu et al., 2019) and semantic similarity scores. Table 13 shows the results. We find that metaphor-only sentences achieve equal or higher similarity and entailment scores compared to hybrid sentences. This suggests that an LLMs ability to interpret a metaphor may decrease when the metaphor is anchored to a metonymic noun.

C Additional Metonymy Downstream Results

Metonymy identification have been shown to struggle under cross-domain settings (Ghosh and Jiang, 2025). To this end, we employ the MetFuse dataset to study its impact in cross-domain metonymy classification. For this experiment, we train BERT on one dataset, and test it on another. We separate the two common noun datasets: ConMeC (Ghosh and Jiang, 2025) and Pedinotti and Lenci (2020), and the two named entity datasets: RelocaR (Gritta et al., 2017) and WimCoR (Nastase and Strube, 2009). When the model is trained on a common noun dataset, it is tested on a common noun dataset as well, and augmented with common noun examples from MetFuse. This goes the same for named entity dataset as well. This is done to ensure a fair setting, as MetFuse has both common noun and named entity examples. We keep the same conditions as before—the augmented MetFuse examples are 50% the size of the original training sample, learning rate was set at 1e-5 over 3 epochs and batch size 8.

Figure 2 shows the result of the cross-domain classification experiment. The test accuracy increases when the model is trained with the MetFuse dataset in all three cases, the only exception being when it is trained with RelocaR and tested on WimCoR. The hybrid examples from MetFuse always has a better performance than purely metonymic examples. The results highlight the usefulness and generalizability nature of the MetFuse dataset. It also shows the metonymic noun can be identified more easily by language models when it is accompanied by a metaphor, supporting the observation made in the main paper.

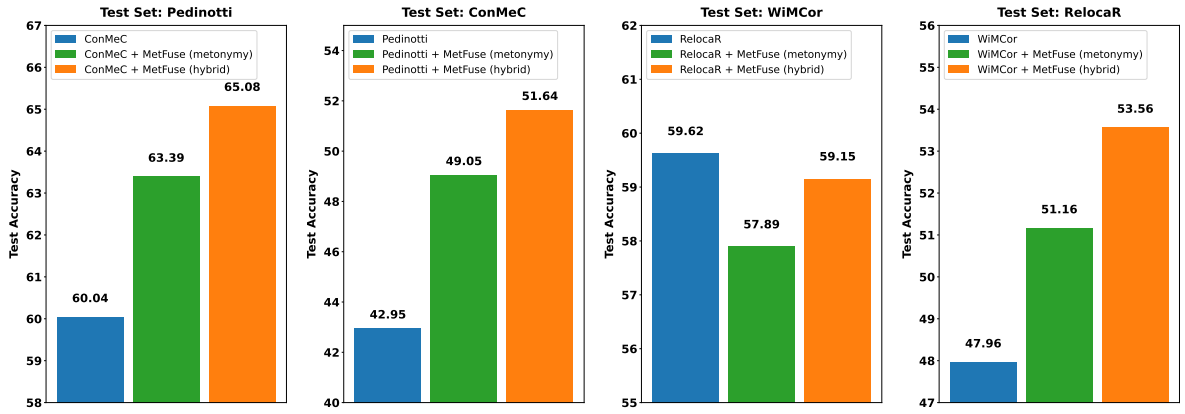
D Surprisal Scores

For further analysis, we calculate the token surprisal scores from the MetFuse dataset. For a token x with model probability $p(x)$, surprisal score is given by:

$$s(x) = -\log p(x).$$

Higher surprisal score indicates the model was not expecting this token, hence is "*surprised*". For the literal, metonymy, metaphor and hybrid sentences, we calculate the surprisal of the noun token (responsible for metonymy) and the verb token (responsible for metaphor).

Table 14 shows the results of this experiment. The bold numbers indicate the instances where the word was changed by our framework to create the figurative expression. It is evident that the surprisal for the noun is high in metonymy and hybrid, while the surprisal for verb is high in metaphor and hybrid.



(a) Train: ConMeC (b) Train: Pedinotti (c) Train: RelocaR (d) Train: WiMCor

Figure 2: Figure showing results of the downstream experiment. Blue bar shows the performance when trained on an existing metonymy dataset. Green bar shows the performance when trained on existing dataset + MetFuse metonymic samples, orange bar is existing datatset + MetFuse hybrid samples.

Surprisal	Lit	Mty	Mtr	Hyb
Noun token	9.01	**12.79**	9.65	**12.81**
Verb token	7.03	9.02	**11.38**	**12.66**

Table 14: Surprisal scores for the noun and verb token in literal (lit), metonymy (Mty), metaphor (Mtr) and hybrid (Hyb) sentences. Bold indicates the token was altered to create the intended figurative expression.

E Discussion - Metonymy vs Metaphor Generation

Our findings highlight an important asymmetry between metaphor and metonymy generation. Metonymy is significantly harder to generate in a controlled fashion. Its constraints come from the fact that metonymic substitutions are restricted to in-domain mappings (Goossens, 1990), which sharply narrows the candidate space. Moreover, the substituted expression must still refer to the same underlying entity, a requirement that is not always straightforward to satisfy For example, "*his father guided him*" cannot simply be replaced by his father's hand guided him without changing the referent).

In contrast, metaphor generation is comparatively more permissive (Kövecses, 2010). Because metaphors involve cross-domain mappings, a wider range of substitutions are tolerated, even in uncontrolled generation. While many generated metaphors may be novel or unconventional, they still tend to preserve intelligibility without the strict referential constraints that metonymy demands.

These observations suggest that while metaphor generation can often succeed through broad lexical substitution, metonymy requires more fine-grained semantic control and discourse awareness. Evidently, table 2 shows 75.2% of the sentences were judged to be metonymic by humans, compared to 84.0% judged to be metaphors. A deeper investigation into when hybrids (metaphor–metonymy blends) outperform pure metaphors or pure metonymies is left for future work.