Benchmarking Metaphor Generation Performance of Modern Large Language Transformer Models

Ada He, Carmen He

Cognitive Science Department, University of California San Diego

ABSTRACT

Multi-headed self attentional transformers, first proposed in 2017, have revolutionized Natural Language Processing (NLP) and resulted in the rise of ubiquitous Large Language Models (LLMs) in recent years. The increasingly sophisticated abilities of LLMs to replicate human language and writing has been a source of concern for many writers and scholars, who fear that these models may threaten human creativity. Metaphors, which capture complex information by mapping seemingly unrelated concrete qualities to abstract phenomena, are an important cornerstone of human language intuition, and provide a lens to analyze the creative language abilities of LLMs. Using the theoretical framework of Cognitive Metaphor Theory (CMT), we tested the metaphor generation capabilities of GPT, DeepSeek, and Gemini with several sentence paraphrasing tasks. Each model was then prompted to rate the fluency, similarity to input sentence, and metaphoricity (abstractness) of the output metaphors. Overall, we found that while models were strong in creating grammatically correct paraphrases that aligned with prompt requirements, their average metaphoricity ratings varied, with Gemini scoring noticeably lower than GPT and DeepSeek in every evaluation category.

1. INTRODUCTION

Metaphors are a pervasive linguistic phenomena critical to human thought and interaction across every culture. In fact, metaphors are so natural to daily communication that individuals often use them without even being aware that they are mapping an abstract event to non-literal concrete experiences. Therefore, the tasks of metaphor recognition, interpretation, and generation are challenging for even humans, not to mention deep learning networks. At the same time, the development and continued improvement of transformer deep learning networks in recent years has rapidly advanced NLP, resulting in a profusion of LLMs that are increasingly sophisticated at “thinking” and “communicating” like real people. As humans seem able to quickly create and understand complex metaphors, assessing performance of state-of-art LLMs on the difficult tasks of metaphor detection, interpretation, and generation offers an approach to operationalize how well transformer-based generative AI has come to demonstrate human natural creative language.

Our experiment focuses specifically on the metaphor generation capabilities of three popular pretrained models – Open AI's GPT-4o mini, DeepSeek V3, and Google's Gemini 2.0 – in a set of paraphrasing tasks. Given a purely literal input sentence, the model is asked to "translate" the original string to a sentence containing a relevant metaphor. Since the definition of a metaphor can vary widely, different perspectives on what makes a metaphor changes the perceived frequency of metaphors in linguistic expressions[1]. To standardize what counts as a metaphor, we apply the cognitive linguistic framework of Conceptual Metaphor Theory (CMT), which defines metaphors as the use of a concrete *source* experience to capture an abstract *target* domain[2]. We chose CMT as our paradigm because it may represent how metaphors are interpreted in the human mind, and it has become the dominant metaphor representation framework since its introduction in 1980 (Lakoff & Johnson 1980). In general, having a standardized definition for metaphors allows us to determine whether models have generated accurate metaphors, and perform evaluations of metaphor quality that we can use to compare the models later.

Finally, we evaluate the strength of each LLM's output metaphors several ways. First, in place of traditional crowdsourcing methods to assess each metaphor's grammatical correctness, semantic meaning preservation from the input, and strength of metaphor creativity, we prompt each model to rate the other models' metaphors in each of our evaluation categories.

2. RELATED WORK

2.1 Metaphoric Paraphrase Generation

Precedence for Transformer model metaphoric paraphrase generation under CMT (Stowe et al. 2021) inspires our experimental design. We adopt the established two method approach to paraphrasing: **free** generation, in which the source domain is not specified, and **controlled** generation, where a source domain is specified. This provides insight on whether deliberate prompt engineering for LLMs improves model "creativity". We also adopt similar automatic evaluation metrics for the output metaphors.

Our project deviates from existing research that explores transformer models tuned specifically for paraphrasing tasks (Goyal & Durrett, 2020) by assessing the performance of widely used LLMs, which, to our knowledge, has not been studied closely.

[1] For instance, in the phrase "defending an argument" some might identify "defending" to be a metaphor mapping the domain of combat to the domain of persuasion and logic, while others may argue that "defense" is not referring to combat but describes a literal action that aligns with "argument".

[2] Consider the example: "He **exploded** with rage", where the metaphor "*exploded"* falls under the mapped domains "Anger (*target*) is fire (*source*)". (Kövecses, 2017)

2.2 Benchmarking AI Assistants

A landscape review of AI Research Assistants that used one of the assessed AI models to rate the responses of all models, including its own (Worsham, 2025), inspired us to replace traditional crowdsourcing approaches to rating metaphors, which can be costly, by having each LLM evaluate the metaphors generated by itself and other models. This expands off Worsham's work with a more robust assessment of whether each LLM is biased towards its own responses.

3 DATA

We use the evaluation pairs from the dataset constructed by Stowe et al. 2021. The original evaluation set consists of fully literal input sentences with the target domain, source domain for controlled generation, index location of the lemma to be paraphrased (assuming 0-based indexing), the grammatical type of the phrase, and the output sentences and output evaluations generated by their models. These pairs are evenly divided into 50 broad metaphor domains and 50 narrow metaphor domains, the distinction of which is a theoretical consideration outside of the scope of our project.

Then, the only additional processing we had to perform was to combine the two sets, dropping all columns not useful to our experiments, and randomly shuffle the data to create a set of 100 samples containing just the target, source, literal input, and indexes of the words to paraphrase.

4 METHODS

4.1 Models

We selected three LLMs that are at the center of the cultural discussion regarding ubiquitous generative AI. All utilize transformer-based models introduced in the landmark paper "Attention is All You Need" (Vaswani et al. 2017), which extracts salient features from tokenized input prompts via multi-head self attention.

GPT-4o mini (ChatGPT) Perhaps the most widely used LLM in the United States, it was imperative for us to test the newest free model. OpenAI has not disclosed the full details of how this model differs from earlier models, except having been trained on both textual and visual data. In general, the latest GPT models employ a traditional decoder-only transformer model with around ~175 billion parameters (Wagh 2023).

DeepSeek v3 Making cultural shock waves due to its lower computational cost and reportedly equal performance to GPT, few studies have been made assessing the "creative" language generation abilities of DeepSeek compared to other models. Like GPT, DeepSeek utilizes transformer blocks with multi-headed self attention layers. However, it deviates from the standard decoder-only transformer model by modifying the feed-forward neural network within each block using a Mixture of Experts (MoE) model. This reduces computational cost by dividing each task into subtasks and only activating necessary subsets of "experts" (which are just smaller feed-forward neural nets with independent parameters) to address each task (Li 2025).

Gemini 2.0 Google has not disclosed many details regarding Gemini's architecture, but experts estimate that it employs an encoder-decoder transformer model, which utilizes a bidirectional attention instead of the casual attention of decoder-only structures. However, there is no conclusive information to date that indicates a significant difference in performance between decoder-only and encoder-only architectures in NLP, so it remains unclear which model should be preferred for our paraphrasing task before running the experiment (Bai 2024).

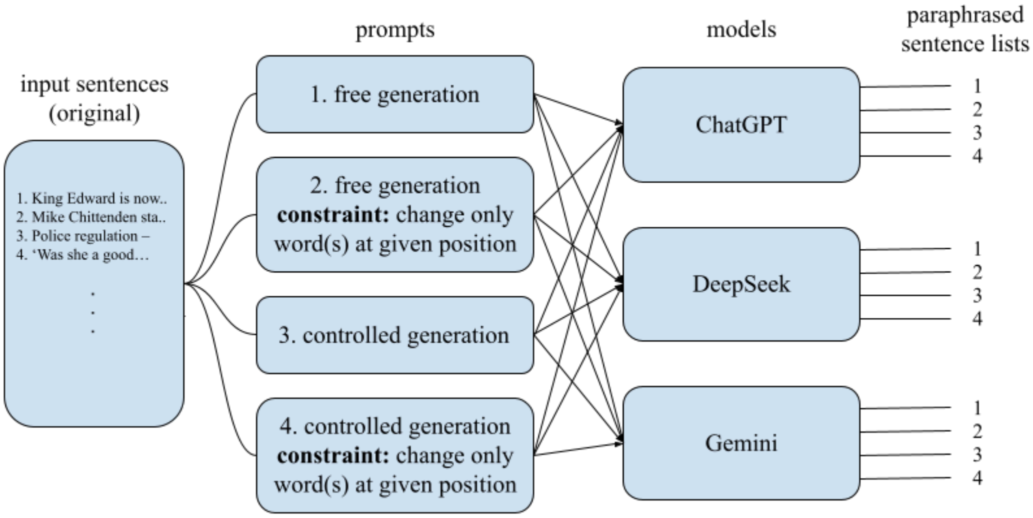
4.2 Experimental Design

Based on the two-method metaphor generation paradigm using CMT used by Stowe and others, we test model performance in **free** and **controlled** metaphor generation, and further subdivide each experiment into two sub-experiments, first giving the model flexibility to paraphrase the sentence with the given information, without constraining it to only paraphrase within indexical positions of the literal words, and again with these index constraints.

Free generation We give the model the freedom to map the given target to any source domain for the paraphrase. In other words, each model is provided only the input and target domain, allowing the model to determine a source domain to paraphrase the model with. For example, with the input sentence "They were married yesterday" we provide the target domain "Action" as an additional input. Each model is prompted to repeat this twice for all 100 input sentences, first without positional constraints, and again with the paraphrase constrained between two indexes.

Controlled generation A source domain is explicitly supplied along with the target in the input sentence. Considering the same example from above, "They were married yesterday", both the target domain "Action" and source domain "Motion along a path" are specified to constrain paraphrased metaphor outputs to metaphors that map the given source to the target. Each model was also asked to metaphorically paraphrase the original list of sentences with no positional restriction, and again with the changes it was allowed to make restricted to a given indexical range. *Figure 1* shows the full experimental paradigm. See the Appendix for the prompts utilized to direct each experiment.

FIGURE 1: Prompt-Driven Metaphor Generation Experimental Paradigm


input sentences (original)
1. King Edward is now..
2. Mike Chittenden sta..
3. Police regulation –
4. 'Was she a good…
prompts
1. free generation
2. free generation **constraint:** change only word(s) at given position
3. controlled generation
4. controlled generation **constraint:** change only word(s) at given position
models
ChatGPT
DeepSeek
Gemini
paraphrased sentence lists
1
2
3
4

The goal of this design is to evaluate the qualities of generated CMT metaphors with varying levels of detail/ provided structure in the prompts. For instance, less constraints, such as in the cases of **free** generation and no positional information, may lead to more metaphoric (abstract and novel) responses but may be more subject to nonsensical AI hallucinations while more constraints, such as in the cases of **controlled** generation and including positional information, may produce more fluent responses at the expense of metaphor novelty.

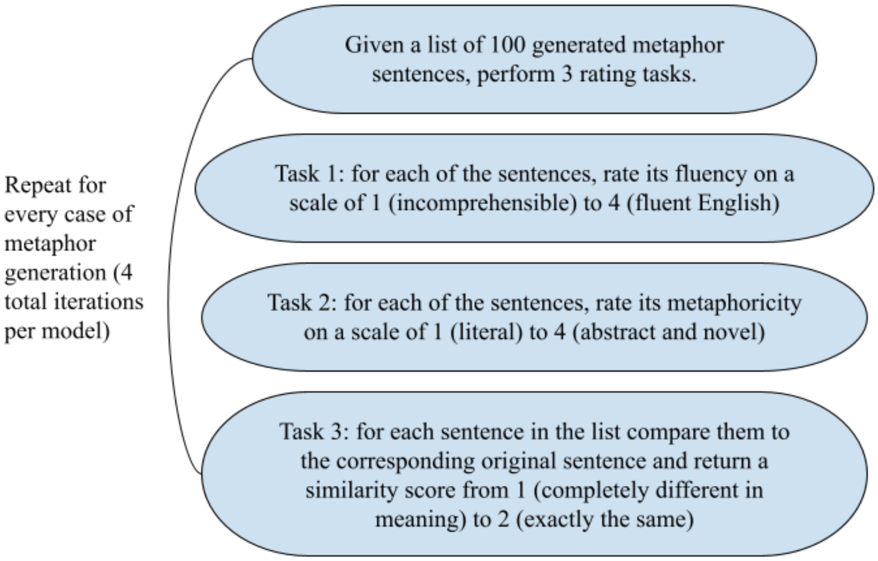
5 EVALUATION

We define a "good" metaphor to be one that is grammatically correct, while balancing understandability with creative use of language. Since creative paraphrasing is a very subjective task and so are these evaluation goals, methodology to score each output in a more objective and consistent manner is called for. Previous work on metaphor evaluation techniques establishes three benchmarks for evaluating a models' creative mastery of language (Stowe et al. 2021):

1. Fluency: how grammatically correct and sensical the paraphrase is.
2. Semantic Similarity: how close the meaning of the paraphrased sentence is to the original.
3. Metaphoricity: how abstract and novel the paraphrased metaphor is.

The quality of a generated metaphor paraphrase can be quantified through assigning it a score in each of these three categories. In practice, researchers did this with large-scale crowdsourcing evaluation methods, hoping that by averaging across many participants' ratings, noise from subjective preference could be reduced and some universal trends would emerge.

FIGURE 2: Prompt-Driven Procedure for Obtaining Evaluation Metrics for Paraphrases


Given a list of 100 generated metaphor sentences, perform 3 rating tasks.
Repeat for every case of metaphor generation (4 total iterations per model)
Task 1: for each of the sentences, rate its fluency on a scale of 1 (incomprehensible) to 4 (fluent English)
Task 2: for each of the sentences, rate its metaphoricity on a scale of 1 (literal) to 4 (abstract and novel)
Task 3: for each sentence in the list compare them to the corresponding original sentence and return a similarity score from 1 (completely different in meaning) to 2 (exactly the same)

It bears note that high scores in each benchmark category are not meant to define "effective" metaphor generation. Rather, the benchmarks obtained are meant to allow users to evaluate which model(s) would be most effective for their specific use cases. For instance, if a user wants to generate creative text from seed text that minimally changes the seed, they may prioritize a model showing higher Semantic Similarity scores over high scores in the other two categories.

5.1 Model "Crowdsourcing" Scores

Lacking the time frame and resources required for typical approaches of using either crowdsourcing to gather subjective ratings or relying on automatic evaluation metrics (such as perplexity and cosine similarity measurements of semantic embeddings) from specially trained deep learning networks like SentBERT (Stowe et al, 2021; DiStefano et al. 2024), we opted for a novel rating procedure inspired by Worsham et. al's use of a tested AI assistant to rank its own results. Stimulating the traditional human crowdsourcing method, we presented each of the three LLMs tested with every list of metaphor paraphrases generated by each model, including itself. The models were given the 12 total lists of metaphoric paraphrases (4 produced by each model) and prompted (see the Appendix for details about the rating scale and exact prompts engineered for this task) to rate each metaphor from a scale of 1 through 4 in each evaluation category. *Figure 2* shows the conversational structure of this evaluation procedure. Fluency, Similarity, and Metaphoricity scores for each set of metaphor paraphrases (free, free with position specified, controlled, controlled with position specified) yielded by each model were then averaged for a general picture of performance.

TABLE 1: Model Performances on Free and Controlled Metaphor Generation Tasks

		Original	Free generation		Controlled generation		Metric Avg.
			w/o position	w/ position	w/o position	w/ position	
ChatGPT	fluency	4	3.9767	3.97	3.9567	3.9667	**3.9675**
	similarity	4	3.5667	3.45	3.4833	3.5333	**3.5083**
	metaphoricity	1.01	3.3833	3.3733	3.3767	3.3767	3.3775
DeepSeek	fluency	3.99	3.8767	3.876	3.8733	3.8733	3.8748
	similarity	4	3.46	3.4667	3.4667	3.35	3.4358
	metaphoricity	1.08	3.75	3.7733	3.7667	3.7667	**3.7642**
Gemini	fluency	3.99	3.9133	3.4567	3.76	3.59	3.68
	similarity	4	3.51	3.1167	3.4533	3.5033	3.3958
	metaphoricity	1.22	2.1567	2.23	1.7433	1.6233	1.938

The “Original” column describes the rating assigned by the model to the input sentences and averaged across all 100 of them, unchanged, as a baseline for comparison. It is not considered in the calculation of metric averages. As expected, all models assigned the input list a average semantic similarity score of 4 (since it was being compared with itself), a fluency score at or very close to 4, and a metaphoricity close to 1. Each value in the table is the average taken across the mean ratings assigned by each of the models, for the specific paraphrase task result (see leftmost column to recognize which model the result being scored is generated by, and the header row to see which task).

6 RESULTS

Overall, all models scored high (close to 4) in all evaluation categories, with the exception of Gemini, which received noticeably lower metaphoricity scores than ChatGPT and DeepSeek. ChatGPT received the highest average fluency and similarity scores, while DeepSeek received the highest average metaphoricity score. Meanwhile, Gemini scored lowest on average in all three categories.

Contrary to expectations, there is also no apparent tradeoff between the similarity and metaphoricity of trials with positional and without positional information. Additionally, only Gemini received lower metaphoricity scores in the controlled generation task than the free generation tasks.

Model Rating Trends and Bias Evaluation
To understand the robustness of using the LLM-based crowdsourcing as an evaluation technique, for each model, we averaged the scores given by each rater in each of the 3 metric categories across the 4 paraphrasing tasks. Tables *2*, *3*, and *4* show the averaged ratings yielded by each LLM for each LLM’s generated paraphrases. Deepseek showed a tendency to assign high rating averages, giving generated paraphrases from every model higher scores on average in all three metric categories than the other 2

models did. The one exception to this trend is its assessment of the metaphoricity of Gemini's paraphrases, for which DeepSeek yielded the lowest across-task rating average out of all the raters.

Gemini shows a tendency to assign lower similarity scores, not assigning similarity scores of over 3.3 across individual tasks in collected data, and showing a maximum average similarity of 3.1675 (taken across all tasks). This different skew of ratings is likely due to differences in reasoning processes across models, as all of them were presented with the exact same rating scales and instructions to follow.

Notably, each model did not necessarily assign its own paraphrases higher scores on average. Deepseek gave ChatGPT very similar scores for fluency and similarity; ChatGPT gives itself a higher fluency score (3.94), whereas its mean fluency rating is about 3.64 for DeepSeek and 3.34 for Gemini, but it gives DeepSeek a higher metaphoricity rating average. For Gemini, though it rated the semantic similarity of its paraphrases higher on average (compared to the averages of its ratings for the other 2 models), it also self-assigned its own sets of paraphrases the lowest metaphoricity scores on average. Thus each LLM does not seem to favor higher scores for its own generated results.

TABLE 2: Ratings for ChatGPT-Generated Paraphrases, Averaged Across Tasks

	ChatGPT	DeepSeek	Gemini
fluency	**3.94**	3.99	3.9725
similarity	**3.405**	3.99	3.13
metaphoricity	**3.335**	3.8075	2.99

TABLE 3: Ratings for DeepSeek-Generated Paraphrases, Averaged Across Tasks

	ChatGPT	**DeepSeek**	Gemini
fluency	3.6425	**3.9925**	3.99
similarity	3.4725	**3.98**	2.855
metaphoricity	3.52	**3.9675**	3.805

TABLE 4: Ratings for Gemini-Generated Paraphrases, Averaged Across Tasks

	ChatGPT	DeepSeek	**Gemini**
fluency	3.3425	3.955	**3.7425**
similarity	3.3225	3.6975	**3.1675**
metaphoricity	2.32	1.54	**1.955**

Conclusions

Our results provide insight for intentional use cases for the assessed models. ChatGPT overall performs strongly in all categories, but only surpasses DeepSeek by a narrow margin in fluency and similarity. DeepSeek also has high fluency and similarity, and seems to produce more abstract metaphors than both GPT and Gemini. In general, the performance of ChatGPT and DeepSeek is close enough that either is a reliable pick for accurate and closely related metaphor paraphrases, though the data suggests that DeepSeek may provide more "creative" metaphors.

7 FUTURE DIRECTIONS

Our investigation develops a basic experimental framework that we hope will inspire future studies on the creative writing and language command capabilities of LLMs. The design we outline in this paper could be improved by running many more trials and averaging across them. This would strengthen the generalizability of our results by accounting for the intrinsic "randomness" built into each LLM that allows it to vary its responses when asked the same question (Guinness, 2024), ensuring that our results hold up beyond an "unusually good" or "unusually bad" output list.

Moreover, the underlying process behind the fluency, semantic similarity, and metaphoricity scores each LLM assigns to each paraphrase during our evaluation step remains unclear because of the complexity and scale of these models. Further exploration that compares these scores with metrics obtained from automatic metrics produced deep learning language models[3] (Stowe et al. 2021, DiStefano et al. 2024) may hint at the underlying approach to NLP scoring tasks that popular LLMs default on. Specifically, we were interested in using SentBERT to calculate the similarity between generated metaphoric paraphrases and the original sentences for a more mathematically based benchmark of the extent to which each LLM changed the original sentences. Furthermore, a social study extension of this project that allows real people to score metaphors generated by LLMs may provide more insight on how "creative" modern AI is to people now, which is relevant to individuals in creative and academic fields.

8 BONUS POINTS

Our project goes above and beyond the assignment requirements with its novel research topic and its detailed experimental design and evaluation.

To our awareness, there has not been a comprehensive study of the creative metaphor "writing" abilities of popularly used LLMs, whose growing use in recent years has been a source of concern for individuals in creative and academic fields. Our work is scientifically relevant for the fields of NLP, Prompt Engineering, Creative Writing, and Creativity Social Research, and provides multiple directions for future studies to build on.

Our experimental process was detailed and labor-intensive: by testing 3 models with 4 experimental prompts and having each model score every single output paraphrase from each model in 3 evaluation categories, we directed a total of 108 conversations with AI to get our results.

[3] Such as *perplexity*, a common method used to evaluate language models' ability to predict the words in a sequence and has been linked to fluency; *SentBERT*, which calculates the cosine similarity across two sentences; and other NLP models specifically trained on human-assigned metaphor novelty scores to rate metaphoricity.

ACKNOWLEDGEMENTS

We would like to thank Professor Zhuowen Tu for his feedback and encouragement during this project.

SOURCES

Amol Wagh. “Architecture of OpenAI ChatGPT & Tips,” In *Medium*, Jul. 02, 2023, Online. Available: https://medium.com/@amol-wagh/open-ai-understand-foundational-concepts-of-chatgpt-and-cool-stuff-you-can-explore-a7a77baf0ee3

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

Doug Worsham. 2025. Comparative Case Study of AI Research Assistants: Primo, Perplexity,Elicit, and Scite. Unpublished, Online. University of California San Diego Library.

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

Kevin Stowe, Nils Beck, and Iryna Gurevych. 2021. Exploring Metaphoric Paraphrase Generation. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 323–336, Online. Association for Computational Linguistics.

Paul DiStefano, John Patterson, and Roger Beaty. 2024. Automatic Scoring of Metaphor Creativity with Large Language Models. In *Creativity Research Journal*, pages 1–15, Online. https://doi.org/10.1080/10400419.2024.2326343

Shirley Li. “DeepSeek-V3 Explained 1: Multi-head latent attention,” In *Medium*, Mar. 19, 2025, Online. Available: https://medium.com/data-science/deepseek-v3-explained-1-multi-head-latent-attention-ed6bee2a67c4

Tanya Goyal and Greg Durrett. 2020. Neural Syntactic Preordering for Controlled Paraphrase Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 238–252, Online. Association for Computational Linguistics.

Yumo Bai. “Why Are Most LLMs Decoder-only?” In *Medium*, Jun. 18, 2024, Online. Available: https://medium.com/@yumo-bai/why-are-most-llms-decoder-only-590c903e4789

Zoltán Kövecses. 2017. Conceptual metaphor theory.

APPENDIX

Experiment Prompts

Free Generation

Without position: “Following the definition of metaphors provided by conceptual metaphor theory, for each provided literal sentence string with a given target domain in the following list (where each line is of the form target domain \t literal sentence string), please paraphrase the sentence to make a metaphor conceptualizing the given target domain while retaining its original grammatical structure. Return the list of paraphrased sentences.”

With position: “Following the definition of metaphors provided by conceptual metaphor theory, for each provided literal sentence string with a given target domain and index start and end positions (using zero based indexing) in the following list (where each line is of the form target domain \t literal sentence string \t [(start index inclusive, end index inclusive)]), please paraphrase only the word(s) between the given index positions inclusive to make a metaphor conceptualizing the given target domain. Return the list of paraphrased sentences.”

Controlled Generation

Without position: “Following the definition of metaphors provided by conceptual metaphor theory, for each provided literal sentence string with a given target domain and source domain in the following list (where each line is of the form target domain \t source domain \t literal sentence string), please paraphrase the sentence to make a metaphor conceptualizing the given target domain using the given source domain while retaining its original grammatical structure. Return the list of paraphrased sentences.”

With position: “Following the definition of metaphors provided by conceptual metaphor theory, for each provided literal sentence string with a given target domain, source domain, and index start and end positions (using zero based indexing) in the following list (where each line is of the form target domain \t source domain \t literal sentence string \t [(start index inclusive, end index inclusive)]), please paraphrase only the word(s) between the given index positions inclusive to make a metaphor conceptualizing the given target domain using the given source domain. Return the list of paraphrased sentences.”

Evaluation Prompts

Fluency

“For each of the given 100 sentences, select how grammatically correct/ fluent the sentence is from a scale of 1 (completely incomprehensible) to 4 (fluent English), ignoring capitalization and punctuation completely. If the sentence is fully grammatically valid, rate it a 4. If it contains some minor spelling or grammatical errors, rate it a 3. If there are multiple errors or it is difficult to understand, rate it a 2. If it is unintelligible or incomprehensible, rate it a 1. Return only the numbered list of ratings, ordered to correspond to the according sentence number you rate.”

Sentence Similarity

“For each of the 100 sentences, select how similar the meaning of the sentence is to a sentence from a second list (which I will provide you) at the same location (i.e. the first sentence from the first list should be compared to the first sentence from the second list) are from a scale of 1 (completely unrelated) to 4 (exactly the same), ignoring capitalization and punctuation entirely. If the sentences have the same meaning, select 4. If the sentences mean the same thing but there is a subtle shift in tone choose 3. If the sentences are not very similar but share some ideas, or if the tone shifts very obviously, choose 2. If the sentences are entirely unrelated, choose 1. Wait for me to provide the list of sentences to compare with the current one, and then only return your rating results as a list from 1-100 without sharing justifications."

Metaphoricity

"For each of the given 100 sentences, select how metaphoric the sentence is from a scale of 1 (most literal) to 4 (most abstract and novel), ignoring capitalization and punctuation entirely. Use the definition of a metaphor specified by Conceptual Metaphor Theory. Return only the numbered list of ratings, ordered to correspond to the according sentence number you rate."