Do Metaphorical Analogies Come Naturally to Language Models?

Anonymous ACL submission

Abstract

The ability to compare by analogy, metaphorically or not, lies at the core on how humans understand the world and communicate. In this paper, we study the ability 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 interpret metaphorical analogies equally well as other types of analogies, and whether the model size has an impact on this ability. The results show that most language models have a higher difficulty at identifying metaphors compared to other types of analogy. Finally, we found that there are relevant differences across models, with the larger models reducing the gap when it comes to analogical understanding, and for distinguishing metaphors from incorrect analogies.

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). At the core of analogical thinking in human cognition are metaphors (e.g. my girlfriend has a *heart of gold*), in which humans tend to 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. The reason for that is in the nature of metaphors, which are in some cases subjective and fall outside of common statistical and linguistic rules.

This may have changed 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 larger size may gain the ability to process complex analogies. Motivated by the recent advances in language modelling, we establish the following two research questions:

Research Question 1 (RQ1). How do language models understand and identify metaphors in comparison to literal and anomalous sentences? In particular, we are interested in knowing how natural LMs find both literal and metaphorical sentences, and in the context of analogies. For this, we are also interested in analysing the differences among model families and, particularly, sizes. This research question is covered in Section 5.

Research Question 2 (RQ2). Assuming differences in the answer of RQ1, we establish the second research question: How do metaphors impact the ability of language models in general analogy tests? Are language models capable of solving analogies when metaphors are involved? Section 6 presents the results for this second experiment.

In order to answer both research questions, we evaluate a broad range of language models on their ability to distinguish anomalies, analogies and metaphors on datasets that were, to our knowledge, previously unused in NLP studies. The results clearly show the marked differences in terms of perplexity between metaphors and literal analogies, in some cases metaphors being processed closer to anomalies. Despite these differences between literal and metaphorical analogies, this difference does not have a noticeable effect on their ability to solve general analogies, and especially as models grow in size.

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. Analogies can be defined as comparisons between different situations or, in the context of this paper, concepts. In this sense, all analogies capture parallels across different concepts. In or-

der to understand an analogy, we must establish a structural alignment between two representations based on their common relational structure. This process is at the core of human cognition (Johnson, 2010), and thus comes naturally to humans, but is generally challenging for computational models because mapping concepts requires to convey implicit semantic attributes. For example, understanding the statement *ketchup is to tomato what guacamole is to avocado* requires to have an internal representation of the relation *x is made of mashed y*.

Metaphors. All metaphors mediate an analogy (Black, 1977), but all analogies are not metaphors. 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). For 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, we hypothesized that language models may struggle to catch the parallelism between the concepts involved in a metaphor in comparison toother types of analogies.

Anomalies. Semantic anomalies resemble metaphors in the sense that they 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). In the context of our research, we are interested in analysing the ability of LMs to differentiate metaphors from anomalies.

3 Related Work

Metaphor processing has seen a regain of interest in the last years, with 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.

Czinczoll et al. (2022) compare the ability of transformers based language models on near analogies and and more creative analogies. They release the SCAN dataset of creative analogies. They find a large gap in the performance of the LM's between near analogies and far or creative ones. In the context of explosion of larger language models last year, 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 allow us to conveniently scale up the experiments with limited computing power.

Webb et al. (2023) study 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 find 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). Those results also suggest that abstract analogical reasoning may be an emergent ability of the larger models. This was also pointed out by (Wei et al., 2022), who showed how there is a sudden improvement in classification of fine-grained figurative language when the models are scaled up. These works were a motivation for the present study. 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 understand complex analogies.

Wijesiriwardene et al. (2023) and Sultan and Shahaf (2023) recently released resources for the identification of analogical pairs of short texts. While Sultan and Shahaf (2023) does not distinguish metaphors from other analogies, Wijesiriwardene et al. (2023) propose a scale of complexity for analogical relations, and finds that the metaphors are at the top. The open research topic of analogical reasoning between documents explored in this previous study is out of the scope of our study. Instead, we frame our experiments to explore the behavior of the models when provided 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 finetuned on analogy datasets, (Griciūtė et al., 2022; Yuan et al., 2023), the question of "natural" representation of figurativeness, as distant paralelism of semantic relations in zero-shot settings, can contribute to a better understanding of the capabilities of large language models. In this respect, the present work is more in line of perplexity-

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	Analogy	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, the number of instances. All datasets are balanced in terms of labels.

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 between models and with human ratings of figurativeness.

4 Models and Perplexity

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 (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 7. In addition to those public LMs, we consider the OpenAI commercial API models. We use GPT-3 (Brown et al., 2020a), GPT3.5 Instruct (Ouyang et al., 2022), GPT3.5 and GPT-4 (Bubeck et al., 2023).[1]

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[2]. 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), 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.

5 Language Model Representation of Metaphoric Analogies

In this section, we aim at understanding how language models 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 metaphor and literal sentences, which are presented in Section 5.1. Then, we rely exclusively on zero-shot experiments, first by computing perplexity scores without a prompt (Section 5.2) and then by studying the abilities of the models to identify metaphors by following instructions (Section 5.3). Finally, to further un-

[1] 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.

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

derstand model's behaviour, the perplexity values are compared to human ratings of figurativeness in Section 5.4.

5.1 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.

Our data have two different formats. The Cardillo and Jankowiak datasets are sentences formed from two concepts 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. The literal sentences of the form x *is-a* y, e.g. *Dogs are animals* are not analogies, when the literal quadruples are analogies. Traditionally, word analogies are represented by four words. In two-word analogies (here metaphors), the relation of interest is implicit. For example, from the sentence *His editing style was a chainsaw*, one can reconstruct the implicit 4-term analogy: *His editing style was to the text what a chainsaw is to a forest*.[3]

Cardillo. This dataset (Cardillo et al., 2010, 2017) was initially created for cognitive science studies and contains 260 pairs of x is-a y instances. Each instance in the pair is composed of one literal and one metaphoric sentence.[4] 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 analysis (see Section 5.4).

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.

Model	Data	%Low Lit.	%Low Met.	%Low Ano.
FLAN-T5$_{XXL}$	Jank.	**77.5**	19.2	3.3
	Green	**80.0**	7.5	12.5
	Cardillo	**82.3**	17.7	-
Llama2-chat$_{70B}$	Jank.	**50.8**	33.3	15.8
	Green	**80.0**	15.0	5.0
	Cardillo	**65.0**	35.0	-
Mistral-Instr$_{8x7B}$	Jank.	**45.8**	38.3	15.8
	Green	**80.0**	12.5	7.5
	Cardillo	**61.5**	38.5	-

Table 2: Percentage of literal, metaphoric and anomalous instances that have the lowest perplexity values of their respective sets in the *no prompt* setting, for a small sample of all the considered models.

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).[5] 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 Are metaphors 'natural'?

Our first experiment defines naturalness for language models, as the perplexity score of each instance, in the absence of a prompt - we refer to this setting as no-prompt (NP). In this case, the input does not contain any instructions, and only the sentences included in the datasets are fed into the model. Then, the perplexity is computed over the sentence fed into the model, as explained in Section 4.

Results. Table 2 shows the percentage of instances for which the literal, metaphoric or anomalous[6] sentences obtain the lowest perplexity for the three largest models of the Flan-T5, Llama2 and Mistral families. Except for a small number of cases, anomalies do not obtain lower perplexity than literal or metaphoric sentences, as we could anticipate. For most instances the lowest perplexity values are achieved by the literal sentences, in all cases above 45% and even as high as 82% in the Cardillo dataset for the Flan-T5$_{XXL}$ model.[7] On the other

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

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

[5] 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.

[6] Cardillo does not contain anomalous sentences, so the results are based on metaphoric and literal sentences only.

[7] The appendix includes the results for all models consid-

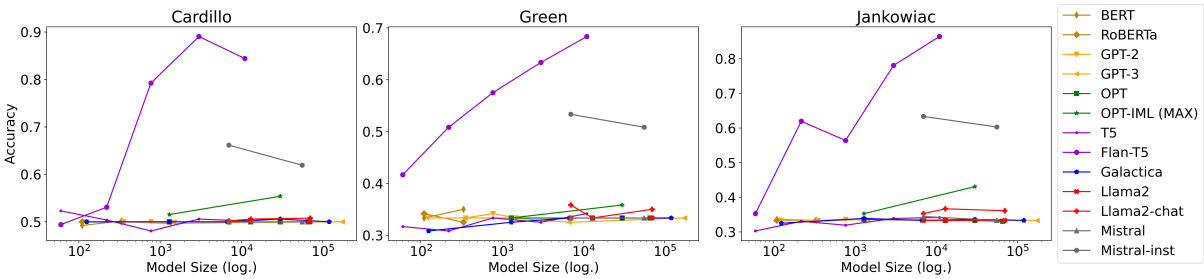

Figure 1: Accuracy of the models for the three datasets in the instruction perplexity setting.

hand, metaphoric sentences display values as low as 7.5% in the Green dataset for Flan-T5. These results clearly highlight how LMs prioritise literal over metaphoric sentences in terms of perplexity. If metaphors are pervasive in language (Lakoff and Johnson, 2003), and if some metaphorical meanings of words might occur more frequently than their literal meanings, they nevertheless remain on average less common than literal utterances.

5.3 Can LMs identify metaphors from literal and anomalous sentences?

In this setting, referred to as instruction (Ins), we explicitly ask the model to provide a response to identify literal, metaphoric and anomalous sentences. This setting is therefore different from the no-prompt setting, as in this case we provide a prompt in the form of a multiple-choice question task. In order to provide answers, models are tested with perplexity scores (Section 5.3.1) or direct generation (Section 5.3.2).

5.3.1 Perplexity experiments

In order to gather automatic responses by each model, we again rely on perplexity. In this case, we provide prompts[8] requesting a model to answer the question of which of the following candidate sentences is literal, metaphoric or anomalous[9]. Then, the perplexity of the possible answers was considered. In order to compute a result, we select the answer with the lowest perplexity among the three options (two in the case of Cardillo).

Figure 1 shows the accuracy results for various model families. Flan-T5 and Mistral models stand out in this task, especially the Flan-T5 models that obtain accuracy above 85% for the Cardillo and Jankoviac datasets, and of 68% on the Green dataset. For other models, the suboptimal results highlight the need for instruction fine-tuning on this task, as well as the importance of the model size, with clear improvements as the number of parameters of the models grows. Analysing the best performing Flan-T5$_{XXL}$, the main source of classification errors occurs between anomalies and metaphors (Figure 2), particularly in classifying anomalies as metaphors (29 instances for Green and 24 for Jankowiak for the largest XXL model). The other error types are much lower, with the lowest being predicting anomalies for literal instances (0 instances for Green and only 2 for Jankowiak for Flan-T5$_{XXL}$). This error analysis again emphasizes the clear preference of these models for literal sentences, which are better understood than metaphors or, at the very least, harder to confuse with anomalies.

5.3.2 Text generation experiments

There is no direct way to label all the elements of a set using perplexity. In the previous section, we tried to address this limitation by classifying the instances with comparing perplexities obtained with multiple prompts, but this method does not apply well to the models other than the latest Flan-T5 and Mixtral models. In this section, we turn to text generation experiments in order to perform a second evaluation of the latest generation of language models that are specialised in generation. Additionally, this setting allows us to integrate the OpenAI models for which perplexity values are not accessible. The prompt for this setting is similar to the previous section with minor variations in other to ensure a single generation output.[10] The main difference of this generation setting lies in the evaluation. Instead of relying on perplexity, 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

ered, which follow the same trend in all cases with the literal instances obtaining the lowest perplexity values overall.

[8] Full prompts available in Appendix C.1.

[9] For Cardillo, the prompts only contained literals and metaphors, as anomalies are not included in the dataset.

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

[11] Generation answers are processed semi-automatically, verifying manually those answers that do not conform exactly with the expected output.

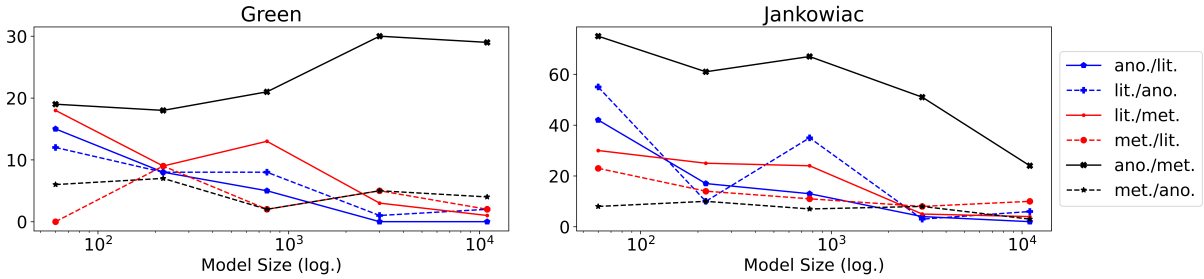

Figure 2: Error analysis of Flan-T5$_{XXL}$ in the instruction-perplexity setting on the Green and Jankowiak datasets. The Y-axis indicates the number of mislabeled instances. In the legend, predicted values are on the left and gold standard values on the right, e.g., ano/lit indicates the literal sentences that were predicted as anomalies by Flan-T5.

Model	Setting	Card.	Jank.	Green
FLAN-T5$_{XXL}$	*ppl*	84.4	86.4	68.3
	gen	78.9	57.4	37.6
Llama2-chat$_{70B}$	*ppl*	50.8	36.1	35
	gen	85.6	73.6	56.4
Mistral-Instr$_{8x7B}$	*ppl*	61.9	60.3	50.8
	gen	76.5	84.1	55.3
GPT-3.5$_{turbo-inst.}$	*gen*	65.9	61.5	38.8
GPT-3.5$_{turbo}$	*gen*	70.5	59.8	41.2
GPT-4	*gen*	**91.8**	**91.4**	**78.6**
Random		50.0	33.3	33.3

Table 3: Accuracy of the generated answers for the three datasets Cardillo, Jankowiak and Green in the instruction setting using perplexity setting (*ppl*) and in the generation setting (*gen*).

Model	**Jank.**		**Green**	
	MA	**LA**	**MA**	**LA**
FLAN-T5$_{XXL}$	**521**	116	220	15
Llama2-chat$_{70B}$	**345**	99	99	94
Mistral-Instr$_{8x7B}$	**141**	75	123	60
GPT-3.5$_{turbo-inst.}$	**433**	138	165	136
GPT-3.5$_{turbo}$	**450**	234	140	143
GPT-4	**89**	18	48	14

Table 4: Error analysis for the Jankowiak and Green datasets in the generation setting (*gen*). The non-directional confusion between the *metaphor* and *anomaly* labels (MA), and the confusion between the *literal* and *anomaly* labels (LA) are shown for all the models evaluation on generation.

sentences within each set (shuffling the order in which literal, metaphoric and anomalous sentences are presenting 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 3, along with the perplexity-based instruction results for completness. In this setting, Flan-T5$_{XXL}$ loses its advantage over the Llama2 and Mistral models. All the models have difficulties to process the Green set, 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 4. 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.

[12]This bias is reported in the Appendix (Tables 14 and 13).

5.4 Correlation between perplexity and figurativeness

Humans often perceive sentences as more of less metaphoric, rather than merely as binary categories. As explained in Section 5.1, Cardillo et al. (2017) enriched their dataset with human scores for each instance according to *figurativeness*. We study the correlation between all the previously obtained perplexities in the *no-prompt* and *instruction settings* and the human judgments of figurativeness using Spearman correlation ρ.

As shown in Figure 3, all models correlate positively with figurativeness in the no-prompt setting. Reminder that this setting computes the perplexity of sentences without any addition. In this case, a positive correlation means that sentence which are more figurative, tend to be less natural according to the LMs. Therefore, even if this correlation is in some cases weak, the fact that it is in all cases positive reinforces the main takeaway from the previous perplexity experiments. For the instruction setting, if we focus on the best performing model, Flan T5, we can observe exactly opposite trends for the metaphor and literal settings. FLAN-T5$_{XXL}$ obtains a Spearman correlation score of 0.69 in the instruction-metaphor setting, which indicates a

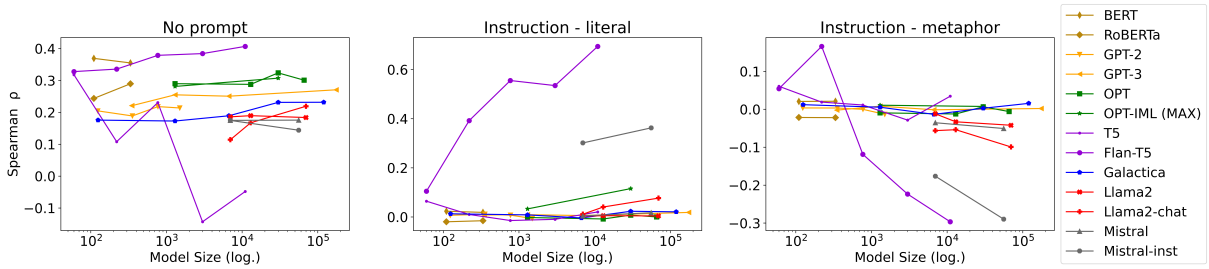

Figure 3: Correlation with human judgment for the no prompt (NP) setting and the instruction setting (*ppl*) on the Cardillo dataset.

strong correlation. The correlation of other models in this setting is weaker, but this is the general case for the overall results (see Section 5.3.1).

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

In the previous section, we tested the capabilities of language models in explicitly recognizing metaphors. The results show how models struggle to understand metaphors or, at the very least, find them less natural than literal sentences. A natural question that may arise is whether this behavior has an effect in 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 no.

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 understand analogies (Brown et al., 2020b; Ushio et al., 2021b,a; Chen et al., 2022; Kumar and Schockaert, 2023). One advantage of this dataset over other benchmarks is that the dataset is not openly available in the web, which mitigates possible concerns of data contamination in LMs. Each set in SAT contains a word pair as input, and five candidates: a correct analogy and four anomalies. The task consists of selecting the correct analogy given the initial word pair as an input.

SAT annotation Each of the 374 questions of the SAT dataset contains four incorrect analogies and a single correct one. Among those right answers, there are metaphoric analogies. 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 analyzed has 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.

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

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 F). 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 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 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 labeling 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

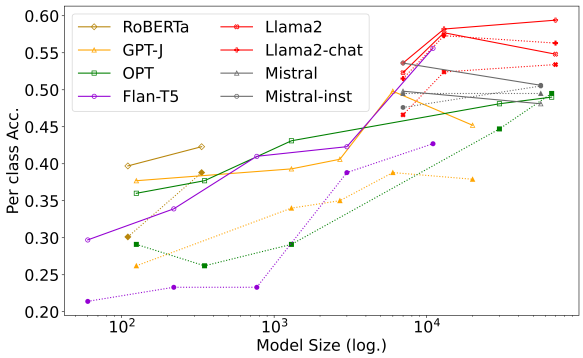

Figure 4: Accuracy results of the no-prompt setting experiment on SAT. The results for the metaphoric class are displayed in the dashed lines, while the results for the non-metaphoric class are shown in the solid lines.

the scores of the four participants. We labeled as non-metaphoric all the quadruples scoring lower to 1 on average and metaphoric all ones scoring above 1. 32 instances with an average score of 1 were filtered out. Table 5 contains an example of a metaphoric instance of the SAT dataset after annotation. In total, 103 instances were labeled as metaphoric, and 239 as non-metaphoric.

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 in a *no-prompt* setting, following the same methodology outlined in Section 5.2. In this case, out of the five choices, the instance with a lowest perplexity is selected as the correct option. In addition, the large instructed LMs are tested through text generation.[13] Then, we simply report the accuracy on the metaphor and non-metaphoric subsets of SAT.

Results. Figure 4 shows the accuracy on the metaphorical and non-metaphorical subsets of SAT in the *no prompt* setting.[14] In general, model performances improve 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.

Table 6 shows the results of the generation experiments for the large instructed models. Unlike in previous experiments, all models perform relatively similarly for both the metaphor and anomaly

[13] As in Section 5.3.2, 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 F.2.

[14] See Table 11 in the Appendix for the full results.

Model	Setting	Non-met	Met
FLAN-T5$_{XXL}$	NP	55.6	42.7
	gen	41.6	44.5
Llama2-chat$_{70B}$	NP	59.4	56.3
	gen	41.0	49.5
Mistral-Instr$_{8x7B}$	NP	50.6	50.5
	gen	50.0	57.5
GPT3.5$_{turbo}$	gen	28.5	32.6
GPT4	gen	72.6	75.0

Table 6: Accuracy results in the perplexity no-prompt (NP) and generation settings (gen) for the non-metaphor and metaphor classes in SAT.

classes in the generation setting. Moreover, we can again observe the best performance for GPT4, although the conclusions that can be drawn from this model are limited due to its closed nature. In general, these results show that, while all the models have preference for literal sentences and consider them to be more natural, this does not have an effect in how the new generation of larger language models identify analogies more generally.

7 Conclusion

In this paper, we have analysed the capabilities of LMs to perceive and identify metaphors. A simple perplexity experiment reveals how metaphors are perceived as less natural (or less likely) as compared to literal sentences. In many contexts, metaphors can be seen closer to anomalous or incorrect sentences than to literal sentences. This behavior is consistent across a wide range of model families and sizes analysed.

As a result of this finding, we also investigated whether these results would be reflected on how models can separate 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, although they still remain as the main source of errors. In other words, large models appear to be able to understand metaphors, even if they would not prioritise over more literal sentences.

These results leave open several follow-up questions not explored in this paper. What is the role of metaphors in generational 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 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 use analogies as proxy to better understand the behaviour of language models, and particularly, in relation to metaphors. Therefore, 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 be also applied to language models. In relation to this, we have not attempted to perform extensive prompt engineering in this work, as we were rather 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 G 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 with 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 are performed only for English, as such, the conclusions may differ for other languages, especially less-resourced and languages from other families. Finally, data contamination may have an effect in the results, which we could not analyse extensively. To mitigate this, we considered datasets not openly available as well as annotated existing ones, having the certainty that these new annotations are therefore not seen by any of the models.

References

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

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.

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.

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

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.

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.

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.

Mark Johnson. 2010. *Metaphor and Cognition*, pages 401–414. Springer Netherlands, Dordrecht.

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.

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. 2003. *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.

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*.

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.

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.

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.

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.

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

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, Ruwan Wickramarachchi, Bimal G. Gajera, Shreeyash Mukul Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. 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.

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 7.

B No-prompt setting experiments result tables

Tables 8 and 9 include the full experimental results of Section 5.2. Specifically, Table 8 showcases the proportions of sets where sentences with literal, metaphoric, and anomalous content exhibit the lowest perplexity in a no-prompt context for both the Cardillo dataset (containing only literal and metaphors) and the Jankowiak dataset. Similar findings for the Green dataset are available in Table 9.

C Instruction setting experiments

In the instruction setting, models are evaluated through perplexity computation and text generation.

C.1 Instruction prompt template

With n the number of distinct labels in a dataset, $\$label$ a variable taking the value each possible label, <*literal*, *metaphoric*, *difficult to interpret*>[15], and $\$i$ the sentence index, it is a template for the creation of n^2 question-answer (QA) pairs for a given set, as exemplified below for the Cardillo dataset:

> Which of the following examples is $\$label$?
>
> (a) His thoughts were a pendulum.
>
> (b) The device was a pendulum.
>
> The answer is $\$i$

C.2 Ratio of each label obtaining the lowest perplexity of a set

The results for those experiments in instruction setting are given in Table 10 for the Cardillo dataset and in Table 12 for the Jankowiak and Green datasets.

D Generation experiments

In this section we provide details for the generation experiments presented in Section 5.3.2.

D.1 Prompt used in the generation experiments (P2)

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

[15] We substitute a more common and less ambiguous English paraphrase to the word *anomalous*

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

D.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 14 and 13. As a consequence, we run the experiments with all possible permutations of the sentences within each set, making distribution of label sequences uniform.

E Experiments on the SAT dataset

Table 11 shows a comparison of the models on the task of solving the analogy questions of SAT in the *no prompt* 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.

F Annotation Guidelines for Adding Metaphoricity 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*.

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

F.1 No-prompt experiments

F.2 Generation experiment prompts

Prompt G2 . The correct answer of the example below is 1., it is classified non metaphoric in SAT.

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

G 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 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 including the annotation and discussion processes.

	Model	Size	Name on HuggingFace
Masked LM	$BERT_{BASE}$	110M	bert-base-cased
	$BERT_{LARGE}$	355M	bert-large-cased
	$RoBERTa_{BASE}$	110M	roberta-base
	$RoBERTa_{LARGE}$	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
	Flan-$T5_{SMALL}$	60M	google/flan-t5-small
	Flan-$T5_{BASE}$	220M	google/flan-t5-base
	Flan-$T5_{LARGE}$	770M	google/flan-t5-large
	Flan-$T5_{XL}$	3B	google/flan-t5-xl
	Flan-$T5_{XXL}$	11B	google/flan-t5-xxl
	Flan-UL2	20B	google/flan-ul2
	UL2	20B	google/ul2
Decoder-only LM	GPT-2_{SMALL}	124M	gpt2
	GPT-2_{BASE}	355M	gpt2-medium
	GPT-2_{LARGE}	774M	gpt2-large
	GPT-2_{XL}	1.5B	gpt2-xl
	GPT-J_{125M}	125M	EleutherAI/gpt-neo-125M
	GPT-$J_{1.3B}$	1.3B	EleutherAI/gpt-neo-1.3B
	GPT-$J_{2.7B}$	2.7B	EleutherAI/gpt-neo-2.7B
	GPT-J_{6B}	6B	EleutherAI/gpt-j-6B
	GPT-J_{20B}	20B	EleutherAI/gpt-neox-20b
	OPT_{125M}	125M	facebook/opt-125m
	OPT_{350M}	350M	facebook/opt-350m
	$OPT_{1.3B}$	1.3B	facebook/opt-1.3b
	OPT_{13B}	13B	facebook/opt-13b
	OPT_{30B}	30B	facebook/opt-30b
	OPT_{66B}	66B	facebook/opt-66b
	OPT-$IML_{1.3B}$	1.3B	facebook/opt-iml-1.3b
	OPT-IML_{30B}	30B	facebook/opt-iml-30b
	OPT-$IML_{M\text{-}1.3B}$	1.3B	facebook/opt-iml-max-1.3b
	OPT-$IML_{M\text{-}30B}$	30B	facebook/opt-iml-max-30b
	$Galactica_{125m}$	125M	facebook/galactica-125m
	$Galactica_{1.3B}$	1.3BM	facebook/galactica-1.3b
	$Galactica_{6.7B}$	6.7B	facebook/galactica-6.7b
	$Galactica_{30B}$	30B	facebook/galactica-30b
	$Galactica_{120B}$	120B	facebook/galactica-120b
	$Bloom_{176B}$	176B	bigscience/bloom
	$Bloomz_{176B}$	176B	bigscience/bloomz
	$Llama2_{7B}$	7B	meta-llama/Llama-2-7b-hf
	$Llama2_{13B}$	13B	meta-llama/Llama-2-13b-hf
	$Llama2_{70B}$	70B	meta-llama/Llama-2-70b-hf
	Llama2-$chat_{7B}$	7B	meta-llama/ Llama-2-7b-chat-hf
	Llama2-$chat_{13B}$	13B	meta-llama/ Llama-2-13b-chat-hf
	Llama2-$chat_{70B}$	70B	meta-llama/ Llama-2-70b-chat-hf
	$Mistral_{7B}$	7B	mistralai/Mistral-7B-v0.1
	Mistral-$Instruct_{7B}$	7B	mistralai/ Mistral-7B-Instr.-v0.2
sMoE	$Mistral_{8x7B}$	56B	mistralai/Mistral-8x7B-v0.1
	Mistral-$Instruct_{8x7B}$	56B	mistralai/ Mistral-8x7B-Instr.-v0.1

Table 7: The model checkpoints used in the LM baselines on HuggingFace model hub. All the model can be obtained at https://huggingface.co.

Model Family	Model	Cardillo		Jankowiak			
		%lit. is lowest	%met is lowest	%lit. is lowest	%met. is lowest	%ano. is lowest	l<m<a (%)
BERT	$BERT_{BASE}$	73.5	26.5	85.8	9.2	5.0	47.5
	$BERT_{LARGE}$	72.3	27.7	80.0	12.5	7.5	45.8
RoBERTa	$RoBERTa_{BASE}$	64.6	35.4	53.3	34.2	12.5	32.5
	$RoBERTa_{LARGE}$	70.0	30.0	43.3	43.3	13.3	30.0
GPT-2	$GPT\text{-}2_{SMALL}$	60.8	39.2	58.3	25.8	15.8	42.5
	$GPT\text{-}2_{MEDIUM}$	62.7	37.3	36.7	50.8	12.5	26.7
	$GPT\text{-}2_{LARGE}$	61.9	38.1	41.7	44.2	14.2	30.0
	$GPT\text{-}2_{XL}$	63.8	36.2	35.8	52.5	11.7	25.8
GPT-J	$GPT\text{-}J_{125M}$	56.5	43.5	21.7	62.5	15.8	15.8
	$GPT\text{-}J_{1.3B}$	61.9	38.1	35.8	53.3	10.8	27.5
	$GPT\text{-}J_{2.7B}$	57.3	42.7	31.7	55.8	12.5	22.5
	$GPT\text{-}J_{6B}$	62.7	37.3	38.3	52.5	9.2	28.3
	$GPT\text{-}J_{20B}$	61.5	38.5	50.0	37.5	12.5	37.5
GPT-3	$GPT\text{-}3_{ada}$	63.1	36.9	54.2	35.8	10.0	40.0
	$GPT\text{-}3_{babbage}$	67.3	32.7	50.0	40.0	10.0	40.0
	$GPT\text{-}3_{curie}$	67.7	32.3	51.7	41.7	6.7	35.8
	$GPT\text{-}3_{davinci}$	67.7	32.3	49.2	43.3	7.5	34.2
OPT	OPT_{125M}	64.2	35.8	44.2	30.0	25.8	21.7
	OPT_{350M}	63.1	36.9	36.7	45.8	17.5	19.2
	$OPT_{1.3B}$	68.5	31.5	40.8	44.2	15.0	25.0
	OPT_{13B}	68.5	31.5	52.5	36.7	10.8	36.7
	OPT_{30B}	68.5	31.5	48.3	40.8	10.8	35.8
	OPT_{66B}	66.9	33.1	43.3	43.3	13.3	27.5
OPT-IML	$OPT\text{-}IML_{1.3B}$	67.3	32.7	40.0	42.5	17.5	26.7
	$OPT\text{-}IML_{30B}$	69.6	30.4	44.2	42.5	13.3	27.5
OPT-IML (MAX)	$OPT\text{-}IML_{M\text{-}1.3B}$	65.8	34.2	41.7	43.3	15.0	25.8
	$OPT\text{-}IML_{M\text{-}30B}$	70.4	29.6	46.7	42.5	10.8	30.8
T5	$T5_{SMALL}$	76.9	23.1	70.8	11.7	17.5	35.0
	$T5_{BASE}$	66.5	33.5	79.2	11.7	9.2	44.2
	$T5_{LARGE}$	67.7	32.3	29.2	34.2	36.7	7.5
	$T5_{3B}$	42.3	57.7	33.3	40.8	25.8	19.2
	$T5_{11B}$	50.8	49.2	49.2	34.2	16.7	25.8
UL2	UL2	61.5	38.5	65.8	22.5	11.7	33.3
Flan-T5	$Flan\text{-}T5_{SMALL}$	78.8	21.2	85.8	9.2	5.0	56.7
	$Flan\text{-}T5_{BASE}$	77.7	22.3	84.2	9.2	6.7	47.5
	$Flan\text{-}T5_{LARGE}$	80.0	20.0	52.5	34.2	13.3	25.0
	$Flan\text{-}T5_{XL}$	77.3	22.7	81.7	15.0	3.3	56.7
	$Flan\text{-}T5_{XXL}$	82.3	17.7	77.5	19.2	3.3	55.0
Flan-UL2	Flan-UL2	80.8	19.2	73.3	21.7	5.0	45.0
Galactica	$Galactica_{125M}$	65.0	35.0	80.8	12.5	6.7	47.5
	$Galactica_{1.3B}$	65.4	34.6	59.2	28.3	12.5	38.3
	$Galactica_{6.7}$	64.2	35.8	58.3	25.8	15.8	39.2
	$Galactica_{30B}$	66.9	33.1	60.8	29.2	10.0	40.8
	$Galactica_{120B}$	64.6	35.4	66.7	25.8	7.5	47.5
Bloom	$Bloom_{176B}$	61.9	38.1	52.5	39.2	8.3	34.2
Bloomz	$Bloomz_{176B}$	66.5	33.5	60.8	30.0	9.2	37.5
Llama2	$Llama2_{7B}$	63.1	36.9	52.5	33.3	14.2	29.2
	$Llama2_{13B}$	63.5	36.5	47.5	35.8	16.7	25.8
	$Llama2_{70B}$	60.8	39.2	50.0	35.0	15.0	27.5
Llama2-chat	$Llama2\text{-}chat_{7B}$	57.3	42.7	50.0	36.7	13.3	26.7
	$Llama2\text{-}chat_{13B}$	63.1	36.9	40.8	45.0	14.2	20.8
	$Llama2\text{-}chat_{70B}$	65.0	35.0	50.8	33.3	15.8	35.0
Mistral	$Mistral_{7B}$	65.0	35.0	45.0	37.5	17.5	26.7
	$Mistral_{8x7B}$	62.7	37.3	48.3	38.3	13.3	27.5
Mistral-inst	$Mistral\text{-}Instruct_{7B}$	64.6	35.4	45.0	36.7	18.3	30.0
	$Mistral\text{-}Instruct_{8x7B}$	61.5	38.5	45.8	38.3	15.8	27.5

Table 8: Ratios of instances for which the literal sentences, the metaphoric sentences and the anomalous sentences, have the lowest perplexity in the Cardillo (for literal examples and metaphors only) and Jankowiak datasets according to model family and size (*no-prompt* setting).

Model Family	Model	%Lit. is lowest	%Met. is lowest	%Ano. is lowest	l<m<a (%)
BERT	$BERT_{BASE}$	65.0	15.0	20.0	30.0
	$BERT_{LARGE}$	70.0	12.5	17.5	47.5
RoBERTa	$RoBERTa_{BASE}$	80.0	2.5	17.5	52.5
	$RoBERTa_{LARGE}$	80.0	10.0	10.0	45.0
GPT-2	$GPT\text{-}2_{SMALL}$	75.0	10.0	15.0	35.0
	$GPT\text{-}2_{MEDIUM}$	70.0	15.0	15.0	42.5
	$GPT\text{-}2_{LARGE}$	72.5	12.5	15.0	45.0
	$GPT\text{-}2_{XL}$	85.0	5.0	10.0	45.0
GPT-J	$GPT\text{-}J_{125M}$	60.0	12.5	27.5	27.5
	$GPT\text{-}J_{1.3B}$	80.0	15.0	5.0	40.0
	$GPT\text{-}J_{2.7B}$	82.5	2.5	15.0	47.5
	$GPT\text{-}J_{6B}$	87.5	5.0	7.5	55.0
	$GPT\text{-}J_{20B}$	85.0	7.5	7.5	47.5
GPT-3	$GPT\text{-}3_{ADA}$	77.5	12.5	10.0	45.0
	$GPT\text{-}3_{BABBAGE}$	77.5	10.0	12.5	45.0
	$GPT\text{-}3_{CURIE}$	87.5	2.5	10.0	47.5
	$GPT\text{-}3_{DAVINCI}$	92.5	2.5	5.0	62.5
OPT	OPT_{125M}	77.5	7.5	15.0	37.5
	OPT_{350M}	77.5	5.0	17.5	40.0
	$OPT_{1.3B}$	92.5	2.5	5.0	52.5
	OPT_{13B}	95.0	2.5	2.5	55.0
	OPT_{30B}	**97.5**	0.0	2.5	60.0
	OPT_{66B}	**97.5**	0.0	2.5	57.5
OPT-IML	$OPT\text{-}IML_{1.3B}$	90.0	2.5	7.5	52.5
	$OPT\text{-}IML_{30B}$	90.0	2.5	7.5	52.5
OPT-IML (MAX)	$OPT\text{-}IML_{M\text{-}1.3B}$	85.0	2.5	12.5	45.0
	$OPT\text{-}IML_{M\text{-}30B}$	97.5	0.0	2.5	57.5
T5	$T5_{SMALL}$	95.0	0.0	5.0	45.0
	$T5_{BASE}$	62.5	17.5	20.0	27.5
	$T5_{LARGE}$	72.5	10.0	17.5	45.0
	$T5_{3B}$	70.0	7.5	22.5	25.0
	$T5_{11B}$	80.0	5.0	15.0	32.5
UL2	UL2	57.5	12.5	30.0	32.5
Flan-T5	$Flan\text{-}T5_{SMALL}$	87.5	0.0	12.5	40.0
	$Flan\text{-}T5_{BASE}$	87.5	5.0	7.5	35.0
	$Flan\text{-}T5_{LARGE}$	85.0	5.0	10.0	30.0
	$Flan\text{-}T5_{XL}$	**92.5**	2.5	5.0	40.0
	$Flan\text{-}T5_{XXL}$	80.0	7.5	12.5	45.0
Flan-UL2	Flan-UL2	85.0	7.5	7.5	55.0
Galactica	$Galactica_{125M}$	65.0	10.0	25.0	32.5
	$Galactica_{1.3B}$	75.0	10.0	15.0	32.5
	$Galactica_{6.7}$	80.0	12.5	7.5	32.5
	$Galactica_{30B}$	85.0	7.5	7.5	45.0
	$Galactica_{120B}$	85.0	10.0	5.0	42.5
Bloom	$Bloom_{176B}$	80.0	5.0	15.0	52.5
Bloomz	$Bloomz_{176B}$	87.5	2.5	10.0	55.0
Llama2	$Llama2_{7B}$	80.0	15.0	5.0	50.0
	$Llama2_{13B}$	82.5	10.0	7.5	60.0
	$Llama2_{70B}$	77.5	17.5	5.0	55.0
Llama2-chat	$Llama2\text{-}chat_{7B}$	82.5	10.0	7.5	60.0
	$Llama2\text{-}chat_{13B}$	90.0	5.0	5.0	62.5
	$Llama2\text{-}chat_{70B}$	80.0	15.0	5.0	62.5
Mistral	$Mistral_{7B}$	82.5	7.5	10.0	52.5
	$Mistral_{8x7B}$	82.5	10.0	7.5	60.0
Mistral-Instr.	$Mistral\text{-}Instr._{7B}$	82.5	10.0	7.5	62.5
	$Mistral\text{-}Instr._{8x7B}$	80.0	12.5	7.5	52.5

Table 9: Ratios of instances for which the literal sentences on the left, and the metaphoric sentences in the middle and the anomalous sentences on the right, have the lowest perplexity in the Green dataset according to model family and size (*no-prompt* setting).

Model	%Lit. is lowest	%Met. is lowest
BERT$_{BASE}$	52.3	48.8
BERT$_{LARGE}$	49.6	50.4
RoBERTa$_{BASE}$	49.6	50.8
RoBERTa$_{LARGE}$	49.6	50.4
GPT-2$_{SMALL}$	50.4	49.6
GPT-2$_{MEDIUM}$	50.4	49.6
GPT-2$_{LARGE}$	50.4	49.6
GPT-2$_{XL}$	50.4	49.6
GPT-J$_{125M}$	50.4	49.6
GPT-J$_{1.3B}$	50.4	49.6
GPT-J$_{2.7B}$	50.4	49.6
GPT-J$_{6B}$	57.3	51.5
GPT-J$_{20B}$	49.6	50.4
GPT-3$_{ADA}$	50.4	49.6
GPT-3$_{BABBAGE}$	50.4	49.6
GPT-3$_{CURIE}$	47.7	49.2
GPT-3$_{DAVINCI}$	47.7	56.2
OPT$_{1.3B}$	50.0	49.6
OPT$_{13B}$	50.4	49.6
OPT$_{30B}$	51.2	48.8
OPT$_{66B}$	50.4	49.6
OPT-IML$_{1.3B}$	50.4	50.4
OPT-IML$_{30B}$	66.2	47.3
OPT-IML$_{M\text{-}1.3B}$	55.0	43.8
OPT-IML$_{M\text{-}30B}$	76.5	46.2
T5$_{SMALL}$	50.4	49.6
T5$_{BASE}$	50.4	49.6
T5$_{LARGE}$	50.0	49.6
T5$_{3B}$	49.2	52.7
T5$_{11B}$	50.0	49.6
UL2	51.5	50.4
Flan-T5$_{SMALL}$	50.8	50.8
Flan-T5$_{BASE}$	67.3	32.3
Flan-T5$_{LARGE}$	76.2	52.3
Flan-T5$_{XL}$	73.8	56.2
Flan-T5$_{XXL}$	87.7	62.3
Flan-UL2	79.2	69.6
Galactica$_{125M}$	50.4	49.6
Galactica$_{1.3B}$	50.4	49.6
Galactica$_{6.7B}$	50.8	49.2
Galactica$_{30B}$	56.2	56.5
Galactica$_{120B}$	45.4	41.2
Bloom$_{176B}$	50.4	43.5
Bloomz$_{176B}$	68.1	71.2
Llama2$_{7B}$	50.4	50.8
Llama2$_{13B}$	51.5	63.8
Llama2$_{70B}$	50.0	66.9
Llama2-chat$_{7B}$	52.3	72.7
Llama2-chat$_{13B}$	59.6	68.8
Llama2-chat$_{70B}$	92.3	78.8
Mistral$_{7B}$	53.1	52.7
Mistral$_{8x7B}$	83.5	68.8
Mistral-Instruct$_{7B}$	86.5	66.9
Mistral-Instruct$_{8x7B}$	**96.2**	80.0

Table 10: Ratios of instances for which the literal sentences, the metaphoric sentences have the lowest perplexity in the Cardillo dataset according to model family and size. Instruction-literal (The prompt contains the question of)ratios are shown in the first column (%Lit. is lowest) and Instruction-metaphor ratios are shown in the second result column (%Met. is lowest).

Model Family	Model	% correct non met.	% correct met.
BERT	BERT$_{BASE}$	34.7	29.1
	BERT$_{LARGE}$	34.3	25.2
RoBERTa	RoBERTa$_{BASE}$	39.7	30.1
	RoBERTa$_{LARGE}$	42.3	38.8
GPT-2	GPT-2$_{SMALL}$	34.3	26.2
	GPT-2$_{MEDIUM}$	36.4	31.1
	GPT-2$_{LARGE}$	38.1	30.1
	GPT-2$_{XL}$	37.7	35.0
GPT-J	GPT-J$_{125M}$	37.7	26.2
	GPT-J$_{1.3B}$	39.3	34.0
	GPT-J$_{2.7B}$	40.6	35.0
	GPT-J$_{6B}$	49.8	38.8
	GPT-J$_{20B}$	45.2	37.9
GPT-3	GPT-3$_{davinci}$	50.6	**55.3**
OPT	OPT$_{125M}$	36.0	29.1
	OPT$_{350M}$	37.7	26.2
	OPT$_{1.3B}$	43.1	29.1
	OPT$_{30B}$	48.1	44.7
	OPT$_{66B}$	49.0	**49.5**
OPT-IML	OPT-IML$_{1.3B}$	43.9	37.9
	OPT-IML$_{30B}$	50.6	44.7
OPT-IML (MAX)	OPT-IML$_{M\text{-}1.3B}$	43.1	35.0
	OPT-IML$_{M\text{-}30B}$	51.9	44.7
T5	T5$_{SMALL}$	29.3	27.2
	T5$_{BASE}$	29.3	20.4
	T5$_{LARGE}$	32.6	28.2
	T5$_{3B}$	36.8	30.1
	T5$_{11B}$	39.7	27.2
Flan-T5	Flan-T5$_{SMALL}$	29.7	21.4
	Flan-T5$_{BASE}$	33.9	23.3
	Flan-T5$_{LARGE}$	41.0	23.3
	Flan-T5$_{XL}$	42.3	38.8
	Flan-T5$_{XXL}$	55.6	42.7
Flan-UL2	Flan-UL2	50.6	47.6
Galactica	Galactica$_{125M}$	23.8	15.5
	Galactica$_{1.3B}$	34.7	28.2
	Galactica$_{6.7}$	37.7	26.2
	Galactica$_{30B}$	42.7	34.0
	Galactica$_{120B}$	39.3	32.0
Llama2	Llama2$_{7B}$	52.3	46.6
	Llama2$_{13B}$	57.7	52.4
	Llama2$_{70B}$	54.8	53.4
Llama2-chat	Llama2-chat$_{7B}$	53.6	51.5
	Llama2-chat$_{13B}$	58.2	57.3
	Llama2-chat$_{70B}$	59.4	56.3
Mistral	Mistral$_{7B}$	49.8	49.5
	Mistral$_{8x7B}$	48.1	**49.5**
Mistral-Instr.	Mistral-Instr.$_{7B}$	53.6	47.6
	Mistral-Instr.$_{8x7B}$	50.6	50.5

Table 11: 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 (no-prompt setting).

Model	Green			Jankowiak		
	%Lit. is lowest	%Met. is lowest	%Ano. is lowest	%Lit. is lowest	%Met. is lowest	%Ano. is lowest
BERT$_{BASE}$	25.0	20.0	32.5	28.3	30.0	38.3
BERT$_{LARGE}$	30.0	30.0	42.5	29.2	36.7	44.2
RoBERTa$_{BASE}$	22.5	25.0	50.0	33.3	33.3	27.5
RoBERTa$_{LARGE}$	22.5	42.5	35.0	39.2	35.8	25.0
GPT-2$_{SMALL}$	42.5	37.5	20.0	30.8	29.2	40.0
GPT-2$_{MEDIUM}$	42.5	37.5	20.0	30.8	29.2	40.0
GPT-2$_{LARGE}$	42.5	37.5	20.0	30.8	29.2	40.0
GPT-2$_{XL}$	42.5	37.5	20.0	30.8	29.2	40.0
GPT-J$_{125M}$	42.5	37.5	20.0	30.8	29.2	43.3
GPT-J$_{1.3B}$	42.5	37.5	20.0	30.8	29.2	40.0
GPT-J$_{2.7B}$	35.0	25.0	40.0	39.2	25.8	36.7
GPT-J$_{6B}$	37.5	25.0	35.0	42.5	34.2	26.7
GPT-J$_{20B}$	42.5	35.0	30.0	36.7	35.8	30.0
GPT-3$_{ADA}$	40.0	37.5	22.5	30.8	29.2	40.0
GPT-3$_{BABBAGE}$	42.5	37.5	20.0	30.8	29.2	40.0
GPT-3$_{CURIE}$	20.0	30.0	32.5	39.2	33.3	34.2
GPT-3$_{DAVINCI}$	40.0	37.5	32.5	32.5	28.3	41.7
OPT$_{1.3B}$	42.5	37.5	27.5	30.0	29.2	41.7
OPT$_{13B}$	37.5	32.5	37.5	20.0	35.0	31.7
OPT$_{30B}$	42.5	37.5	22.5	30.8	29.2	39.2
OPT$_{66B}$	42.5	37.5	20.0	30.8	29.2	40.0
OPT-IML$_{1.3B}$	47.5	40.0	20.0	30.8	36.7	40.0
OPT-IML$_{30B}$	70.0	27.5	17.5	66.7	34.2	46.7
OPT-IML$_{M\text{-}1.3B}$	60.0	40.0	5.0	53.3	24.2	29.2
OPT-IML$_{M\text{-}30B}$	65.0	40.0	25.0	83.3	34.2	67.5
T5$_{SMALL}$	42.5	37.5	20.0	30.8	28.3	40.0
T5$_{BASE}$	40.0	37.5	20.0	30.8	29.2	40.0
T5$_{LARGE}$	32.5	37.5	35.0	30.8	29.2	40.0
T5$_{3B}$	32.5	22.5	42.5	27.5	35.8	42.5
T5$_{11B}$	40.0	37.5	20.0	33.3	29.2	35.0
UL2	37.5	32.5	35.0	33.3	30.8	38.3
Flan-T5$_{SMALL}$	35.0	27.5	37.5	50.0	26.7	20.8
Flan-T5$_{BASE}$	55.0	22.5	32.5	62.5	34.2	26.7
Flan-T5$_{LARGE}$	85.0	30.0	30.0	67.5	56.7	47.5
Flan-T5$_{XL}$	82.5	17.5	12.5	84.2	46.7	50.0
Flan-T5$_{XXL}$	92.5	20.0	47.5	88.3	69.2	72.5
Flan-UL2	75.0	37.5	30.0	72.5	73.3	69.2
Galactica$_{125M}$	42.5	37.5	20.0	30.8	29.2	40.8
Galactica$_{1.3B}$	42.5	37.5	20.0	30.8	29.2	40.0
Galactica$_{6.7B}$	52.5	22.5	17.5	27.5	40.0	29.2
Galactica$_{30B}$	60.0	40.0	20.0	59.2	42.5	21.7
Galactica$_{120B}$	42.5	22.5	27.5	34.2	30.8	20.0
Bloom$_{176B}$	45.0	20.0	17.5	30.8	30.8	31.7
Bloomz$_{176B}$	65.0	55.0	40.0	49.2	68.3	36.7
Llama2$_{7B}$	42.5	37.5	20.0	30.8	29.2	40.0
Llama2$_{13B}$	45.0	35.0	20.0	30.8	33.3	44.2
Llama2$_{70B}$	42.5	60.0	22.5	30.0	51.7	53.3
Llama2-chat$_{7B}$	35.0	55.0	15.0	60.8	66.7	38.3
Llama2-chat$_{13B}$	35.0	70.0	35.0	75.0	63.3	51.7
Llama2-chat$_{70B}$	57.5	60.0	57.5	92.5	75.8	51.7
Mistral$_{7B}$	42.5	37.5	20.0	37.5	46.7	41.7
Mistral$_{8x7B}$	42.5	67.5	50.0	80.0	77.5	58.3
Mistral-Instruct$_{7B}$	75.0	42.5	65.0	80.8	75.8	73.3
Mistral-Instruct$_{8x7B}$	72.5	52.5	67.5	92.5	79.2	78.3

Table 12: Ratios of instances for which the literal sentences, the metaphoric sentences and the anomalous sentences have the lowest perplexity in the Green and Jankowiak datasets according to model family and size Instruction-literal ($P1_L$) ratios are shown in the first column (%Lit. is lowest), Instruction-metaphor ($P1_M$) ratios are shown in the second result column (%Met. is lowest) and Instruction-anomaly ($P1_A$) ratios are shown in the third result column (%Ano. is lowest) for both sets.

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
	Mistral-Instr.$_{8x7B}$	**42.5**	35.0	4.6	6.2	0.8	2.1
	GPT-3.5$_{turbo\text{-}instr.}$	**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
	Mistral-Instr.$_{8x7B}$	20.1	18.5	18.9	16.9	8.1	13.9
	GPT-3.5$_{turbo\text{-}instr.}$	**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 the same number of times. Flan-T5$_{XXL}$ label distribution is far from summing to 100 because according to the table because it outputs a large proportion of incorrect sequences such as [M,M,M], not shown here.

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

Table 14: 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 the same number of times in each position.