How Are Metaphors Processed by Language Models? The Case of Analogies

Joanne Boisson[1], **Asahi Ushio**[2], **Hsuvas Borkakoty**[1], **Kiamehr Rezaee**[1], **Dimosthenis Antypas**[1], **Zara Siddique**[1], **Nina White**[1] and **Jose Camacho-Collados**[1]
[1]Cardiff NLP, School of Computer Science and Informatics
Cardiff University, United Kingdom
[2]Amazon, Japan
{boissonjc,borkakotyh,rezaeek,antypasd,siddiquezs2,camachocolladosj}
@cardiff.ac.uk

Abstract

The ability to compare by analogy, metaphorically or not, lies at the core of how humans understand the world and communicate. In this paper, we study the likelihood of metaphoric outputs, and the capability of a wide range of pretrained transformer-based language models to identify metaphors from other types of analogies, including anomalous ones. In particular, we are interested in discovering whether language models recognise metaphorical analogies equally well as other types of analogies, and whether the model size has an impact on this ability. The results show that there are relevant differences using perplexity as a proxy, with the larger models reducing the gap when it comes to analogical processing, and for distinguishing metaphors from incorrect analogies. This behaviour does not result in increased difficulties for larger generative models in identifying metaphors in comparison to other types of analogies from anomalous sentences in a zero-shot generation setting, when perplexity values of metaphoric and non-metaphoric analogies are similar.

1 Introduction

Analogical reasoning is critical to deep language understanding, as it is a core mechanism of human generalization and creativity (Holyoak and Thagard, 1996; Hofstadter, 2001). Analogical thinking includes figurativeness (e.g. The mind is a *sponge*.), in which humans naturally express relationships based on non-literal connections. Traditionally, metaphors have been challenging to model from a computational perspective (Veale et al., 2016) and in the context of NLP. This is due to their proteiform nature, conventional or creative, concise or structurally more complex.

Some limitations might have been lifted given the new wave of language models (LMs) that have revolutionalised the field of NLP and beyond (Chowdhery et al., 2022; Ouyang et al., 2022; Touvron et al., 2023). Indeed, recent studies on the last generation of large transformer-based LMs show enhanced abilities to perform analogical reasoning (Webb et al., 2023), suggesting that models of a larger size may gain the ability to process complex analogies.

As a conceptual innovation device, figurative analogies have also been studied in relation to the fluency, creativity and originality of students' writing (Kao, 2020). Creative writing support tools specialising in metaphor generation have been developed, such as Metaphoria (Gero and Chilton, 2019). The emergence of LLMs as writing assistants has further highlighted the importance of understanding how metaphors are processed by LMs, especially given some limitations pointed by their users related to the generation of poor metaphors and overly predictable endings, to name a few (Chakrabarty et al., 2024).

Motivated by the recent advances in language modeling and the need for understanding how LMs process metaphors, we establish the following two research questions:

Research Question 1 (RQ1). How do language models distinguish metaphors from literal and anomalous sentences? In particular, we are interested in determining if the likelihood of metaphors compared to both literal and anomalous sentences is consistent across models. For this, we are also interested in analysing the differences among model families and, particularly, sizes. This research question is addressed in Section 5.

Research Question 2 (RQ2). Assuming differences in the answer to RQ1, we aim to address the following complementary questions: how do metaphors impact the performance of language models in general analogy tests? Are language models capable of solving analogies when metaphors are involved? Our findings are presented

Proceedings of the 28th Conference on Computational Natural Language Learning, pages 365–387
November 15-16, 2024

in Section 6.

In order to answer both research questions, we evaluate a broad range of language models on their ability to distinguish anomalous, metaphoric and non-metaphoric sentences on datasets from psycholinguistics, that were, to our knowledge, previously unused in NLP studies. The results clearly show the marked differences in terms of perplexity between attributive metaphors and other literal attributive structures, where, in some cases metaphors are processed more similarly to anomalies, whereas in other cases, they are processed more similarly to literal examples. A last experiment on the SAT analogy test dataset allows a comparison of the models in open-generation tasks for challenging metaphors and analogies. We observed differences between perplexity and generation-based approaches, with an enhanced ability of the models to deal with metaphors in the generation setting[1].

2 Background

In this section, we provide more details on the relation between analogies and metaphors, and discuss other terminology used across the paper.

Analogies. Analogy is a type of similarity in which the same system of relations holds across different sets of elements (Gentner and Smith, 2012). The analogies that we consider express parallels across pairs of concepts captured minimally through attributive structures *A is-a B* or more explicitly with comparisons of the form *A is to B what C is to D*. Mapping conceptual structures to understand or create analogies comes naturally to humans, but it is generally challenging for computational models because it conveys implicit semantic attributes and relations. For example, understanding the statement *ketchup is to tomato what guacamole is to avocado* involves an internal representation of the relation *x is made of mashed y*.

In two-word analogies, the relation of interest is implicit. For example, from the sentence *His editing style was a chainsaw*, one can reconstruct an implicit 4-term analogy: *His editing style was to the text what a chainsaw is to a forest*.[2]

[1]The code and datasets used in our experiments can be found at `https://github.com/Mionies/Metaphors_and_Analogies`.

[2]Such reconstructions may leave room for interpretation as they are generally underdefined. For instance, *forest* may not be the only choice in the example.

Metaphors. Within Conceptual Metaphor Theory (CMT), a metaphor is defined as a mapping process between broad conceptual domains (Lakoff and Johnson, 1980), which occur at the level of thought and manifests through language. In order to study the ability of models to identify metaphoric mappings, we experiment on linguistic expressions constrained in form. In this paper, a metaphor is defined as a word (or a set of related words), that can be understood through the prism of another distant word (or another paired set of related words), without relying on additional explicit context. We feed minimal metaphoric sentences that almost only contain the words forming mappings into the models, to gain a better understanding of how they are represented by the LMs.

According to Black (1977), all metaphors mediate an analogy , but not all analogies are metaphors. The relation between metaphors and analogies has been much debated. Researchers who refer to shared features and structural analogies as the basis of metaphors disagreed with some conceptual mapping theorists who have argued that similarity is not the basis for metaphors (Grady, 1999). Gentner et al. (2001) and Bowdle and Gentner (2005) introduce a framework that intends to unify both views. The present study adopts this theoretical framework. Metaphors are treated here as a species of analogies. More recently, Wijesiriwardene et al. (2023a) proposed a taxonomy of analogies where the metaphors included in our dataset would be classified as *semantic and pragmatic analogies*, i.e. the two most complex types of analogies, which require good semantic representations, and sometimes pragmatic knowledge, to be processed accurately.

Among all analogies, we hypothesise that metaphors might be even harder to process, because they are more structurally variable than other types of analogy. The attribute and relation conveyed are partial matches. They can even violate structural consistency (Gentner et al., 1988). According to Tourangeau and Sternberg (1982), a good metaphor is one that involves two very different domains. It is not an absolute criterion, but good metaphors are often cross-domain (far) analogies, which adds to the complexity. Another specificity of metaphors is that the mapping is not reversible (Ortony, 1993), i.e., metaphors have directionality. For example *The acrobat is a hippopotamus* suggests a clumsy acrobat and *The hippopotamus is an acrobat* suggests a graceful hippopotamus.

For these two reasons, LMs may struggle to catch capture the parallelism between the concepts involved in a metaphor in comparison to other types of analogies.

Anomalies. Semantic anomalies can resemble metaphors in the sense that they may eventually bring together concepts that are distant from each other. Unlike metaphors, the two concepts do not share any obvious properties. For example, *A chair is a syllogism* can be considered to be an anomaly (Black, 1977). Fallacious analogies made of two word pairs in the *A is to B what C is to D* structures are constructed by mapping words that are not connected by the same relation. For example, having the first pair linked by a *part of* relation and the second pair by a *made of* relation.[3]

3 Related Work

Automatic metaphor processing research has seen a garnered increased in recent years, partially due to the encouraging performance of language models on existing benchmarks (Leong et al., 2020). However, there have been almost no studies on metaphors in the context of analogies.

3.1 Analogies

Czinczoll et al. (2022) compared the performance of transformer-based language models on near analogies and more creative ones. They reported a large gap in the performance of the LMs between the two categories and released the SCAN dataset of creative analogies. In the context of the recent multiplication of larger language models, we can now say that their study is limited to relatively small models, BERT and GPT2, and in the framework of fine-tuning experiments. In contrast, we study the zero-shot abilities of the model, which allows us to conveniently scale up the experiments with limited computing power. The SCAN dataset does not contain anomalies or distinguish between metaphoric and non-metaphoric analogies. Therefore, integrating it into our our experimental setting would require additional annotations.

Webb et al. (2023) studied the performance of the GPT3-davinci models on a large range of different analogies, from geometric patterns to short pieces of text. All the experiments are compared with the performance of humans on the same task. The authors observed a sudden improvement with the davinci-003 model, which corresponds to the beginning of the release of instruction-tuned models by OpenAI (Ouyang et al., 2022). These results also suggest that abstract analogical reasoning may be an emergent ability of the larger models. This was also demonstrated by Wei et al. (2022), who observed a sudden improvement in the classification of fine-grained figurative language when the models are scaled up. These works were a motivation for the present study in the context of metaphorical analogies. We tested a large number of models of different sizes, including open-source ones, to better understand how the sizes and model types impact their ability to recognise complex analogies.

Wijesiriwardene et al. (2023b) and Sultan and Shahaf (2023) recently released resources for the identification of analogical pairs of short texts. While Sultan and Shahaf (2023) do not distinguish metaphors from other analogies, Wijesiriwardene et al. (2023b) proposed a scale of complexity for analogical relations, with metaphors occupying the highest level. The open research topic of analogical reasoning between documents explored in this previous study beyond the scope of our study. Instead, we frame our experiments to explore the behavior of the models when they are provided with the minimal linguistic information necessary to create an analogy and a metaphor, in zero-shot settings.

While good performance can be achieved when the models are fine-tuned on analogy datasets, (Griciūtė et al., 2022; Yuan et al., 2023), we are interested in understanding how LMs represent metaphors without explicit fine-tuning. In this respect, the present work is more in line of perplexity-based experiments of Ushio et al. (2021b). In contrast, we do not focus on improving the perplexity metrics but on the comparison between vanilla perplexity scores across models.

3.2 Metaphors

Metaphor processing in NLP comprises many methods developed for metaphor identification (Turney et al., 2011; Tsvetkov et al., 2014; Mao et al., 2019; Wachowiak and Gromann, 2023), but also generation (Veale, 2016; Stowe et al., 2021; Chakrabarty et al., 2021b), textual (Mao et al., 2018) and multimodal (Kulkarni et al., 2024) interpretation, metaphor understanding through entailment (Agerri et al., 2008; Chakrabarty et al., 2021a; Stowe et al., 2022), among other tasks. Ge et al.

[3] In the rest of this paper, we refer to the sentences that are not figurative, and not semantically anomalous as literal. Table 1 shows examples of 2-terms literal sentences, that are not analogies, and 4-terms sentences that are analogies.

Dataset	Format	n_sent	n_set	n_ins	Labels	Example
Cardillo	2-term	520	2	260	Literal	The murder weapon was a chainsaw.
					Metaphor	His editing style was a chainsaw.
Jankowiak	2-term	360	3	120	Literal	These marks are bruises.
					Metaphor	Failures are bruises.
					Anomaly	Bottles are bruises.
Green	4-term	120	3	40	Literal	Answer is to riddle what solution is to problem
					Metaphor	Answer is to riddle what key is to lock
					Anomaly	Answer is to riddle what jersey is to number

Table 1: Analogy datasets included in the experiments: n_sent indicate the number of sentences; n_set, the number of sentences per instance; and n_ins, and the number of instances. All datasets are balanced in terms of labels.

(2023) provide a comprehensive recent survey on the topic.

An early approach to metaphoric mapping detection that resonates with our perplexity-based study is the measurement of the preference of predicates for semantic classes of arguments (Fass and Wilks, 1983), formalized by Resnik (1997) as a WordNet based selectional preference (SP) and SP strength measure. Mason (2004); Shutova et al. (2010); Li et al. (2013) rely on the assumption that metaphoric verb-object pairs will tend to appear with lower association strength than literal compositions. More recently, Zhang and Liu (2022) models SP violations as incongruity between target words and their contexts.

In a similar work to ours, Pedinotti et al. (2021) investigated the plausibility of metaphoric associations for LMs. BERT's ability to identify the boundaries of metaphoric creativity is studied with literal sentences, conventional metaphors, creative metaphors and nonsensical sentences, and observed that the average pseudo-likelihood scores decreases in this order for the four considered categories, in accordance with human ratings of semantic plausibility. We expand the analysis to additional models and datasets, including 4-term analogies, and compare perplexity-based results to generation-based results for instructed models.

4 Experimental Details: Model Selection and Perplexity Computation

Our aim in this paper is to evaluate a wide range of diverse LMs in terms of architecture and size, which are presented below.

Models. In our experiments, we consider the masked language models BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), decoder-only LM GPT-2 (Radford et al., 2019), GPT-J (Wang and Komatsuzaki, 2021), OPT (Zhang et al., 2022), OPT-IML (Iyer et al., 2022), Galactica (Taylor et al., 2022), Bloom (Hasanain and Elsayed, 2022) and Bloomz (Muennighoff et al., 2023) , Llama-2 and Llama-3 (Touvron et al., 2023) , and the encoder-decoder LM T5 (Raffel et al., 2020), Flan-T5 (Chung et al., 2022), Flan-UL2 (Tay et al., 2023). Finally, we consider the recent Mistral (Jiang et al., 2023) and Sparse Mixture of Experts Mixtral models (Jiang et al., 2024). All the model weights are taken from HuggingFace, where the complete list of the models we used can be found in Appendix 6. In addition to those open-source LMs, we consider the OpenAI commercial API models. We use GPT-3 (Brown et al., 2020a), GPT-3.5 Instruct (Ouyang et al., 2022), GPT-3.5 and GPT-4 (Bubeck et al., 2023).[4]

Perplexity. Perplexity measures how well a LM predicts a given sentence. In that respect, this measure can provide a good proxy to compare how natural or likely different types of sentences are. Following previous work (Brown et al., 2020a; Ushio et al., 2021b), for comparing the sentence likelihood we compute perplexity on each candidate sentence and choose the one with the lowest perplexity[5]. For decoder-only LMs such as GPT (Radford et al.), we compute the perplexity of a tokenized sentence $\boldsymbol{x} = [x_1 ... x_m]$ as

$$f(\boldsymbol{x}) = \exp\left(-\frac{1}{m}\sum_{j=1}^{m}\log P_{\text{lm}}(x_j|\boldsymbol{x}_{j-1})\right) \quad (1)$$

where $P_{\text{lm}}(x|\boldsymbol{x})$ is the likelihood of the next token given the precedent tokens. For masked language models (MLM) such as BERT (Devlin et al., 2019),

[4] In the main body of the paper we provide results for the largest models, as well as representative models for all families in the size experiments, but in the appendix we include results for all models.

[5] We use `https://github.com/asahi417/lmppl` to compute perplexity.

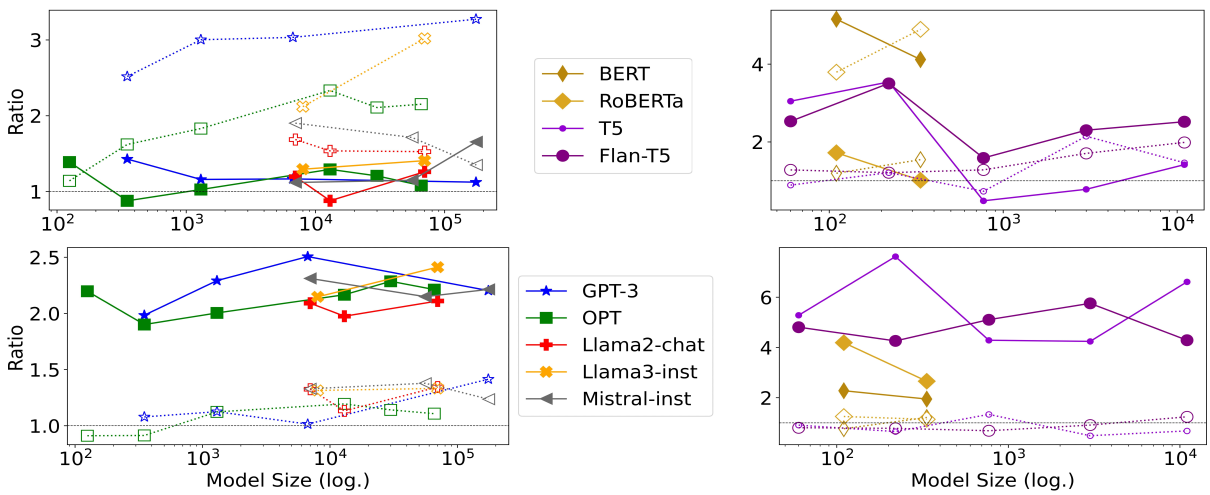

Figure 1: Medians of the ratios between the perplexities of the metaphoric and literal instances (solid lines) and between the anomalous and metaphoric instances (dashed lines) for decoder only models on the left, masked and encoder-decoder models on the right, for the Jankowiak dataset (upper plots) and Green dataset (lower plots).

pseudo-perplexity (Salazar et al., 2020) is used instead, which replaces the likelihood P in Equation 1 by $P_{\text{mask}}(x_j|\boldsymbol{x}_{\setminus j})$, the pseudo-likelihood (Wang and Cho, 2019) to predict the masked token x_j. For encoder-decoder LMs such as T5 (Raffel et al., 2020), we compute P_{lm} on the decoder, which is conditioned by the encoder. We should emphasize that perplexity values are model-dependent. Thus, in this work we have not attempted to measure perplexity values across LMs, but only for comparing sentences within the same LM. [6]

5 Language Model Representation of Metaphoric Analogies

In this section, we aim to understand how LMs identify metaphors in comparison to other types of analogies or literal statements, and how models can identify them from semantically anomalous sentences. To this end, we rely on three datasets containing sets of metaphoric and literal sentences, which are presented in Section 5.1. Following this, we rely exclusively on zero-shot experiments, first by computing perplexity scores (Section 5.2) and then by studying the abilities of the models to identify metaphors by following instructions (Section 5.3).

5.1 Metaphors and analogy datasets

In our evaluation, we focus on datasets that contain metaphors. Because of this, we exclude other well-known analogy datasets such as Google-analogies (Mikolov et al., 2013) or BATS (Gladkova et al., 2016), as they include analogies directly linked to well-defined lexical relations (e.g. capital-of). The three datasets considered in our experiments are summarized in Table 1. They are all composed of sets within which one element of the pairs remains identical and the second one varies.

Our data have two different formats. The Cardillo and Jankowiak datasets are sentences formed from two concepts based on the pattern x *is-a* y, where the problem to solve is the nature of the relation between x and y. The Green data are quadruples of the form $\{(x_i, x_j), (y_i, y_j)\}$ where the relation of interest stands between (x_i, x_j) and (y_i, y_j). Green and Jankowiak contain metaphoric, anomalous and literal sentences, while Cardillo only contains metaphoric and literal sentences.

Cardillo. This dataset (Cardillo et al., 2010, 2017) was initially created for studies within experimental psychology and contains 260 pairs of x is-a y instances. Each instance in the pair is composed of one literal and one metaphoric sentence.[7] We group the initial dataset from Cardillo et al. (2010) with the extension released in Cardillo et al. (2017). In addition to the set of instance pairs, each sentence has been annotated by a large number of participants on a scale of figurativeness that we also consider in our perplexity analysis.

[6] In the following experiments, due to computational resource limitation, we use the bitsandbytes python module to load the models larger than 13B parameters with quantization.

[7] Liu et al. (2022) created a large dataset of x is-a y metaphoric pairs but they do not contain negative examples.

Jankowiak. The Jankowiak dataset (Jankowiak, 2020) results from a similar study. In addition to literal and metaphorical sentences, it contains anomalous x is-a y sentences. It contains 120 sets of three sentences sharing the same concrete end word y, and the start words x are in the same range of frequencies.

Green. The Green dataset (Green et al., 2010) contains 120 quadruples organised in 40 sets. Each set contains one incorrect analogy (referred to as *anomaly*), one near analogy, and one far analogy (metaphor in our context). [8] For this dataset consisting of word pairs and not full sentences, we construct minimal sentences of the form A *is to* B *what* C *is to* D, where (A, B) is the first pair and (C, D) is the second pair.

5.2 Perplexity analysis

The metaphoric, anomalous and literal sentences from each dataset are fed into the model, and the perplexity is computed over each sentence, as explained in Section 4.

Results. For all datasets and for the vast majority of models, the median of the perplexities of metaphoric examples is higher than the median of literal ones, which is similar to the findings of Pedinotti et al. (2021) when analysing BERT-like models.[9] Full results and statistical significance of the difference in perplexity scores between the three classes are shown in Tables 7,8 and 9 in Appendix, Section B.2.

Figure 1 shows the variation of the perplexity ratios between metaphoric and literal examples and between anomalous and metaphoric examples, for the Jankowiak and the Green dataset. For the Green dataset, model perplexities are closer between metaphors and anomalies than between metaphors and literal instances. The ratios remain relatively stable when the size of the models increase, but we observe that the gap between metaphors and anomaly values increases for the largest decoder-only models. In contrast, in the Jankowiack dataset, metaphoric examples have closer perplexity scores to the literal ones than to the anomalous ones among most decoder-only models, and show unstable trends among the masked

[8] Kmiecik et al. (2019) released a similar corpus with 720 quadruples divided into near, far and incorrect analogies, but unlike Green, the far analogies were not all metaphors.

[9] Perplexity scores distributions for Llama3-Inst$_{70B}$ can be found in the Appendix Figure 5 as an example.

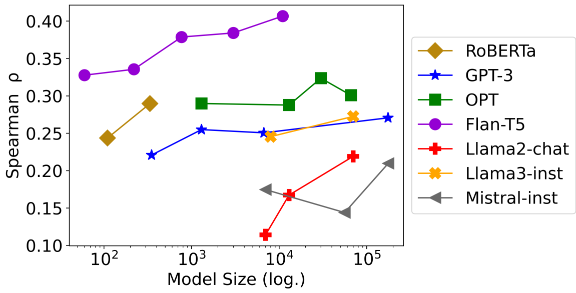

Figure 2: Correlation with human judgment for the perplexity setting on the Cardillo dataset.

and encoder-decoder models.

Finally, as an example of the impact of instruction tuning on the representation of metaphors, we see that T5 and Flan-T5 models show different score distributions, particularly in the Jankowiak dataset. More comparison between instructed and non instructed version of the models can be found in Section B.2 of the Appendix. Across all the considered datasets, Flan-T5 models score the literal examples of each set lower than the other classes in a large majority of cases. This specificity on Flan-T5 models appears in the next experiment.

Correlation between perplexity and figurativeness. Humans perceive sentences as more or less metaphoric, rather than merely as binary categories. As explained in Section 5.1, Cardillo et al. (2010, 2017) enriched their dataset with human ratings for each instance according to *figurativeness*. We study the correlation between all the previously obtained perplexities and the human judgments of figurativeness using Spearman correlation ρ. As shown in Figure 2, all models correlate positively with figurativeness. This means that sentences which are more figurative, tend to be have a lower pseudo log-likelihood according to the LMs.

FLAN-T5$_{\text{XXL}}$ obtains the highest Spearman correlation ρ of .41, and the Flan-T5 family correlation improves with the model size. BERT$_{\text{BASE}}$ and BERT$_{\text{LARGE}}$ also obtain competitive correlations, respectively .37 and .35. There is a weaker correlation for all other models including the largest ones (see the complete results in the Appendix, Table 11). The relatively low correlation between perplexity and figurativeness can be explained by the various levels of conventionality or creativity of the metaphors in the Cardillo dataset. Some frequently encountered metaphors are still perceived as very figurative. For example *The exhibition was a smash.* is both common and judged highly figurative.

5.3 Can LMs identify metaphors from literal and anomalous sentences?

In this setting, we explicitly ask the models specialised in generation to produce a response to identify literal, metaphoric and anomalous sentences of each set at once with a prompt[10], in the form of multiple-choice question tasks. This allows us to integrate OpenAI models for which perplexity values are not accessible. We process the generated answers by each model[11] and provide the overall results based on accuracy. We run the experiments with all possible permutations of the sentences within each set (shuffling the order in which literal, metaphoric and anomalous sentences are presented in the prompt) because we identified a bias toward the generation of some sequences of labels in the models.[12]

Results. Accuracy scores for the models analysed in this setting are shown in Table 2. In this setting, Flan-T5$_{\text{XXL}}$ loses its advantage over the Llama2 and Mistral models. Unlike the other models, its generated answers do not always contain distinct labels for the elements of a set, especially for the Cardillo and Jankowiak datasets that contain three sentences per set. For those two datasets, the gap in accuracy with the other models is above 16 points. All the models have difficulties processing the Green dataset, made of 4-term instances, with the exception of GPT-4 that reaches an accuracy of 78.6%.

Error analysis. An error analysis of the results on the Green and Jankowiak datasets evaluated through the generation setting is shown in Table 3. For both datasets and all models, we observe that the confusion between literal and anomalous sentences is significantly less frequent than the confusion between metaphors and anomalies. With GPT-4, the confusion between metaphors and anomalies drops significantly for both datasets on all error types.

[10]An example prompt is available in Appendix C.1.

[11]The default hyper-parameters are used for all models. The minimum or maximum output length are adjusted to ensure a complete answer. Generation answers are processed semi-automatically, verifying manually those answers that do not conform exactly with the expected output.

[12]This bias is reported in the Appendix (Tables 12 and 13).

Model	Card.	Jank.	Green
FLAN-T5$_{\text{XXL}}$	78.9	57.4	37.6
Llama2-chat$_{\text{70B}}$	85.6	73.6	56.4
Llama3-Instr$_{\text{70B}}$	88.7	89	64.3
Mixtral-Instr$_{\text{8x7B}}$	76.5	84.1	55.3
Mixtral-Instr$_{\text{8x22B}}$	82	81.9	67.1
GPT-3.5$_{\text{turbo-inst.}}$	65.9	61.5	38.8
GPT-3.5$_{\text{turbo}}$	70.5	59.8	41.2
GPT-4	**91.8**	**91.4**	**78.6**
Random	50.0	33.3	33.3

Table 2: Accuracy of the generated answers for the three datasets Cardillo, Jankowiak and Green in the instruction generation setting (gen).

Model	Jank.			Green		
	LM	MA	LA	LM	MA	LA
FLAN-T5$_{\text{XXL}}$	282	**521**	116	214	**220**	15
Llama2-chat$_{\text{70B}}$	127	**345**	99	86	**111**	92
Llama3-Instr$_{\text{70B}}$	80	**117**	41	**111**	92	54
Mixtral-Instr$_{\text{8x7B}}$	127	**141**	75	**130**	123	60
Mixtral-Instr$_{\text{8x22B}}$	90	**253**	45	37	**153**	35
GPT-3.5$_{\text{turbo-inst.}}$	260	**433**	138	140	**165**	136
GPT-3.5$_{\text{turbo}}$	179	**450**	234	137	140	**143**
GPT-4	79	**89**	18	**92**	48	14

Table 3: Error analysis for the Jankowiak and Green datasets in the generation setting (gen). The non-directional confusion between *literal* and *metaphor* (LM), *metaphor* and *anomaly* (MA) and *literal* and *anomaly* (LA) labels are shown for all the models evaluation on generation.

6 Do Metaphors Have an Impact on How LMs Solve Analogies?

In the previous section, we tested the capabilities of language models in explicitly recognising metaphors. The results show how models find them less likely than literal sentences. A natural question that may arise is whether this behavior has an impact on how LMs solve analogies more generally. In particular, our aim is to understand whether LMs are capable of solving analogies irrespective of whether they are metaphorical or not.

6.1 Data

We rely on the SAT analogy dataset (Turney, 2006) for our experiments. SAT is composed of 374 multiple-choice word analogy questions from the SAT college entrance exam in the US. This dataset has been used in the context of NLP to evaluate how models recognise analogies (Brown et al., 2020b; Ushio et al., 2021b,a; Chen et al., 2022; Kumar and Schockaert, 2023). One advantage of this dataset

Input: ***weave is to fabric what ...***	**Label: Met.**
1) illustrate is to manual	4) bake is to oven
2) hang is to picture	⇒ **5) write is to text**
3) sew is to thread	

Table 4: Example set of the SAT dataset where the correct analogy *5)* has been labeled as a metaphor.

over other benchmarks is that the dataset was not openly available on the internet, which mitigates possible concerns of data contamination in LMs. Each set in the SAT contains a stem word pair, and five other candidate pairs, forming a correct analogy and four anomalies with the stem pair. The task consists of selecting the correct analogy.

SAT annotation Each of the 374 questions of the SAT dataset contains a single correct analogy, and a subset of them are metaphoric analogies, as in the example presented Table 4. Our aim is to divide SAT correct analogies between metaphoric and non-metaphoric ones. This extended annotation enables a new experiment in which we assess the SAT performance of different types of analogies, metaphorical or not. Moreover, in the unlikely case that any of the closed language models that we analysed had been trained with the original SAT analogies, this information was not available to the model. Given the difficulty of the task, the annotation process required two rounds of annotation, detailed in the Appendix Section E.1.

A common reason for disagreement after the first round was that, sometimes, annotators could not think of a context in which two pairs of concepts could be used metaphorically. When one annotator had a clear example in mind, he or she was usually able to convince the others that an analogy was metaphoric during the discussions. For instance, the example *playwright is to actor what composer is to musician*, is easier to label after seeing the example *The playwright made him the gong in the symphony of his play*. Disagreement often occurred with the analogies when concrete domains were not very distant from each other[13]. We therefore asked all annotators to suggest and share examples prior to the second round of annotations. In total, 103 instances were labelled as metaphoric, and 239 as non-metaphoric.

[13]This difficulty is related to the practical delimitation and granularity of domains.

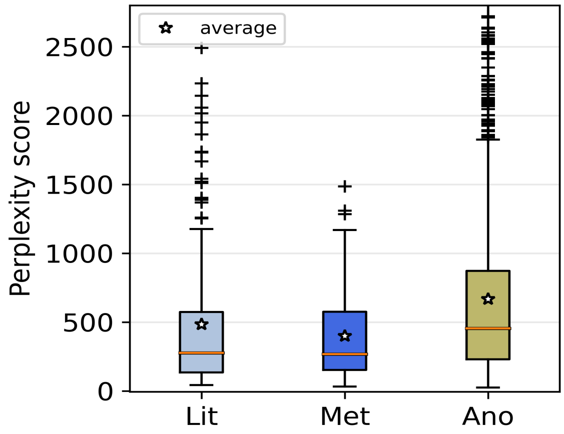

Figure 3: Boxplot showing the distribution of the perplexity scores for the three classes literal sentences (Lit), metaphor (Met) and anomalies (Ano) for the $\text{Llama3}_{\text{70B-instr}}$ model in SAT. Results for all models can be found in the Appendix, Table 10.

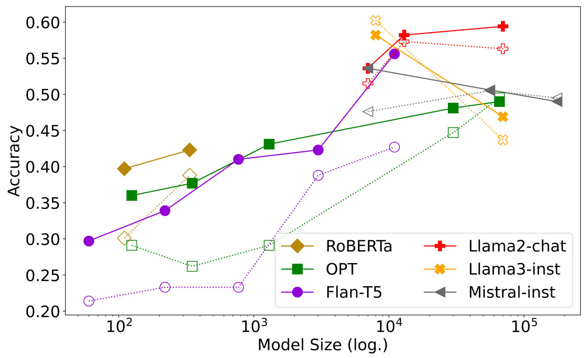

Figure 4: Accuracy results of the *perplexity* setting experiment on SAT. The results for the metaphoric class are displayed in the dashed lines, while the results for the non-metaphoric class are shown in the solid lines.

6.2 Experimental results

Experimental setting. The experimental setting is similar to the ones set out in the previous section. In particular, we test LMs using perplexity, following the same methodology outlined in Section 5.2. In this case, out of the five choices, the instance with the lowest perplexity is selected as the correct option. In addition, the large instructed LMs are tested through text generation, prompted to output the correct answer among the five choices[14]. Then, we simply report the accuracy on the metaphoric and non-metaphoric subsets of SAT.

[14]As in Section 5, experiments are run on all possible permutations of the correct answer position to neutralise the effect of sentence position bias in the prompt. The prompt used for this experiment is available in Appendix E.3.

Perplexity analysis. The SAT* perplexity scores of the metaphoric and non-metaphoric analogies are in the same range of values for most models (Figure 3 shows the results for Llama3$_{70B\text{-instr}}$). A Mann-Whitney U rank test on two independent samples for the two classes (two-sided, $p<0.05$) shows significance in the difference between the two groups for only 6 of the 51 models tested (see Table 10 in the Appendix). In fact, a majority of models have slightly larger perplexity scores on average for the non-metaphoric analogies than for the metaphoric ones. The SAT dataset is designed to be a difficult test, containing infrequent words and non-obvious analogies. This allows us to study the behavior of the models and their ability to identify correct analogies when presented with metaphoric and non-metaphoric far analogies with a similar level of perplexity.

Results. Figure 4 shows the accuracy on the metaphorical and non-metaphorical subsets of SAT in the *perplexity* setting.[15] In general, model performance improves with size. Smaller models show a gap in accuracy between questions involving metaphors and other types. This gap diminishes when the model size increases until the accuracy for the metaphor class becomes similar to that of the simple analogy class in the larger models. We observe a decrease of the performance of the largest Llama3$_{70B\text{-inst}}$ and Mixtral$_{8x22B}$ models that might eventually be caused by more constrained expectations on the input format (e.g. special input tokens for Mixtral models and system prompt for Llama3).

Table 5 shows the results of the generation experiments for the large instructed models in comparison with the perplexity setting. While models tend to perform better for non-metaphoric analogies in the perplexity setting, they obtain better results on the metaphors in the generation setting. A possible explanation for this result is that the metaphors of SAT* have in fact more chances to appear in natural sentences than the artificially constructed non-metaphoric analogies. Llama3$_{70B\text{-inst}}$ and Mixtral$_{8*22\text{-inst}}$ perform better in the generation than in the perplexity setting, reinforcing the hypothesis that perplexity may not be the best metric when using these models in applications, even for the task of detecting plausible sentences or analogies. Moreover, we can observe again that GPT-4 performed better than the other models, although the conclusions that can be drawn from this model are limited due to its closed nature.

[15] See Table 10 in the Appendix for the full results.

Model	**PPL**		**GEN**	
	Lit	**Met**	**Lit**	**Met**
FLAN-T5$_{XXL}$	*55.6	42.7	41.6	44.5
Llama2-chat$_{70B}$	**59.4**	**56.3**	41.0	*49.5
Llama3-Instr$_{70B}$	46.9	43.7	55.8	*62.5
Mixtral-Instr$_{8x7B}$	50.6	50.5	45.4	47.6
Mixtral-Instr$_{8x22B}$	49.0	49.5	50.5	*55.7
GPT-3.5$_{turbo}$			28.5	32.6
GPT-4			**72.6**	**75.0**

Table 5: Accuracy results in the perplexity (PPL) and generation settings (GEN) for the literal and metaphor classes in SAT. Bold numbers show the highest accuracy scores overall. The statistical significance of the gap between literal and metaphoric accuracy scores is calculated with a two independent samples t-test ($p<0.05$), and indicated with * on the higher score in the table.

7 Conclusion

In this paper, we have analysed the capabilities of LMs to perceive and identify metaphors. Using perplexity as a proxy to measure plausibility in LMs, we observe that, in general, LMs perceive metaphors as less likely, and are often perceived closer to anomalous sentences than literal ones. In general, LMs struggle more often to distinguish metaphors from anomalous sentences even when instructed to do so, although this gap diminishes with newer and larger models.

As a result of this finding, we also investigated whether these results would be reflected in how models can distinguish metaphors from anomalies in a wider context. The results show that, at least for the new generation of LM-based conversational agents, this does not appear to be as problematic.

Several follow-up questions remain unaddressed in spite of these findings. What is the role of metaphors in generative models? Do LMs generate (new) metaphors in the context of a conversation, or do they resort to existing expressions and literal sentences? In the context of computational linguistics and semantics, it would be interesting to better understand how metaphors are internally represented or encoded in this new generation of LMs.

Limitations

There is a body of work in the literature that has questioned analogy evaluation as a reliable way to probe NLP models, and, in particular, word embeddings (Linzen, 2016; Schluter, 2018; Nissim et al., 2020). In our paper, we are not interested in analogy as an evaluation benchmark, and rather as input data to extract insights. Nonetheless, some of the criticism of the aforementioned papers with respect to word analogies can also be applied to language models. In relation to this, we have not attempted to perform extensive prompt engineering in this work, as we were interested in knowing the trends and raw behaviour of models rather than obtaining the best results. This was also prompted due to computational constraints (see Appendix F for details on the computational resources and time). It is likely, however, that some results may differ if other prompts or evaluation protocols were considered.

In this work, we did not study the model behavior in relation to the frequency of the semantic associations in corpora. Since some metaphors are more common than other literal associations, this extended control analysis may reveal other behavior patterns not captured in our experiments. Our experiments focus solely on English corpora, therefore findings may differ for other languages, especially less-resourced and languages from other families. Finally, data contamination may have an impact on the results, which we could not analyse extensively. To mitigate this, we considered datasets that are not openly available and enriched existing data, thereby ensuring that these new annotations had not been seen by any of the models.

Ethical considerations

We have not identified any potential misuse of this research. No personal data was required in the annotation of the SAT analogy dataset and all the annotators are co-authors of this paper.

Acknowledgments

We thank the anonymous reviewers for their invaluable feedback, Thomas Green at the Advanced Research Computing at Cardiff (ARCCA) for giving us access to the computational resources, Eileen Cardillo for authorizing us to re-share the datasets attached to Cardillo et al. (2010, 2017), Daniel Bouçanova Loureiro, Nico Potyca and Yi Zhou for the helpful discussions. Jose Camacho-Collados and Dimosthenis Antypas are supported by a UKRI Future Leaders Fellowship.

References

Rodrigo Agerri, John Barnden, Mark Lee, and Alan Wallington. 2008. Textual entailment as an evaluation framework for metaphor resolution: A proposal. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 357–363. College Publications.

Max Black. 1977. More about metaphor. *Dialectica*, 31(3/4):431–457.

Brian Bowdle and Dedre Gentner. 2005. The career of metaphor. *Psychological review*, 112:193–216.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Eileen R. Cardillo, Christine Watson, and Anjan Chatterjee. 2017. Stimulus needs are a moving target: 240 additional matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior Research Methods*, 49(2):471–483.

E.R. Cardillo, G.L. Schmidt, A. Kranjec, and A. Chatterjee. 2010. Stimulus design is an obstacle course: 560 matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behav. Res. Methods*, 42(3):651–664.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021a. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024. Creativity support in the age of large language models: An empirical study involving emerging writers.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021b. 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*, pages 4250–4261, Online. Association for Computational Linguistics.

Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022. E-KAR: A benchmark for rationalizing natural language analogical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3941–3955, Dublin, Ireland. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2022. Scientific and creative analogies in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2094–2100, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dan Fass and Yorick Wilks. 1983. Preference semantics, ill-formedness, and metaphor. *American Journal of Computational Linguistics*, 9(3-4):178–187.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, 56(02):1829–1895.

D. Gentner, B. Falkenhainer, and J. Skorstad. 1988. Viewing metaphor as analogy.

Dedre Gentner, Brian Bowdle, Phillip Wolff, and Consuelo Boronat. 2001. Metaphor is like analogy. *Metaphor Is Like Analogy*.

Dedre Gentner and L. Smith. 2012. *Analogical Reasoning*, pages 130–136.

Katy Ilonka Gero and Lydia B. Chilton. 2019. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Joseph Grady. 1999. A typology of motivation for conceptual metaphor: correlation vs. resemblance. In *Metaphor in Cognitive Linguistics*. John Benjamins.

Adam E Green, David J M Kraemer, Jonathan A Fugelsang, Jeremy R Gray, and Kevin N Dunbar. 2010. Connecting long distance: semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cereb Cortex*, 20(1):70–76.

Bernadeta Griciūtė, Marc Tanti, and Lucia Donatelli. 2022. On the cusp of comprehensibility: Can language models distinguish between metaphors and nonsense? In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 173–177, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Maram Hasanain and Tamer Elsayed. 2022. Cross-lingual transfer learning for check-worthy claim identification over twitter.

Douglas Hofstadter. 2001. Epilogue: Analogy as the core of cognition. In Dedre Gentner, Keith J. Holyoak, and Boicho N. Kokinov, editors, *The Analogical Mind: Perspectives from Cognitive Science*, pages 499–538. MIT Press.

Keith J Holyoak and Paul Thagard. 1996. *Mental leaps: Analogy in creative thought*. MIT press.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization.

Katarzyna Jankowiak. 2020. Normative data for novel nominal metaphors, novel similes, literal, and anomalous utterances in polish and english. *Journal of Psycholinguistic Research*, 49(4):541–569.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Chen-Yao Kao. 2020. How figurativity of analogy affects creativity: The application of four-term analogies to teaching for creativity. *Thinking Skills and Creativity*, 36:100653.

Matthew J. Kmiecik, Ryan J. Brisson, and Robert G. Morrison. 2019. The time course of semantic and relational processing during verbal analogical reasoning. *Brain and Cognition*, 129:25–34.

Shreyas Kulkarni, Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2024. A report on the FigLang 2024 shared task on multimodal figurative language. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 115–119, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Nitesh Kumar and Steven Schockaert. 2023. Solving hard analogy questions with relation embedding chains. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6224–6236, Singapore. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. *Metaphors we Live by*. University of Chicago Press, Chicago.

Chee Wee (Ben) 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, Online. Association for Computational Linguistics.

Hongsong Li, Kenny Q. Zhu, and Haixun Wang. 2013. Data-driven metaphor recognition and explanation. *Transactions of the Association for Computational Linguistics*, 1:379–390.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany. Association for Computational Linguistics.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia. Association for Computational Linguistics.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

Zachary J. Mason. 2004. CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System. *Computational Linguistics*, 30(1):23–44.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Ortony. 1993. *Metaphor and Thought*, 2 edition. Cambridge University Press.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021. A howling success or a working sea? testing what BERT knows about metaphors. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Tagging Text with Lexical Semantics: Why, What, and How?*

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Natalie Schluter. 2018. The word analogy testing caveat. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 242–246, New Orleans, Louisiana. Association for Computational Linguistics.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010, Beijing, China. Coling 2010 Organizing Committee.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. 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 (Volume 1: Long Papers)*, pages 6724–6736, Online. Association for Computational Linguistics.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Oren Sultan and Dafna Shahaf. 2023. Life is a circus and we are the clowns: Automatically finding analogies between situations and processes.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science.

Roger Tourangeau and Robert J. Sternberg. 1982. Understanding and appreciating metaphors. *Cognition*, 11(3):203–244.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2021a. Distilling relation embeddings from pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9044–9062, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021b. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.

Tony Veale. 2016. Round up the usual suspects: Knowledge-based metaphor generation. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 34–41, San Diego, California. Association for Computational Linguistics.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. *Metaphor: A computational perspective*. Morgan & Claypool Publishers.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Thilini Wijesiriwardene, Amit Sheth, Valerie L. Shalin, Amitava Das, and Amit Sheth. 2023a. Why do we need neurosymbolic ai to model pragmatic analogies? *IEEE Intelligent Systems*, 38(5):12–16.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal G. Gajera, Shreeyash Mukul Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023b. Analogical – a novel benchmark for long text analogy evaluation in large language models.

Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2023. Analogykb: Unlocking analogical reasoning of language models with a million-scale knowledge base.

Shenglong Zhang and Ying Liu. 2022. Metaphor detection via linguistics enhanced Siamese network. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

A Language models used in our experiments

The models evaluated in the experiments, along with their sizes and corresponding HuggingFace links, are presented in Table 6.

	Model	Size	Name on HuggingFace
Masked LM	$\text{BERT}_{\text{BASE}}$	110M	bert-base-cased
	$\text{BERT}_{\text{LARGE}}$	355M	bert-large-cased
	$\text{RoBERTa}_{\text{BASE}}$	110M	roberta-base
	$\text{RoBERTa}_{\text{LARGE}}$	355M	roberta-large
Encoder-Decoder LM	T5_{SMALL}	60M	t5-small
	T5_{BASE}	220M	t5-base
	T5_{LARGE}	770M	t5-large
	T5_{3B}	3B	t5-3b
	T5_{11B}	11B	t5-11b
	$\text{Flan-T5}_{\text{SMALL}}$	60M	google/flan-t5-small
	$\text{Flan-T5}_{\text{BASE}}$	220M	google/flan-t5-base
	$\text{Flan-T5}_{\text{LARGE}}$	770M	google/flan-t5-large
	$\text{Flan-T5}_{\text{XL}}$	3B	google/flan-t5-xl
	$\text{Flan-T5}_{\text{XXL}}$	11B	google/flan-t5-xxl
	Flan-UL2	20B	google/flan-ul2
	UL2	20B	google/ul2
Decoder-only LM	GPT-2	124M	gpt2
	$\text{GPT-2}_{\text{MEDIUM}}$	355M	gpt2-medium
	$\text{GPT-2}_{\text{LARGE}}$	774M	gpt2-large
	GPT-2_{XL}	1.5B	gpt2-xl
	$\text{GPT-J}_{\text{125M}}$	125M	EleutherAI/gpt-neo-125M
	$\text{GPT-J}_{\text{2.7B}}$	2.7B	EleutherAI/gpt-neo-2.7B
	GPT-J_{6B}	6B	EleutherAI/gpt-j-6B
	$\text{GPT-J}_{\text{20B}}$	20B	EleutherAI/gpt-neox-20b
	OPT_{125M}	125M	facebook/opt-125m
	OPT_{350M}	350M	facebook/opt-350m
	$\text{OPT}_{\text{1.3B}}$	1.3B	facebook/opt-1.3b
	OPT_{13B}	13B	facebook/opt-13b
	OPT_{30B}	30B	facebook/opt-30b
	OPT_{66B}	66B	facebook/opt-66b
	$\text{OPT-IML}_{\text{1.3B}}$	1.3B	facebook/opt-iml-1.3b
	$\text{OPT-IML}_{\text{30B}}$	30B	facebook/opt-iml-30b
	$\text{OPT-IML}_{\text{M-1.3B}}$	1.3B	facebook/opt-iml-max-1.3b
	$\text{OPT-IML}_{\text{M-30B}}$	30B	facebook/opt-iml-max-30b
	$\text{Bloom}_{\text{176B}}$	176B	bigscience/bloom
	$\text{Bloomz}_{\text{176B}}$	176B	bigscience/bloomz
	$\text{Llama2}_{\text{7B}}$	7B	meta-llama/Llama-2-7b-hf
	$\text{Llama2}_{\text{13B}}$	13B	meta-llama/Llama-2-13b-hf
	$\text{Llama2}_{\text{70B}}$	70B	meta-llama/Llama-2-70b-hf
	$\text{Llama2-chat}_{\text{7B}}$	7B	meta-llama/ Llama-2-7b-chat-hf
	$\text{Llama2-chat}_{\text{13B}}$	13B	meta-llama/ Llama-2-13b-chat-hf
	$\text{Llama2-chat}_{\text{70B}}$	70B	meta-llama/ Llama-2-70b-chat-hf
	$\text{Llama3-Inst}_{\text{8B}}$	8B	meta-llama/ Meta-Llama-3-8b-Instruct
	$\text{Llama3-Inst}_{\text{70B}}$	70B	meta-llama/ Meta-Llama-3-70b-Instruct
	$\text{Mistral}_{\text{7B}}$	7B	mistralai/Mistral-7B-v0.1
	$\text{Mistral-Inst}_{\text{7B}}$	7B	mistralai/ Mistral-7B-Instr.-v0.2
sMoE	$\text{Mixtral}_{\text{8x7B}}$	56B	mistralai/Mixtral-8x7B-v0.1
	$\text{Mixtral-Inst}_{\text{8x7B}}$	56B	mistralai/ Mixtral-8x7B-Instr.-v0.1
	$\text{Mixtral-Inst}_{\text{8x22B}}$	176B	mistralai/ Mixtral-8x22B-Instr.-v0.1

Table 6: The model checkpoints used in the LM baselines on HuggingFace model hub. All the models can be obtained at https://huggingface.co.

B Perplexity setting experiments result

B.1 Graphics of the perplexity experiment results

The boxplots of the metaphoric, literal and anomalous instances for Llama3-Inst$_{70B}$ perplexity scores for the three datasets are shown Figure 5.

B.2 Result tables for all models

Tables 7, 8 and 9 include the full experimental results of Section 5.2. They show the proportions of sets where sentences with literal, metaphoric, and anomalous content exhibit the lowest perplexity for all the datasets, and the statistical significance test results for the differences in perplexity scores obtained by the metaphoric, literal and anomalous instances.

B.3 Correlation between perplexity scores and human ratings of figurativeness

Table 11 shows the correlation with human ratings of figurativeness for the Cardillo dataset with all studied models.

C Generation experiments

In this section we provide details for the generation experiments presented in Section 5.3.

C.1 Prompt used in the generation experiments

An example prompt used for text generation in order to label all the sentences of a set at once.

> **Example : Green**
>
> I will give you three sentences and I would like you to tell me which one is "anomalous", which one is "literal", and which one is a "metaphor". There is exactly one anomalous sentence, one metaphor, and one literal sentence among the three provided sentences. Here are the three sentences:
>
> 1. flock is to goose what wolfpack is to wolf
> 2. flock is to goose what constellation is to star
> 3. flock is to goose what pond is to turtle
>
> Please provide the answer in separate lines for each sentence.
> Answer:
> Sentence 1) is

C.1.1 Specificities of the Mixtral and Llama-3 models prompts.

Mixtral models. The use of special tokens is recommended in the Mixtral models prompts to obtain the best performances [16]. We modify the prompt according to the guideline.

> **<s> [INST]** I will give you three sentences and I would like you to tell me which one is "anomalous", which one is "literal", and which one is a "metaphor". There is exactly one anomalous sentence, one metaphor, and one "literal sentence among the three provided sentences. Here are the three sentences:
>
> **{SENTENCES LIST}**
>
> Please provide the answer in separate lines for each sentence. **[/INST]** Answer:
> Sentence 1) is

Llama3 models. The output of the Llama-3 models with the original prompt did not contain the expected answer to the task. We added the following system prompt to the original prompt. The results presented for Llama3 were all generated after the integration of this system prompt.

> You always answer in three lines, with one sentence index (for example "1)","2)" or "3)") followed by the words "is metaphoric", "is literal" or "is anomalous" on each line.

C.2 Bias of the models toward label sequences

We run a first batch of generation experiments using our generation prompt, and find that all the models are biased toward some sequences of sentence-label pairs. For example, in the case of the Cardillo dataset, all the models tend to answer that the first sentence of the set is *metaphoric* and the second is *literal* much more often than the opposite. This bias of the models is presented in Appendix Tables 12 and 13. As a consequence, we ran the experiments with all possible permutations of the sentences within each set, making distribution of label sequences uniform.

D Experiments on the SAT dataset

E Annotation Guidelines for Adding Metaphorical Labels in SAT

The proportional analogies to label are made of exactly four words x_i, x_j, y_i and y_j. The relation between the four words can be paraphrased by the sentence x_i *is to* x_j *what* y_i *is to* y_j. For example, *Dancing is to walking what singing is to talking*.

[16] see https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

Model Family	Model	%Lit. is lowest	p_{value}	p<0.05	Med. M/L
BERT	BERT$_{\text{BASE}}$	73.5	.0	T	2.14
	BERT$_{\text{LARGE}}$	72.3	.0	**T**	1.7
RoBERTa	RoBERTa$_{\text{BASE}}$	64.6	.0	T	2.22
	RoBERTa$_{\text{LARGE}}$	70.0	.0	**T**	2.31
T5	T5$_{\text{SMALL}}$	76.9	.0	**T**	2.62
	T5$_{\text{BASE}}$	66.5	.0	**T**	.36
	T5$_{\text{LARGE}}$	67.7	.0	**T**	.2
	T5$_{\text{3B}}$	42.3	.9992	F	0.0
	T5$_{\text{11B}}$	50.8	.2257	F	0.0
UL2	UL2	61.5	.0002	**T**	0.17
Flan-T5	Flan-T5$_{\text{SMALL}}$	78.8	.0	**T**	2.49
	Flan-T5$_{\text{BASE}}$	77.7	.0	**T**	2.35
	Flan-T5$_{\text{LARGE}}$	80.0	.0	**T**	2.44
	Flan-T5$_{\text{XL}}$	77.3	.0	**T**	2.14
	Flan-T5$_{\text{XXL}}$	82.3	.0	**T**	2.59
Flan-UL2	Flan-UL2	80.8	.0	**T**	2.37
GPT-2	GPT-2	60.8	.0	**T**	1.5
	GPT-2$_{\text{MEDIUM}}$	62.7	.0	**T**	1.45
	GPT-2$_{\text{LARGE}}$	61.9	.0	**T**	1.43
	GPT-2$_{\text{XL}}$	63.8	.0	**T**	1.49
GPT-J	GPT-J$_{\text{125M}}$	56.5	.0039	**T**	1.39
	GPT-J$_{\text{2.7B}}$	57.3	.019	**T**	1.3
	GPT-J$_{\text{6B}}$	62.7	.0	**T**	1.6
	GPT-J$_{\text{20b}}$	61.5	.0	**T**	1.45
GPT-3	GPT-3$_{\text{ada}}$	63.1	.0	**T**	1.54
	GPT-3$_{\text{babbage}}$	67.3	.0	**T**	1.63
	GPT-3$_{\text{curie}}$	67.7	.0	**T**	1.68
	GPT-3$_{\text{davinci}}$	67.7	.0	**T**	1.75
OPT	OPT$_{\text{125M}}$	64.2	.0	**T**	1.5
	OPT$_{\text{350M}}$	63.1	.0	**T**	1.4
	OPT$_{\text{1.3B}}$	68.5	.0	**T**	1.51
	OPT$_{\text{13B}}$	68.5	.0	**T**	1.53
	OPT$_{\text{30B}}$	68.5	.0	**T**	1.59
	OPT$_{\text{66B}}$	66.9	.0	**T**	1.54
OPT-IML	OPT-IML$_{\text{1.3B}}$	67.3	.0	**T**	1.54
	OPT-IML$_{\text{30B}}$	69.6	.0	**T**	1.54
OPT-IML	OPT-IML$_{\text{M-1.3B}}$	65.8	.0	**T**	1.49
(MAX)	OPT-IML$_{\text{M-30B}}$	70.4	.0	**T**	1.59
Bloom	Bloom$_{\text{175B}}$	61.9	.0	**T**	1.36
Bloomz	Bloomz$_{\text{175B}}$	66.5	.0	**T**	1.49
Llama2	Llama2$_{\text{7B}}$	63.1	.0	**T**	1.34
	Llama2$_{\text{13B}}$	63.5	.0	**T**	1.38
	Llama2$_{\text{70B}}$	60.8	.0	**T**	1.36
Llama2-Chat	Llama2-Chat$_{\text{7B}}$	57.3	.0007	**T**	1.26
	Llama2-Chat$_{\text{13B}}$	63.1	.0	**T**	1.32
	Llama2-Chat$_{\text{70B}}$	65.0	.0	**T**	1.45
Llama3-Inst	Llama3-Inst$_{\text{8B}}$	66.5	.0	**T**	1.51
	Llama3-Inst$_{\text{70B}}$	68.8	.0	**T**	1.88
Mistral	Mistral$_{\text{7B}}$	65.0	.0	**T**	1.4
	Mixtral$_{\text{8x7B}}$	62.7	.0	**T**	1.37
Mistral-Inst	Mistral-Inst$_{\text{7B}}$	64.6	.0	**T**	1.47
	Mixtral-Inst$_{\text{8x7B}}$	61.5	.0	**T**	1.28
	Mixtral-Inst$_{\text{8x22B}}$	66.9	.0	**T**	1.36

Table 7: Ratios of instances for which the literal sentences have a lower perplexity than the metaphoric sentences in the Cardillo dataset according to model family and size (*perplexity* setting). The following two columns show the significance in the difference of perplexity scores between the set of literal sentences and metaphoric sentences. A paired samples Wilcoxon test is used (p<0.05). The last column shows the median of the ratios between the score of the metaphoric and literal sentences in each set.

Model	%L is lowest	%M is lowest	%A is lowest	% L<M<A	p_{value} L-M	$p_{L\text{-}M}$ <0.05	p_{value} M-A	$p_{M\text{-}A}$ <0.05	Med. M/L	Med. A/M
BERT$_{BASE}$	85.8	9.2	5.0	47.5	.0	**T**	.102	F	5.156	1.192
BERT$_{LARGE}$	80.0	12.5	7.5	45.8	.0	**T**	.0362	**T**	4.118	1.543
RoBERTa$_{BASE}$	53.3	34.2	12.5	32.5	.0002	**T**	.0001	**T**	1.719	3.793
RoBERTa$_{LARGE}$	43.3	43.3	13.3	30.0	.2513	F	.0	**T**	1.012	4.894
T5$_{SMALL}$	70.8	11.7	17.5	35.0	.0	**T**	.6776	F	3.047	.887
T5$_{BASE}$	79.2	11.7	9.2	44.2	.0	**T**	.3309	F	3.541	1.209
T5$_{LARGE}$	29.2	34.2	36.7	7.5	.9918	F	.9646	F	.483	.728
T5$_{3B}$	33.3	40.8	25.8	19.2	.6729	F	.1363	F	.779	2.145
T5$_{11B}$	49.2	34.2	16.7	25.8	.0602	F	.0136	**T**	1.413	1.462
UL2	65.8	22.5	11.7	33.3	.0	**T**	.0667	F	2.58	1.322
Flan-T5$_{SMALL}$	85.8	9.2	5.0	56.7	.0	**T**	.0011	T	2.529	1.278
Flan-T5$_{BASE}$	84.2	9.2	6.7	47.5	.0	**T**	.0806	F	3.505	1.207
Flan-T5$_{LARGE}$	52.5	34.2	13.3	25.0	.002	**T**	.085	F	1.585	1.279
Flan-T5$_{XL}$	81.7	15.0	3.3	56.7	.0	**T**	.0	**T**	2.3	1.704
Flan-T5$_{XXL}$	77.5	19.2	3.3	55.0	.0	**T**	.0	**T**	2.518	1.986
Flan-UL2	73.3	21.7	5.0	45.0	.0	**T**	.0002	**T**	2.37	1.636
GPT-2	58.3	25.8	15.8	42.5	.0	**T**	.0	**T**	1.592	1.945
GPT-2$_{MEDIUM}$	36.7	50.8	12.5	26.7	.9724	F	.0	**T**	0.755	2.911
GPT-2$_{LARGE}$	41.7	44.2	14.2	30.0	.1797	F	.0	**T**	.979	2.698
GPT-2$_{XL}$	35.8	52.5	11.7	25.8	.6566	F	.0	**T**	.87	3.006
GPT-J$_{125M}$	21.7	62.5	15.8	15.8	.9999	F	.0	**T**	.662	3.269
GPT-J$_{2.7B}$	31.7	55.8	12.5	22.5	.9956	F	.0	**T**	.593	4.341
GPT-J$_{6B}$	38.3	52.5	9.2	28.3	.8928	F	.0	**T**	.763	3.795
GPT-J$_{20b}$	50.0	37.5	12.5	37.5	.2047	F	.0	**T**	1.047	2.885
GPT-3$_{ada}$	54.2	35.8	10.0	40.0	.0013	**T**	.0	**T**	1.427	2.517
GPT-3$_{babbage}$	50.0	40.0	10.0	40.0	.0474	**T**	.0	**T**	1.158	3.002
GPT-3$_{curie}$	51.7	41.7	6.7	35.8	.0399	**T**	.0	**T**	1.165	3.033
GPT-3$_{davinci}$	49.2	43.3	7.5	34.2	.0806	F	.0	**T**	1.122	3.273
OPT$_{125M}$	44.2	30.0	25.8	21.7	.0001	**T**	.3836	F	1.387	1.14
OPT$_{350M}$	36.7	45.8	17.5	19.2	.585	F	.0006	**T**	.876	1.62
OPT$_{1.3B}$	40.8	44.2	15.0	25.0	.3443	F	.0	**T**	1.025	1.83
OPT$_{13B}$	52.5	36.7	10.8	36.7	.0039	**T**	.0	**T**	1.291	2.332
OPT$_{30B}$	48.3	40.8	10.8	35.8	.0227	**T**	.0	**T**	1.205	2.107
OPT$_{66B}$	43.3	43.3	13.3	27.5	.2122	F	.0	**T**	1.077	2.151
OPT-IML$_{1.3B}$	40.0	42.5	17.5	26.7	.3224	F	.0	**T**	.99	1.684
OPT-IML$_{30B}$	44.2	42.5	13.3	27.5	.0519	F	.0	**T**	1.118	1.999
OPT-IML$_{M\text{-}1.3B}$	41.7	43.3	15.0	25.8	.3501	F	.0	**T**	1.016	1.794
OPT-IML$_{M\text{-}30B}$	46.7	42.5	10.8	30.8	.0476	**T**	.0	**T**	1.11	2.059
Bloom$_{175B}$	52.5	39.2	8.3	34.2	.0079	**T**	.0	**T**	1.225	2.524
Bloomz$_{175B}$	60.8	30.0	9.2	37.5	.0	**T**	.0041	**T**	1.928	1.558
Llama2$_{7b}$	52.5	33.3	14.2	29.2	.0022	**T**	.0021	**T**	1.334	1.229
Llama2$_{13B}$	47.5	35.8	16.7	25.8	.0926	F	.0012	**T**	1.192	1.398
Llama2$_{70B}$	50.0	35.0	15.0	27.5	.0283	**T**	.001	**T**	1.259	1.35
Llama2-Chat$_{7B}$	50.0	36.7	13.3	26.7	.0143	**T**	.0004	**T**	1.195	1.685
Llama2-Chat$_{13B}$	40.8	45.0	14.2	20.8	.8471	F	.0	**T**	.877	1.535
Llama2-Chat$_{70B}$	50.8	33.3	15.8	35.0	.0094	**T**	.0001	**T**	1.259	1.525
Llama3-Inst$_{8B}$	52.5	39.2	8.3	37.5	.0114	**T**	.0	**T**	1.293	2.119
Llama3-Inst$_{70B}$	51.7	38.3	10.0	37.5	.0012	**T**	.0	**T**	1.406	3.019
Mistral$_{7B}$	45.0	37.5	17.5	26.7	.1122	F	.006	**T**	1.133	1.413
Mixtral$_{8x7B}$	48.3	38.3	13.3	27.5	.079	F	.0065	**T**	1.171	1.473
Mistral-Inst$_{7B}$	45.0	36.7	18.3	30.0	.0727	F	.0006	**T**	1.118	1.901
Mixtral-Inst$_{8x7B}$	45.8	38.3	15.8	27.5	.2222	F	.0006	**T**	1.147	1.712
Mixtral-Inst$_{8x22B}$	54.2	27.5	18.3	33.3	.0	**T**	.0115	**T**	1.653	1.349

Table 8: The first three columns show the ratios of sets for which the literal (L), metaphoric (M) and anomalous (A) sentences have the lowest perplexity in the Jankowiak dataset according to model family and size (*perplexity* setting).%L<M<A shows the ratio of sets for which perplexity scores follow this order. The following four columns show the significance in the difference of perplexity scores between the set of literal and metaphoric sentences, and then between the set of metaphoric and anomalous sentences. A paired samples Wilcoxon test is used (p<0.05).

Model	%L is lowest	%M is lowest	%A is lowest	% L<M<A	p_{value} L-M	$p_{L\text{-}M}$ <0.05	p_{value} M-A	$p_{M\text{-}A}$ <0.05	Med. M/L	Med. A/M
BERT$_{\text{BASE}}$	65.0	15.0	20.0	30.0	.0	**T**	.592	F	2.277	0.765
BERT$_{\text{LARGE}}$	70.0	12.5	17.5	47.5	.0001	**T**	.0544	F	1.946	1.213
RoBERTa$_{\text{BASE}}$	80.0	2.5	17.5	52.5	.0	**T**	.4236	F	4.189	1.251
RoBERTa$_{\text{LARGE}}$	80.0	10.0	10.0	45.0	.0	**T**	.1702	F	2.654	1.139
T5$_{\text{SMALL}}$	95.0	0.0	5.0	45.0	.0	**T**	.6721	F	5.281	.907
T5$_{\text{BASE}}$	62.5	17.5	20.0	27.5	.0	**T**	.9016	F	7.618	.651
T5$_{\text{LARGE}}$	72.5	10.0	17.5	45.0	.0	**T**	.4709	F	4.287	1.335
T5$_{\text{3B}}$	70.0	7.5	22.5	25.0	.0	**T**	.8841	F	4.242	.487
T5$_{\text{11B}}$	80.0	5.0	15.0	32.5	.0	**T**	.9231	F	6.613	.677
UL2	57.5	12.5	30.0	32.5	.0	**T**	.8298	F	4.139	.907
Flan-T5$_{\text{SMALL}}$	87.5	0.0	12.5	40.0	.0	**T**	.8587	F	4.807	.805
Flan-T5$_{\text{BASE}}$	87.5	5.0	7.5	35.0	.0	**T**	.805	F	4.261	.78
Flan-T5$_{\text{LARGE}}$	85.0	5.0	10.0	30.0	.0	**T**	.9861	F	5.106	.684
Flan-T5$_{\text{XL}}$	92.5	2.5	5.0	40.0	.0	**T**	.823	F	5.756	.91
Flan-T5$_{\text{XXL}}$	80.0	7.5	12.5	45.0	.0	**T**	.255	F	4.288	1.24
Flan-UL2	85.0	7.5	7.5	55.0	.0	**T**	.0265	**T**	4.345	1.278
GPT-2	75.0	10.0	15.0	35.0	.0	**T**	.6624	F	1.937	.913
GPT-2$_{\text{MEDIUM}}$	70.0	15.0	15.0	42.5	.0	**T**	.2996	F	2.096	1.057
GPT-2$_{\text{LARGE}}$	72.5	12.5	15.0	45.0	.0	**T**	.075	F	2.299	1.202
GPT-2$_{\text{XL}}$	85.0	5.0	10.0	45.0	.0	**T**	.2101	F	2.211	.98
GPT-J$_{\text{125M}}$	60.0	12.5	27.5	27.5	.0	**T**	.8733	F	1.98	.864
GPT-J$_{\text{2.7B}}$	82.5	2.5	15.0	47.5	.0	**T**	.408	F	1.959	0.975
GPT-J$_{\text{6B}}$	87.5	5.0	7.5	55.0	.0	**T**	.1668	F	1.891	1.202
GPT-J$_{\text{20b}}$	85.0	7.5	7.5	47.5	.0	**T**	.0916	F	1.945	1.315
GPT-3$_{\text{ada}}$	77.5	12.5	10.0	45.0	.0	**T**	.3425	F	1.984	1.078
GPT-3$_{\text{babbage}}$	77.5	10.0	12.5	45.0	.0	**T**	.3573	F	2.292	1.125
GPT-3$_{\text{curie}}$	87.5	2.5	10.0	47.5	.0	**T**	.3184	F	2.506	1.014
GPT-3$_{\text{davinci}}$	92.5	2.5	5.0	62.5	.0	**T**	.0341	**T**	2.203	1.414
OPT$_{\text{125M}}$	77.5	7.5	15.0	37.5	.0	**T**	.7741	F	2.197	.911
OPT$_{\text{350M}}$	77.5	5.0	17.5	40.0	.0	**T**	.6957	F	1.901	.913
OPT$_{\text{1.3B}}$	92.5	2.5	5.0	52.5	.0	**T**	.195	F	2.004	1.123
OPT$_{\text{13B}}$	95.0	2.5	2.5	55.0	.0	**T**	.085	F	2.166	1.194
OPT$_{\text{30B}}$	97.5	.0	2.5	60.0	.0	**T**	.0385	**T**	2.286	1.141
OPT$_{\text{66B}}$	97.5	.0	2.5	57.5	.0	**T**	.0879	F	2.212	1.107
OPT-IML$_{\text{1.3B}}$	90.0	2.5	7.5	52.5	.0	**T**	.2423	F	1.964	1.032
OPT-IML$_{\text{30B}}$	90.0	2.5	7.5	52.5	.0	**T**	.1159	F	2.246	1.121
OPT-IML$_{\text{M-1.3B}}$	85.0	2.5	12.5	45.0	.0	**T**	.3279	F	1.951	.988
OPT-IML$_{\text{M-30B}}$	97.5	0.0	2.5	57.5	.0	**T**	.0624	F	2.166	1.136
Bloom$_{\text{175B}}$	80.0	5.0	15.0	52.5	.0	**T**	.1877	F	2.084	1.167
Bloomz$_{\text{175B}}$	87.5	2.5	10.0	55.0	.0	**T**	.0446	**T**	2.161	1.185
Llama-2$_{\text{7b}}$	80.0	15.0	5.0	50.0	.0	**T**	.0341	**T**	1.747	1.245
Llama-2$_{\text{13B}}$	82.5	10.0	7.5	60.0	.0	**T**	.0184	**T**	1.713	1.202
Llama-2$_{\text{70B}}$	77.5	17.5	5.0	55.0	.0001	**T**	.0011	**T**	1.785	1.322
Llama2-Chat$_{\text{7B}}$	82.5	10.0	7.5	60.0	.0	**T**	.0018	**T**	2.091	1.325
Llama2-Chat$_{\text{13B}}$	90.0	5.0	5.0	62.5	.0	**T**	.0204	**T**	1.975	1.132
Llama2-Chat$_{\text{70B}}$	80.0	15.0	5.0	62.5	.0	**T**	.0001	**T**	2.11	1.344
Llama3-Inst$_{\text{8B}}$	95.0	2.5	2.5	65.0	.0	**T**	.0043	**T**	2.147	1.314
Llama3-Inst$_{\text{70B}}$	82.5	10.0	7.5	60.0	.0	**T**	.0139	**T**	2.412	1.332
Mistral$_{\text{7B}}$	82.5	7.5	10.0	52.5	.0	**T**	.0191	**T**	2.153	1.136
Mixtral$_{\text{8x7B}}$	82.5	10.0	7.5	60.0	.0	**T**	.0041	**T**	1.976	1.313
Mistral-Inst$_{\text{7B}}$	82.5	10.0	7.5	62.5	.0	**T**	.0003	**T**	2.311	1.329
Mixtral-Inst$_{\text{8x7B}}$	80.0	12.5	7.5	52.5	.0	**T**	.0019	**T**	2.149	1.378
Mixtral-Inst$_{\text{8x22B}}$	77.5	7.5	15.0	60.0	.0	**T**	.0024	**T**	2.214	1.236

Table 9: The first three columns show the ratios of sets for which the literal (L), metaphoric (M) and anomalous (A) sentences have the lowest perplexity in the Green dataset according to model family and size (*perplexity* setting).%L<M<A shows the ratios of sets for which perplexity scores follow this order. The following four columns show the significance in the difference of perplexity scores between the set of literal and metaphoric sentences, and then between the set of metaphoric and anomalous sentences. A paired samples Wilcoxon test is used (p<0.05).

Model	p_{value} L<A	$p_{L<A}$ <0.05	p_{value} M<A	$p_{M<A}$ <0.05	p_{value} L<M	$p_{L<M}$ <0.05	p_{value} Acc. L-M	$p_{Acc.\ L\text{-}M}$ <0.05	%Lit. is lowest	%Met. is lowest
BERT$_{BASE}$	.0	**T**	.0	**T**	.6787	F	.3057	F	34.7	29.1
BERT$_{LARGE}$	.0	**T**	.0	**T**	.1583	F	.0879	F	34.3	25.2
RoBERTa$_{BASE}$	.0	**T**	.0001	**T**	.7688	F	.083	F	39.7	30.1
RoBERTa$_{LARGE}$	.0	**T**	.025	**T**	.1813	F	.555	F	42.3	38.8
T5$_{SMALL}$	.0	**T**	.0	**T**	.4271	F	.6919	F	29.3	27.2
T5$_{BASE}$	.0	**T**	.0	**T**	.0973	F	.0742	F	29.3	20.4
T5$_{LARGE}$	.0	**T**	.0	**T**	.6088	F	.4069	F	32.6	28.2
T5$_{3B}$	.0	**T**	.0	**T**	.862	F	.2241	F	36.8	30.1
T5$_{11B}$	.0	**T**	.0	**T**	.1066	F	.0216	**T**	39.7	27.2
Flan-T5$_{SMALL}$	.0	**T**	.0	**T**	.2046	F	.0981	F	29.7	21.4
Flan-T5$_{BASE}$	.0	**T**	.0	**T**	.0135	**T**	.0425	**T**	33.9	23.3
Flan-T5$_{LARGE}$	.0	**T**	.0	**T**	.0024	**T**	.0009	**T**	41.0	23.3
Flan-T5$_{XL}$	.0001	**T**	.0016	**T**	.8248	F	.555	F	42.3	38.8
Flan-T5$_{XXL}$	.169	F	.0473	**T**	.9573	F	.0286	**T**	55.6	42.7
Flan-UL2	.1298	F	.2246	F	.0963	F	.606	F	50.6	47.6
GPT-2	.0	F	.0	**T**	.0398	**T**	.1305	F	34.3	26.2
GPT-2$_{MEDIUM}$	.0	**T**	.0	**T**	.235	F	.3371	F	36.4	31.1
GPT-2$_{LARGE}$	.0	**T**	.0	**T**	.2262	F	.1503	F	38.1	30.1
GPT-2$_{XL}$	.0	**T**	.0001	**T**	.3808	F	.6338	F	37.7	35.0
GPT-J$_{125M}$	.0	**T**	.0	**T**	.085	F	.0342	**T**	37.7	26.2
GPT-J$_{1.3B}$	.0	**T**	.0	**T**	.236	F	.3456	F	39.3	34.0
GPT-J$_{6B}$	.0127	**T**	.0168	**T**	.132	F	.0609	F	49.8	38.8
GPT-J$_{20b}$	.0077	**T**	.0153	**T**	.4348	F	.207	F	45.2	37.9
GPT-3$_{davinci}$	.0604	F	.5223	F	.7779	F	.4249	F	50.6	55.3
OPT$_{125M}$	.0	**T**	.0	**T**	.9119	F	.2114	F	36.0	29.1
OPT$_{350M}$	.0	**T**	.0	**T**	.293	F	.0342	**T**	37.7	26.2
OPT$_{1.3B}$	.0024	F	.0002	**T**	.5922	F	.0122	**T**	43.1	29.1
OPT$_{30B}$	.096	F	.147	F	.7362	F	.5581	F	48.1	44.7
OPT$_{66B}$	.1401	F	.3924	F	.9563	F	.9246	F	49.0	49.5
OPT-IML$_{1.3B}$	.0006	**T**	.0003	**T**	.7934	F	.2951	F	43.9	37.9
OPT-IML$_{30B}$	.0666	F	.0781	F	.5541	F	.3125	F	50.6	44.7
OPT-IML$_{M\text{-}1.3B}$	.0003	**T**	.0004	**T**	.7761	F	.1552	F	43.1	35.0
OPT-IML$_{M\text{-}30B}$	.1221	F	.0676	F	.4202	F	.2218	F	51.9	44.7
Llama-2$_{7b}$	.5361	F	.0685	F	.0053	**T**	.3357	F	52.3	46.6
Llama-2$_{13B}$	.8903	F	.3698	F	.0985	F	.3682	F	57.7	52.4
Llama-2$_{70B}$	.7528	F	.4882	F	.0565	F	.8109	F	54.8	53.4
Llama2-Chat$_{7B}$	.5661	F	.4373	F	.1395	F	.7228	F	53.6	51.5
Llama2-Chat$_{13B}$	.9327	F	.9185	F	.298	F	.8809	F	58.2	57.3
Llama2-Chat$_{70B}$	.946	F	.6535	F	.1786	F	.5965	F	59.4	56.3
Llama3-Inst$_{8B}$	.8849	F	.9923	F	.7068	F	.7263	F	58.2	60.2
Llama3-Inst$_{70B}$	.0089	**T**	.0454	**T**	.9355	F	.5902	F	46.9	43.7
Mistral$_{7B}$	.2952	F	.2511	F	.0561	F	.9628	F	49.8	49.5
Mixtral$_{8x7B}$	.1081	F	.1586	F	.0164	**T**	.8135	F	48.1	49.5
Mistral-Inst$_{7B}$	.3453	F	.4334	F	.0385	**T**	.3124	F	53.6	47.6
Mixtral-Inst$_{8x7B}$	.2714	F	.3441	F	.1188	F	.9809	F	50.6	50.5
Mixtral-Inst$_{8x22B}$	.2538	F	.2993	F	.7561	F	.9246	F	49.0	49.5

Table 10: The first four columns shows significance in the gap of perplexity scores between the anomalies that has the lowest perplexity of the four incorrect options in each set (A) and the literal instances (L) or the metaphoric instances (M). A paired samples Wilcoxon test is used ($p<0.05$). The next two columns show the the statistical significance between the set of perplexity values of the literal and the metaphoric instances using a Mann-Whitney U test. This test is used because metaphoric and non-metaphoric analogies are not paired in the SAT. The following two columns , p_{value} Acc. L-M show the result of two independent samples t-tests to show if the accuracy of the models for non-metaphoric examples is significantly better than its accuracy on metaphoric examples. The last two columns show the ratios of instances for which the non-metaphoric analogy on the left, and the metaphoric analogy on the right, have the lowest perplexity of their set in the SAT dataset, according to model family and size (perplexity setting).

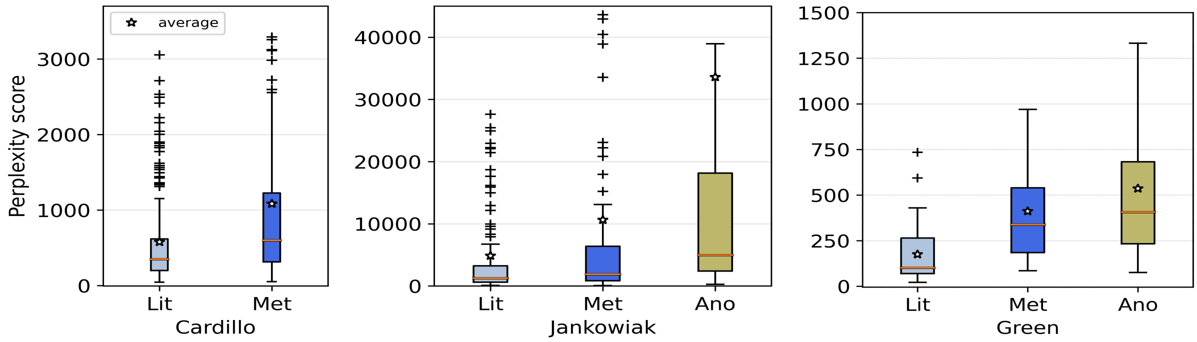

Figure 5: Boxplots of the Llama3-Inst$_{70B}$ perplexity scores for the three datasets and three classes: literal (Lit), metaphoric (Met) and anomalous (Ano). Outliers with the highest scores do not appear in the plots.

We want to decide if x_i and x_j can form a metaphoric mapping with $y_i$and y_j.

Given four words x_i, x_j, y_i and y_j :

1. Find the relation between the two elements of each pair. You can imagine relevant contexts in which they can be used. For example, *dancing* implies steps that follow a music, and *singing* often implies saying words following a music.

   If a word has multiple senses, consider its meaning in the context of the pair. For example, in the following analogy, *Abash is to embarrassment what annoy is to irritation*, the word *irritation* is polysemic. It it may take the meaning of an inflammation of the skin or be a near synonym of annoyance. Here, in the context of the word *annoy*, its emotional meaning is the only one to consider. This usage of the word may be a metaphoric sense, but it should not influence the label. We are only interested in the relation between the provided words.

   (a) Try to infer the relation between x_i and x_j

   (b) Try to infer the relation between y_i and y_j

   The relations should be similar.

2. Consider the relation between the two pairs (x_i, x_j) and (y_i, y_j).

   - Do they belong to the same domain? If x_i and y_i or x_j and y_j are either near synonyms or antonyms, then it is not a metaphor. For example, *worry is to panic what happiness is to bliss* is not a metaphor.
   - Try to recombine the pairs and form sentences using x_i and y_j or y_i and x_j. If one of the two combinations work, it may be a metaphor. For example, given *invest money* and *pour liquid*, you can construct the metaphor *pour money*.
   - Try to talk about x_i and x_j using y_i and y_j and then to talk about y_i and y_j using x_i and x_j. If you cannot think of a natural sentence, then do not label it as a metaphor.

3. Label the quadruple :

   - 0 : analogy that is not a metaphor
   - 2 : analogy that is also a metaphor
   - 1 : unsure

E.1 SAT annotations

First annotation round. Three annotators including two native speakers and two with a background in metaphor studies and linguistics labeled the 374 analogies of SAT after an initial training session and presentation of the guidelines (Appendix E). The labels were 0 for *non-metaphoric*, 1 for *unsure* and 2 for *metaphoric*. At the end of this process, in spite of the training sessions and provided guidelines, the pairwise agreement between annotators was low (Spearman $\rho = 0.17$; std$= 0.16$).

Second annotation round. In the second annotation round, we included an additional qualified native speaker and first asked all participants to place analogies in context. The source of disagreement was mainly due to the difficulty of imagining a relevant context where the 4-term analogy could be used to make a meaningful metaphor. The four participants were asked to create sentences whenever they thought that a metaphoric sentence

Model Familly	Model	Spearman ρ
BERT	$\text{BERT}_{\text{BASE}}$	.37
	$\text{BERT}_{\text{LARGE}}$	.35
RoBERTa	$\text{RoBERTa}_{\text{BASE}}$	.24
	$\text{RoBERTa}_{\text{LARGE}}$	.29
T5	T5_{SMALL}	.32
	T5_{BASE}	.11
	T5_{LARGE}	.23
	T5_{3B}	-.14
	T5_{11B}	-.05
UL2	UL2	.09
Flan-T5	$\text{Flan-T5}_{\text{SMALL}}$	.33
	$\text{Flan-T5}_{\text{BASE}}$	.34
	$\text{Flan-T5}_{\text{LARGE}}$	.38
	$\text{Flan-T5}_{\text{XL}}$	.38
	$\text{Flan-T5}_{\text{XXL}}$	**.41**
Flan-UL2	Flan-UL2	.39
GPT-2	GPT-2	.2
	$\text{GPT-2}_{\text{MEDIUM}}$	.19
	$\text{GPT-2}_{\text{LARGE}}$	.22
	GPT-2_{XL}	.21
GPT-J	$\text{GPT-J}_{\text{125M}}$	.1
	$\text{GPT-J}_{\text{2.7B}}$	.12
	GPT-J_{6B}	.21
	$\text{GPT-J}_{\text{20b}}$	.21
GPT-3	$\text{GPT-3}_{\text{ada}}$	.22
	$\text{GPT-3}_{\text{babbage}}$	.25
	$\text{GPT-3}_{\text{curie}}$	.25
	$\text{GPT-3}_{\text{davinci}}$	.27
OPT	OPT_{125M}	.29
	OPT_{13B}	.29
	OPT_{30B}	.32
	OPT_{66B}	.3
OPT-IML	$\text{OPT-IML}_{\text{1.3B}}$	.29
	$\text{OPT-IML}_{\text{30B}}$	.3
OPT-IML (MAX)	$\text{OPT-IML}_{\text{M-1.3B}}$	.28
	$\text{OPT-IML}_{\text{M-30B}}$	.31
Bloom	$\text{Bloom}_{\text{175B}}$	.19
Bloomz	$\text{Bloomz}_{\text{175B}}$	.27
Llama2	$\text{Llama-2}_{\text{7b}}$	.19
	$\text{Llama-2}_{\text{13B}}$	.19
	$\text{Llama-2}_{\text{70B}}$	.18
Llama2-Chat	$\text{Llama2-Chat}_{\text{7B}}$	.11
	$\text{Llama2-Chat}_{\text{13B}}$	.17
	$\text{Llama2-Chat}_{\text{70B}}$	.22
Llama3-Inst	$\text{Llama3-Inst}_{\text{8B}}$	.25
	$\text{Llama3-Inst}_{\text{70B}}$	.27
Mistral	$\text{Mistral}_{\text{7B}}$	.17
	$\text{Mixtral}_{\text{8x7B}}$	.18
Mistral-Inst	$\text{Mistral-Inst}_{\text{7B}}$	.17
	$\text{Mixtral-Inst}_{\text{8x7B}}$	.14
	$\text{Mixtral-Inst}_{\text{8x22B}}$	.21

Table 11: Spearman ρ correlation between human ratings of figurativeness and peplexity scores for the instances of the Cardillo dataset, according to model family and size (*perplexity* setting).

Answer	[M, L]	[L, M]	[M, M]	[L, L]
$\text{Flan-T5}_{\text{XXL}}$	**61.2**	29.4	9.4	0
$\text{Llama2-chat}_{\text{70B}}$	**57.1**	42.9	0	0
$\text{Llama3-Instr.}_{\text{70B}}$	**58.7**	38.3	3.1	0
$\text{Mixtral-Instr.}_{\text{8x7B}}$	**71.0**	24.6	3.5	0.6
$\text{Mixtral-Instr.}_{\text{8x22B}}$	**67.3**	31.3	1.3	0
$\text{GPT-3.5}_{\text{turbo-instr.}}$	**78.7**	15.8	0.2	0
$\text{GPT-3.5}_{\text{turbo}}$	**78.1**	21.7	0	0
GPT-4	**57.9**	41.5	0.6	0

Table 12: Imbalance of the models' answers on the Cardillo dataset. Experiments are run with all possible permutations of sentence within each set, with each correct sequence appearing an equal number of times in each position.

could be created. For example, given the two pairs $(sap, tree)$ and $(blood, mammal)$, one can imagine telling a kid who is damaging a tree *"Be careful, you are hurting it. Look, it is bleeding"*. The sentences were shared among all the participants and a new labelling task was completed, leading to a significant pairwise inter-annotator agreement (Spearman $\rho = 0.48$; std$= 0.17$).

The final SAT labels were obtained by averaging the scores of the four participants. We labeled as non-metaphoric all the quadruples scoring lower to 1 on average and metaphoric all those scoring above 1. 32 instances with an average score of 1 were filtered out. Table 4 contains an example of a metaphoric instance of the SAT dataset after annotation. In total, 103 instances were labelled as metaphoric, and 239 as non-metaphoric.

E.2 SAT* perplexity experiments

Table 10 shows a comparison of the models on the task of solving the analogy questions of SAT in the *perplexity* setting. The sentence in each set with the lowest perplexity is selected as the correct analogy. Accuracy is shown in two distinct columns for metaphoric and non-metaphoric analogies.

E.3 Generation experiment prompts

Prompt G2 . The correct answer of the example below is 1., it is classified as non-metaphoric in SAT. Identical modification to the prompt as the ones described in Appendix section C.1.1 are applied to Mixtral and Llama3 models.

Answer		[M, L, A]	[M, A, L]	[A, L, M]	[A, M, L]	[L, A, M]	[L, M, A]
Green	Flan-T5$_{XXL}$	0	0	0.4	7.5	0	0.4
	Llama2-chat$_{70B}$	16.2	6.2	14.6	4.2	**24.2**	17.9
	Llama3-Instr.$_{70B}$	23.8	**39.6**	14.6	18.3	0.8	0.8
	Mixtral-Instr.$_{8x7B}$	**42.5**	35.0	4.6	6.2	0.8	2.1
	Mixtral-Instr.$_{8x22B}$	**33.3**	19.6	1.7	0.4	12.9	16.2
	GPT-3.5$_{turbo-instr.}$	**75.8**	17.1	0.4	0.4	1.2	4.6
	GPT-3.5$_{turbo}$	**73.3**	3.3	7.9	5.0	0.4	8.8
	GPT-4	19.6	**28.8**	21.2	13.8	9.2	7.5
Jankowiak	Flan-T5$_{XXL}$	0.8	0.7	9.7	**34.2**	1.5	7.4
	Llama2-chat$_{70B}$	8.5	6.4	**34.0**	22.8	15.3	12.9
	Llama3-Instr.$_{70B}$	**19.2**	18.6	14.6	16.7	12.1	13.6
	Mixtral-Instr.$_{8x7B}$	20.1	18.5	18.9	16.9	8.1	13.9
	Mixtral-Instr.$_{8x22B}$	**27.9**	18.5	9.3	8.8	10.4	18.8
	GPT-3.5$_{turbo-instr.}$	**46.5**	25.6	1.7	3.1	6.7	11.5
	GPT-3.5$_{turbo}$	**53.5**	9.9	4.7	6.0	4.9	20.3
	GPT-4	22.4	21.5	13.5	13.5	13.8	14.0

Table 13: Imbalanced distribution of the sequence of labels in the models' answers on the Green and Jankowiak datasets. Experiments are run with all possible permutations of the sentences within each set, with each possible sequence of labels being the correct answer an equal number of times. Flan-T5$_{XXL}$ label distribution does not sum to 100 in the table because the model outputs a large proportion of incorrect sequences such as [M,M,M], not shown here.

Prompt 3: Find the correct analogy
Example: SAT

Answer the question by choosing the correct option. Which of the following is an analogy?

1. beauty is to aesthete what pleasure is to hedonist
2. beauty is to aesthete what emotion is to demagogue
3. beauty is to aesthete what opinion is to sympathizer
4. beauty is to aesthete what seance is to medium
5. beauty is to aesthete what luxury is to ascetic

The answer is

F Computational and Annotation Time

Computation time. In terms of experiments, we have run a wide range of models of different sizes and settings, leading to a high computational cost. Most of the experiments have been run on a 4 40GB A100 GPUs.

We estimate the total execution time to be 100 hours overall in this infrastructure, with some experiments for small models having been run on local GPUs as well.

Annotation time. In order to annotate the SAT dataset, four annotators that have contributed as authors of the paper have dedicated an overall 80 hours, which includes the annotation and discussion processes.