Analogy Generation by Prompting Large Language Models: A Case Study of InstructGPT

Bhavya Bhavya[1], Jinjun Xiong[2] and ChengXiang Zhai[1]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign
[1]{bhavya2, czhai}@illinois.edu
[2]Department of Computer Science and Engineering, University at Buffalo
[2]jinjun@buffalo.edu

Abstract

We propose a novel application of prompting Pre-trained Language Models (PLMs) to generate analogies and study how to design effective prompts for two task settings: generating a source concept analogous to a given target concept (aka Analogous Concept Generation or ACG), and generating an explanation of the similarity between a given pair of target concept and source concept (aka Analogous Explanation Generation or AEG). We found that it is feasible to prompt InstructGPT to generate meaningful analogies and the best prompts tend to be precise imperative statements especially with a low temperature setting. We also systematically analyzed the sensitivity of the InstructGPT model to prompt design, temperature, and injected spelling errors, and found that the model is particularly sensitive to certain variations (e.g., questions vs. imperative statements). Further, we conducted human evaluation on 1.4k of the generated analogies and found that the quality of generations varies substantially by model size. The largest InstructGPT model can achieve human-level performance at generating meaningful analogies for a given target while there is still room for improvement on the AEG task.[1]

1 Introduction

Large Pre-trained Language Models (PLMs) such as BERT(Devlin et al., 2018) and GPT(Brown et al., 2020) have been applied to many tasks of text generation (e.g., summarization, dialogue system) with promising results (Li et al., 2021). However, no existing work has studied how to apply PLMs to generate different kinds of textual analogies, such as conceptual metaphors (e.g.,"Life is a journey[2]"), and instructional analogies (e.g., "A red blood cell is like a truck in that they both transport essential supplies"(Newby et al., 1995)).

[1]Our code and datasets are available for public use: https://github.com/Bhaavya/InstructGPT-Analogies

[2]https://en.wikipedia.org/wiki/Conceptual_metaphor

Table 1: Selected prompts and InstructGPT-generated analogies for *natural selection*

Prompt (P7):	*What is analogous to natural selection?*
InstructGPT Output:	The analogous process to natural selection is artificial selection. **(9 words)**
Prompt (P2):	*Explain natural selection using a well-known analogy.*
InstructGPT Output:	Imagine that you have a jar of mixed nuts ... If you shake the jar ...the big nuts will fall out first ... analogy is that natural selection is like a sieve that separates the fit from the unfit... **(136 words)**

Generating analogies has a wide range of applications, such as explaining concepts and scientific innovation, and analogies play a crucial role in human cognition. Analogical matching and reasoning enables humans to understand and learn unfamiliar concepts (aka target concepts) by means of familiar ones (aka source concepts) and to make scientific innovations. Unsurprisingly, analogy modeling and generation has been a long-standing goal of AI (Mitchell, 2021). This is a challenging problem because it often requires computing deep semantic similarities that are beyond the surface-level similarity. For example, the Bohr's atom model and the solar system are analogous due to their structural and relational similarities (i.e., atoms orbit around the nucleus like planets around the sun).

Much work has been done to compute such analogical similarities between concepts. However, existing approaches mostly rely on structured representations, thus, they can only where such representations already exist. For example, one of the most popular models is Structural Mapping Engine (SME) (Forbus et al., 2017), which aligns *structured representations* of the target and source concepts using predicate logic. Moreover, they cannot *generate* analogies in natural language.

Proceedings of the 15th International Conference on Natural Language Generation, pages 298 - 312

Inspired by the recent success in applying PLMs to many NLP tasks (e.g., (Li et al., 2021)), we propose and study the application of PLMs to analogy generation. We consider two typical application scenarios of analogy generation: 1) Analogous Concept Generation (ACG): given a target concept (e.,g, bohr's model), generate a source concept analogous to the target concept (e.g., solar system), possibly with an explanation of their similarities; 2) Analogy Explanation Generation (AEG): given a target concept and an analogous source concept, generate an explanation of their similarities.

By noting the similarity of the two tasks defined above to other text generation problems, and being inspired by the recent success of using prompted PLMs for text generation, we propose analogy generation by using a PLM with appropriately designed prompts. We adopt the promising emerging paradigm of prompting language models (Liu et al., 2021) that uses textual prompts with unfilled slots and directly leverages the language models to fill those slots and obtain the desired output. For example, Table 1 shows sample prompts and PLM-generated outputs for ACG from our experiments.

Specifically, we study the following main research questions: RQ1) How effective is a modern PLM such as InstructGPT in generating meaningful analogies? RQ2) How sensitive are the generated analogies to prompt design, the temperature hyperparameter, and spelling errors? RQ3) How does the model size impact the quality of generated analogies?

To study these questions, we design several experiments on analogies generated from the InstructGPT (Ouyang et al., 2022) model. First, we manually validate whether InstructGPT can generate meaningful analogies for ten well-known analogies in the science domain. Next, we design and systematically vary prompt variants (e.g., imperative statements vs. questions) and temperature, and investigate the corresponding variations in the generated text by comparing them to a reference dataset of science analogies. Finally, we study the impact of model size on the quality of generated analogies both by automatically comparing against the reference data and using human evaluation.

Our experimental results show that PLMs (specifically, InstructGPT) offer a promising general approach to generating analogies with properly designed prompts. Furthermore, the InstructGPT model is found to be sensitive to the prompt design, temperature, and spelling errors for this task, particularly to the prompt style (i.e., question vs. imperative statement). Precise imperative statements in low-temperature setting are found to be the best prompts. Finally, the quality of the generated analogies depends heavily on the model size. While the largest model can achieve human-level performance on the ACG task, the smallest model barely generates any meaningful analogies. The AEG task proved to be more challenging based on human evaluation and could be a better test of the analogical reasoning capabilities of PLMs especially for explaining analogies not seen during training.

2 Related Work

2.1 Computational Models of Analogies

There has been a lot of work on computational modeling of analogies (Mitchell, 2021). The SME model (Forbus et al., 2017) is one of the most popular symbolic model that finds the *mapping*, or connections between structured representations of source and target concepts and their attributes. However, such methods cannot generate new analogous source concepts with analogical explanation.

The recent deep learning-based approaches, including using pre-trained language models (Mikolov et al., 2013; Rossiello et al., 2019; Ushio et al., 2021), are able to *generate* analogies to some extent, but are currently limited to simple word-level and proportional analogies, such as (ostrich:bird :: lion:?). In contrast, we aim to generate and explain more complex analogies of concepts, e.g. instructional analogies (Newby et al., 1995).

Another line of work is on finding analogous documents for scientific innovation, such as product descriptions and research papers, based on their semantic similarities (Kittur et al., 2019). In contrast, we operate in a generative task setup.

To the best of our knowledge, none of the existing work has studied the problem of automatically generating complex analogies in natural language. Recently, research on more "generative" analogy-making tasks has been recommended (Mitchell, 2021). Along this direction, we believe that our proposed task is challenging and more practically useful than the existing text-based generative analogical tasks including letter-string (e.g., if "abc" changes "abd"), what does "pqrs" change to?) and word-level analogies.

.

2.2 Prompting Language Models

Recently, prompts have been either manually created or learned to successfully leverage PLMs for several natural language tasks (Liu et al., 2021). Our work is closest to prompting for lexical and proportional analogy generation (Ushio et al., 2021). But, none of the existing work has performed an in-depth study on prompting PLMs for both generating analogous concepts given a single query concept and explaining the analogical similarities between two query concepts.

3 Problem Formulation

Motivated by the practical applications of this task (e.g., explaining concepts), we study analogy generation in the following settings.

1. Analogous Concept Generation (ACG) or **No Source (NO_SRC)**: Here, only the target concept is provided as the input. The goal is to generate an analogous source concept or scenario, along with some explanation to justify the analogy. For example, "Explain Bohr's atomic model using an analogy."

2. Analogy Explanation Generation (AEG) or **With Source (WSRC)**: Here, in addition to the target, the source concept is also a part of input. The goal is to generate an explanation of how the target and source are analogous. For example, "Explain how Bohr's atomic model is analogous to the solar system."

Our problem setup is similar to the use of PLMs for text generation (Li et al., 2021), and is most closely related to single-relation analogy generation (e.g., ostrich : bird :: animal : lion) (Ushio et al., 2021), where the input is a pair of query concept (e.g., ostrich : bird), and the task is to choose an analogical pair from a pre-defined list of candidate pairs. But, our proposed task is still different in nature and much more challenging (e.g., requiring more creativity in some cases). First, both of our inputs and outputs are different. For example, in the proposed ACG setup, our input is a single concept (e.g., "bohr's model"), not a pair of concepts. Our task is to identify another concept (or scenario) that has an equivalence to the query concept based on their deep and non-trivial semantic similarities. No previous work has studied this kind of "single-concept-based" analogy generation with pre-trained language models. Even in the proposed AEG setup where we also use a pair of concepts as input, they are different from the pair used in the previous work. For example, our input could be a pair (e.g., "bohr's model" and "solar system") and the output is an explanation of their analogical relations (e.g., how their structures are similar). Second, we do not have a pre-defined finite list of candidates to choose from, which is a more realistic and interesting setting than previous work from application perspectives, and is also much more challenging for evaluation.

4 Experiment Setup

In this section, we discuss InstructGPT PLM and datasets used in our experiments.

InstructGPT Model: Recently, several PLMs have been developed and trained on massive web data (Devlin et al., 2018; Brown et al., 2020; Raffel et al., 2019). In this study, we probe the aligned GPT-3 models, InstructGPT. These are GPT-3 models that have been optimized to follow instructions better (Ouyang et al., 2022). InstructGPT has four variants depending on the model size (number of parameters), namely Ada (350 M), Babbage (1.3 B), Curie (6.7 B) , and Davinci (175 B)[3]. Unless otherwise mentioned, we use the Davinci model for the experiments as it is expected to have the best performance.

We used the Open AI API [4] to generate all analogies. Main hyperparameters are described in Section 5.2.2 and rest in the Appendix A

Dataset: As the task of analogy generation, as defined in this paper, has not been previously studied, there is no existing data set available to use directly for evaluation. We thus opted to create new data sets for evaluation. Table 2 shows sample data from these datasets.

Standard Science Analogies (STD): As far as we could find, the closest dataset consisting of conceptual analogies is from (Turney, 2008). It consists of ten standard science analogies. However, these only contain the source and target concepts but not any explanation in natural language.

Science analogies from academic Q&A sites SAQA: We searched for quiz questions that asked to create analogies on academic Q&A sites like Chegg.com, Study.com [5] by using search queries

[3] https://blog.eleuther.ai/gpt3-model-sizes/

[4] https://beta.openai.com/docs/api-reference/completions/create

[5] https://chegg.com/, https://study.com/. We manually inspected the data and found no personal identifiers or offensive content. We manually compiled the datasets, no scraping was done.

like ‘create an analogy’, ‘analogy to explain’, and manually downloaded the relevant questions and answers. After manually removing irrelevant data, 75 unique question-answer pairs were obtained. Next, we manually extracted the analogies from answers, i.e., target and source concepts, and the explanation of the analogical similarity.

There are total 109 concepts (about high-school science) with 148 English analogies. The average word length of analogies is 62.25 words.

Table 2: Sample analogies from STD and SAQA.

Dataset	Target	Source	Explanation
STD	atom	solar system	-
SAQA	ligase	sewing machine	... Ligase is similar to a sewing machine, as it binds two elements ... **(25 words)**

5 Experiment Results

In this section, we present our experiment results and examine each of the three research questions introduced earlier.

5.1 Feasibility Analysis

We first examine RQ1 and investigate whether InstructGPT is capable of generating analogies with simple prompts by looking at the results on the smaller STD dataset which contains well-known analogies. Here, we seek standard analogies, so we designed prompts with keywords such as "well-known analogy", "often used to explain", etc. The full list of prompts is in Table 17, Appendix C).

We observed that all the prompts were successful in retrieving natural language analogies to some extent but they differed in several aspects. Table 1 shows sample analogies generated by two of our prompts (P7 and P2, Table 17) for the target concept “natural selection.” In this case, the reference answer in the STD dataset is “artifical selection,” which P7 successfully retrieved, while P2 generated a different but also valid analogy. Such variations indicate both the potential of using different prompts to generate (multiple) different analogies and the model sensitivity to prompt design, which we further investigate in Section 5.2.

To quantify the effectiveness of different prompts, we manually evaluated the source concepts mentioned in the generated analogies (if any). Table 3 shows the number of exact matches of generated source concepts to those in the reference STD dataset, along with the number of “valid” source concepts generated. Valid means a reasonable analogy that is either commonly known (e.g., easily available on the internet [6]) or contains a meaningful justification. All prompts generated valid analogies in most cases, even if they didn’t exactly match the reference source concept further suggesting the promise of InstructGPT for generating meaningful analogies. Note that the low number of exact matches with the reference dataset is expected to some extent because there are several possible “valid” analogies for a given source concept and so there is a small chance that the model would generate exactly the same analogous concept as in the reference.

Table 3: Number of analogies that match the ground truth or are otherwise meaningful, out of the total ten analogies generated for STD target concepts by the seven prompts (P1-P7).

	P1	P2	P3	P4	P5	P6	P7
# Match	3	3	6	4	3	5	3
# Valid	6	9	9	8	7	10	10

5.2 Robustness analyses

As observed in many other applications of prompted PLMs, the performance of a task tends to be sensitive to the prompts used and the temperature parameter (Lu et al., 2021; Zhao et al., 2021). Moreover, many PLMs are known to be vulnerable to the presence of spelling errors (Pruthi et al., 2019; Ma et al., 2020). Thus, it is important to experiment with variations of both the prompts and the temperature parameter (with frequency_penalty, Section 5.2.2), and spelling errors and study how they impact the generated analogy (RQ2).

For these analyses, we need to compare the model performance in a large number of configurations, which makes human evaluation impossible. Thus, we rely on automatic metrics. Automatic evaluation of natural language generation is known to be challenging (e.g., long-form question answering (Krishna et al., 2021)) and automatic metrics generally have low correlation with human judgment (Callison-Burch et al., 2006; Raffel et al., 2019). Evaluation of analogies is even more chal-

[6]Note that commonly known does not necessarily mean available on the internet. We use it only as a proxy here since there is no good way to determine what is common knowledge.

lenging especially because a target concept could have several valid analogies with seemingly different meanings (e.g., "artificial selection" vs. "sieve" from Section 5.1). Thus, before using existing methods, we designed sanity checks and found that those methods behave as we expect (e.g., analogies have a higher score than non-analogies, see Appendix B). We note that our sanity checks are only the necessary and not the sufficient requirements of a good metric for evaluating analogies as they do not evaluate creativity or reasoning. However, we use them as an approximation only for relative comparison between methods on the same task as they are unlikely to favor any single method.

We use three representative measures of automatic evaluation of generated text: BLEURT (Sellam et al., 2020) (B), METEOR (Lavie and Agarwal, 2007) (M), ROUGE-L (R)[7] (Lin, 2004)[8]. BLEURT (B) is used as the primary metric for evaluation since it is a recent machine learning-based metric that has been shown to capture semantic similarities between texts (Sellam et al., 2020).

Similar average BLEURT values would indicate that the prompts are equally good (or bad) on a task, but not necessarily in the same way. On the other hand, Kendall's Tau (Kendall, 1938) indicates how well the ranks of two variables are correlated. This would suggest that those prompts have similar strengths and weaknesses. Thus, we analyze both scores to get a more complete picture of hyperparameter sensitivity.

5.2.1 Analysis of prompts

To study the effectiveness and robustness of different prompts for analogy generation in the unsupervised setting, we manually designed several prompts for all the problem settings. The different prompt variants are all paraphrases that are semantically similar. The main ways they differ are: 1. *Questions vs. Imperative Statements* (e.g., P5 vs. P2, Table 5); 2. *Synonyms* (e.g., P2 vs. P3, Table 5); 3. *Word Ordering* (e.g., P1 vs. P3, Table 4). We only study the zero-shot setting mainly because the choice/number of examples in few-shot could make an impact on the generated analogies and make it harder to interpret our experiment results.

Prompts for the NO_SRC and WSRC settings are in Tables 4,5, respectively. Here, <target>, <src> are target and source concept placeholders.

Our major findings are as follows:

[7]https://pypi.org/project/rouge-score/

[8]https://www.nltk.org/api/nltk.translate.meteor_score.html

Table 4: Prompts for NO_SRC

Id	Prompt
P1	Explain <target> using an analogy.
P2	Create an analogy to explain <target>.
P3	Using an analogy, explain <target>.
P4	What analogy is used to explain <target>?
P5	Use an analogy to explain <target>.

Table 5: Prompts for WSRC

Id	Prompt
P1	Explain <target> using an analogy involving <src>.
P2	Explain how <target> is analogous to <src>.
P3	Explain how <target> is like <src>.
P4	Explain how <target> is similar to <src>.
P5	How is <target> analogous to <src>?
P6	How is <target> like <src>?
P7	How is <target> similar to <src>?

Questions and statements are significantly different: The question prompts are P4, Table 4 and P5-P7, Table 5. From Tables 6 and 7, questions have significantly different and lower scores than statements. This could be an artifact of how the InstructGPT models were trained and should be further investigated.

Table 6: Comparison of performances of different prompts and temperatures in NO_SRC. $*$ and † mean statistically significant compared to the best performing setting at p<0.1 and p<0.05 respectively based on a two-tailed t-test.

	B	R	M
$P1_{tl}$	0.46	0.187	0.154
$P1_{th}$	$0.448^{\dagger}$	$0.181^{\dagger}$	0.167
$P2_{tl}$	0.451	0.193	0.154
$P2_{th}$	0.45^{*}	0.184	0.161
$P3_{tl}$	**0.462**	**0.196**	0.164
$P3_{th}$	0.452	0.188	**0.171**
$P4_{tl}$	$0.427^{\dagger}$	$0.170^{\dagger}$	$0.126^{\dagger}$
$P4_{th}$	$0.431^{\dagger}$	$0.179^{\dagger}$	0.156
$P5_{tl}$	0.451	0.188	0.154
$P5_{th}$	0.449^{*}	0.183^{*}	0.163

Impact of synonyms and word order: Prompt performances vary based on synonyms and word order. For example, some synonymous prompt pairs (e.g, P2-P4, P5-P7 in WSRC) are more correlated than others (e.g., P2-P3, P5-P6 in WSRC). This could be because "analogous to" and "similar to" share a word unlike the other synonym "like". As expected,

prompts with the most different meanings (e.g., P1 in WSRC – involving <src> is not necessarily the same as analogous to <src>) are least correlated with others. However, from Table 7, the average performances of synonymous prompts (e.g., $P2_{tl}$ and $P3_{tl}$, $P2_{tl}$ and $P5_{tl}$) are not significantly different. Overall, this suggests that InstructGPT is more robust to synonyms/word-order than to the prompt style (question/imperative statements) for this task. The overall best-performing prompts (P3 in NO_SRC, P2 in WSRC) contain some form of the word "analogy" rather than its synonyms, confirming that precise and direct prompts are better.

Table 7: Comparison of performances of different prompts and temperatures in WSRC. $*$ and † mean statistically significant at p<0.1 and p<0.05 compared to the best performing setting respectively based on a two-tailed t-test.

	B	R	M
$P1_{tl}$	0.504	0.223	$0.187^{\dagger}$
$P1_{th}$	$0.497^{\dagger}$	$0.212^{\dagger}$	0.199
$P2_{tl}$	**0.515**	0.217	0.203
$P2_{th}$	0.502^{*}	$0.210^{\dagger}$	**0.208**
$P3_{tl}$	0.504	**0.229**	0.191
$P3_{th}$	0.504	0.216	0.203
$P4_{tl}$	0.506	0.214	0.197
$P4_{th}$	$0.497^{\dagger}$	$0.206^{\dagger}$	0.2
$P5_{tl}$	0.499^{*}	0.217	$0.18^{\dagger}$
$P5_{th}$	$0.496^{\dagger}$	$0.211^{\dagger}$	0.191^{*}
$P6_{tl}$	0.500^{*}	0.216	$0.176^{\dagger}$
$P6_{th}$	$0.494^{\dagger}$	$0.212^{\dagger}$	$0.183^{\dagger}$
$P7_{tl}$	$0.497^{\dagger}$	$0.208^{\dagger}$	$0.179^{\dagger}$
$P7_{th}$	$0.492^{\dagger}$	$0.204^{\dagger}$	$0.186^{\dagger}$

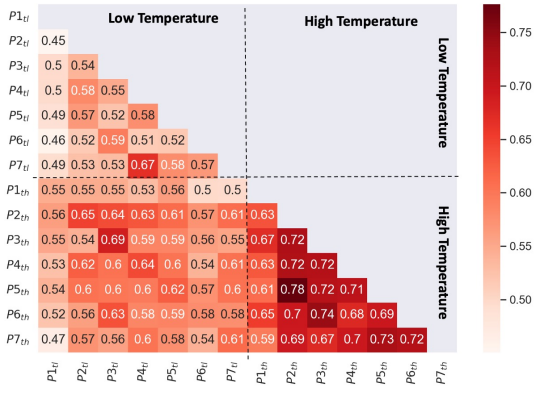

Figure 1: Kendall's Tau correlation between BLEURT scores of various prompts and temperatures in WSRC

5.2.2 Analysis of temperature

Higher temperature increases the randomness in the generated text and is often suggested for creative tasks (Lucy and Bamman, 2021). Since some analogies require creativity, we are especially interested in studying the impact of this hyperparameter.

We explore two settings. **Low Temperature (tl)**: this is a deterministic setting, where temperature = frequency_penalty = presence_penalty = 0. **High Temperature (th)**: Here temperature is set to 0.85. To avoid repetition of words and topics, we set frequency_penalty = 1.24 and presence_penalty = 1.71. These hyperparameters were selected based on initial qualitative exploration. To account for the randomness, we set best_n = 3, i.e., select the best response out of three generated responses, and generate 5 such best responses. In all experiments, we report the average performance of all 5 responses.

From the results in lower-right corner of Figures 1 and 3 (in Appendix D), we make the following two observations:

Firstly, at high temperature, prompts are generally well-correlated with each other suggesting lesser sensitivity to prompt design at high temperatures. This requires further investigation because we expect higher randomness to generate a variety of different analogies, and thus have lower correlations in general.

Secondly, the overall performances of the high-temperature variants are generally lower than their low-temperature counterparts. To investigate when high temperature could help, we further looked into a case in the WSRC setting where the high-temperature version of the best prompt, ($P2_{th}$), performed much better. The results are shown in Table 8. In this case, unlike $P2_{hl}$, $P2_{tl}$ fails on identifying the target and also generates incorrect facts, ("rubber of your lungs"). This shows some evidence of high temperature prompts working better for more complex and creative analogies, which should be investigated further in the future.

5.2.3 Analysis of synthetic spelling errors

Since obtaining natural spelling errors can be difficult, injecting synthetic errors is a common way to test the robustness of models (e.g., (Jayanthi et al., 2020). Thus, following previous work (Sakaguchi et al., 2017; Jayanthi et al., 2020), we injected the following four types of character-level errors to the internal characters of the target concept in the prompt: Delete (delete one randomly chosen character), Permute (switch two randomly chosen ad-

Table 8: Analogies generated in the high and low-temperature setting by the prompt: *Explain how breathing mechanism of frogs is analogous to blowing of the balloon.*

Temp.	Generated Analogy
High	... When you blow into a balloon, the air pressure ... The frog's respiratory system works in a similar way. When you breathe in, you increase the air pressure inside your lungs. This increased air pressure causes the walls ...
Low	When you blow up a balloon, the air inside the balloon is pushing outwards on the rubber of the balloon. This is the same thing that happens when you breathe in: the air inside your lungs is pushing outwards on the rubber of your lungs.

jacent characters in the string), Insert (insert one random alphabet at a random position), and Replace (replace one randomly chosen character in the string with a random alphabet). Target concepts with length less than 3 were kept unchanged.

Average BLEURT scores from three different runs for all prompts in the low-temperature setting in NO_SRC are shown in Table 9. Overall, the performance decreases, indicating the sensitivity of language models to spelling errors. Further, Replace generally leads to the biggest performance drop for all prompts ($\sim 3-7\%$ relative decrease). The model is generally most robust to Insert, similar to the results reported in previous work on word recognition using neural networks (Sakaguchi et al., 2017).

Table 9: Impact of injecting Delete (D), Permute (P), Insert (I) and Replace (R) errors to the target concept in the prompt compared to the original (O) prompt based on BLEURT scores. $*$ and † mean statistically significant at p<0.1 and p<0.05 respectively based on a two-tailed t-test.

	D	P	I	R	O
P1	$0.438^{\dagger}$	$0.437^{\dagger}$	$0.436^{\dagger}$	$0.429^{\dagger}$	0.46
P2	$0.431^{\dagger}$	$0.434^{\dagger}$	0.442	$0.427^{\dagger}$	0.451
P3	$0.444^{\dagger}$	$0.445^{\dagger}$	0.447^{*}	$0.44^{\dagger}$	0.462
P4	0.423	0.424	0.428	0.416	0.427
P5	0.438^{*}	0.437^{*}	0.441	$0.435^{\dagger}$	0.451

5.3 Analysis of model size

Finally, we examine RQ3, i.e., how does the model size impact the quality of the generated analogies. In general, models with more parameters can be expected to perform better. We now study whether the same holds for this task and how much the model size impacts the performance.

Figure 2 shows the BLEURT scores of various models on both the task setups. As expected, the performance increases significantly with model size in both WSRC and NO_SRC, suggesting that larger models are better at generating analogy-like text for the given targets. Further, the biggest improvement is seen as the number of parameters increases from 0.3B to 1.3B in both settings (19.17% and 15.34% relative improvements, respectively).

Similar to what we observed in the case of the 175B Davinci model, the performance in WSRC is higher than that in NO_SRC for other models too. This confirms that all models have some capacity to incorporate the source provided in the prompt.

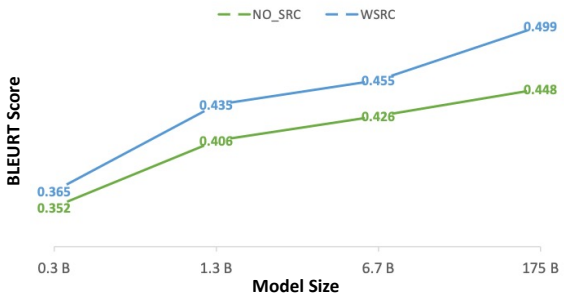

Figure 2: Average performances of various InstructGPT models based on BLEURT scores.

5.4 Human evaluation

To further validate the generated analogies more comprehensively, we also conducted human evaluation as described below.

5.4.1 Annotation Setup

We conducted the study on Amazon Mechanical Turk. Based on manual evaluation of responses to screening tests (Appendix E), we selected 17 workers for the main study.

Further, we created a sample dataset for evaluating analogies generated both in the NO_SRC and WSRC settings. In total, we generated 13k analogies [9] in NO_SRC and 18k analogies [10] in WSRC.

[9] 6 analogies (5 in high temperature and 1 low temperature) *109 target concepts*5 prompts*4 models

[10] 6 analogies (5 in high temperature and 1 low temperature) *109 target concepts*7 prompts*4 models

From this data, we randomly selected 42 concepts for the NO_SRC setup and 21 of them were selected for the wsrc setup (to have comparabale number of analogies in both settings). The analogies for the selected concepts, generated by all the models using all the prompts in the low temperature setting were selected for evaluation since low temperature was better based on automatic evaluation.

In total, 1407 unique analogies (576 from WSRC, 770 from NO_SRC, and 61 human-generated from SAQA) were evaluated by 3 workers each, which is common in previous work on evaluation of automatically generated text (van der Lee et al., 2021). The main study had one question asking workers to evaluate whether the shown candidate analogy was meaningful for the target concept (Yes/No/Can't decide) and provide a text input for explaining their choice (Figure 6, Appendix F). Please refer to Appendix E for more details of the study design.

5.4.2 Quantitative Results

Table 10 shows the percentage of analogies rated as meaningful, based on majority vote, for the various models and the human references from SAQA. There were <2% ties or cases with 'Can't decide' as the majority, which were discarded. The Fleiss' kappa (Fleiss, 1971) inter-annotator agreement was 0.347 in case of WSRC (plus human references for the selected concepts for wsrc concepts), indicating fair agreement and 0.553 in case of NO_SRC (plus human references for the selected concepts for NO_SRC concepts) indicating moderate agreement.

We observe that the percentage of meaningful analogies increases with model size, again confirming that larger models have a higher capacity to generate analogies. Interestingly, in the NO_SRC setting, the largest model has comparable performance to humans. We note that this doesn't necessarily mean that those models are creative or have commonsense reasoning skills as they could have simply memorized those analogies, which a known problem of such models (Bender et al., 2021). It requires further research to test whether the models generate novel analogies unseen during training.

Moreover, upon inspection, we found that the human-generated analogies sometimes had minor issues, such as grammatical errors, which could impact their rating by annotators. So, it is possible that analogies written by experts, such as science instructors proficient in English, might be rated higher. Nevertheless, these results are quite encouraging as the model seems to have comparable performance to general online users who wrote the analogies in our reference dataset.

In the WSRC setting, the performance of InstructGPT is lower than human performance. This could be because there is a lesser likelihood of seeing the exact same analogy, i.e., the one asked to explain in the prompt, during training, compared to seeing *any* analogies for the target concept as required in the NO_SRC setting. So, WSRC might require more "analogical reasoning" from the models, especially for explaining analogies not seen during training. This highlights the importance of human evaluations for such tasks because otherwise, based on automatic evaluation alone, we would conclude that this is an easier setting. This is because metrics like BLEURT cannot assess the soundness of the generated reasoning.

We also compute the NO_SRC performance on the 21 shared concepts (NO_SRC_{21}, Table 10) for a fair comparison between the two settings. It is interesting to note that the performances of smaller models increase while that of larger models go down in the WSRC setting. This could be because the provided source in the prompt helps provide some guidance to the smaller models. For example, even by copying parts of the prompt (i.e., source and target), they could generate meaningful analogies (e.g., <source> is like <target>) in a few cases. Since their performance in the NO_SRC setting is very poor, even minor help or "tricks" would lead to performance improvement. On the other hand, the larger models that already performed very well, likely do not have much to gain from such help and, in fact, perform worse due to the analogical reasoning argument made above.

Overall, this highlights some limitations of the InstructGPT model for analogical reasoning, which requires further research for improvement.

5.4.3 Error Analysis

The annotators were also asked to explain their answer choice (i.e, meaningful analogy or not). By inspection, we identified the following major themes based on the workers' explanations for choosing "not meaningful" across all models/tasks. These themes are not mutually exclusive and multiple themes were often found for one wrong generation.

1. No Analogy: This is one of the most common cases where the model failed to generate any analogy at all. Instead, it mostly generated a simple description/definition of the target concept. In a

Table 10: Percentage of meaningful analogies generated by various InstructGPT models and humans based on human evaluation. Highest value per row is underlined.

	0.3B	1.3B	6.7B	175B	Human
NO_SRC	1.90	15.61	48.29	<u>70.05</u>	66.67
WSRC	8.97	29.05	38.46	53.79	<u>71.88</u>
NO_SRC_{21}	0	12.0	47.0	66.99	<u>71.88</u>

few cases, it also generated a tautology or an example. For example, *"The b-lymphocytes are similar to the white blood cells."*

2. Irrelevant to target: The generated text contained little to none relevant information pertaining to the target. One interesting reason behind this was capitalization for abbreviations. For example, since the targets in the prompt were lowercased (e.g., nadh), smaller models were unable to identify abbreviations, while the larger models succeeded at this. Another reason observed was that of an ambiguous target, e.g., computer "mouse" misidentified as a rodent. In more insidious cases, the text looked correct but presented incorrect facts.

3. Incorrect source or explanation: Here, important details about the source concept were either incorrect or missing, or the provided explanation was insufficient, making the analogy completely wrong or weak at best. For example, *"A molecule of DNA is like a drop of water. It has a specific shape and size, and it can carry the genetic instructions for making a particular organism."*

Some error types found in other natural language generations from GPT-3 (Dou et al., 2021), e.g., incoherence and grammar, were also found in our task. Further research is required to quantify them for analogical generation and attempt to fix them.

6 Limitations

A major limitation of our study is that we only studied analogies on a small reference dataset in one domain (high-school science). Our newly created reference data sets are relatively small due to limited resources found online. But, the sample size of the automatically generated analogies we evaluated was large ($\sim$ 31k automatically evaluated, and $\sim$ 1.4k manually evaluated) thereby mitigating some concerns about bias due to small dataset size. Moreover, as our research questions study an open-ended generation task, having a pre-defined list of reference candidates is not ideal for evaluation. Thus, future research is required to more thoroughly evaluate the generated analogies and investigate the generalizability of the findings to other domains.

Further, the manual evaluation was conducted by a selected group of people in the US and might not reflect the opinions of a more diverse group. Moreover, our kappa scores of 0.3-0.5, although common in previous NLG evaluation work (van der Lee et al., 2021), are not on the higher end. In general, thresholds to determine what counts as high or low kappa scores tend to be open to interpretation (van der Lee et al., 2021). Thus, we've released our annotated and full datasets online, as also suggested in (van der Lee et al., 2021), and invite other researchers to further investigate them.

7 Conclusion

In this study, we proposed and studied the novel task of generating analogies by prompting InstructGPT. Our experiments showed that the InstructGPT is effective on this task when precise prompts are used, thus offering a promising new way to generate analogies, which can break the limitation of the traditional analogy generation methods in requiring a pre-generated structured representation.

By evaluating the performances of the various designed prompts in multiple temperature settings and in the presence of synthetic spelling errors, we found that the InstructGPT model is sensitive to those variations (e.g., question vs. imperative-style prompts). Additionally, based on human evaluation, we found that the quality of the generated analogies substantially depends on the model size. The largest model was found to achieve human-level performance at generating analogies for given target concepts. There is still much room for improvement at the challenging task of explaining the analogical similarity between the given target and source concepts.

Our work opens up many exciting opportunities for future work both for application-oriented and foundational research on PLMs for analogy generation. For example, conducting more robustness analyses based on prompt perturbations (e.g., natural spelling mistakes, grammar, length, etc.). Also, in addition to the unsupervised approaches we explored in this paper, it is interesting to develop supervised approaches for this task including by fine-tuning PLMs on our created datasets.

8 Acknowledgments

This work is supported in part by the IBM-Illinois Center for Cognitive Computing Systems Research (C3SR) as an IBM AI Horizon's Network.

9 Ethical Considerations

The risks associated with using PLMs for analogy generation are similar to those of NLG tasks, such as bias, toxicity, and misinformation (Bender et al., 2021; Weidinger et al., 2021). Accordingly, these should be carefully evaluated before deploying the models for any practical applications, such as education.

Furthermore, there is a steep monetary and environmental cost associated with using the GPT-3 models, especially Davinci. The OpenAI API charges $0.06 /1K tokens. Including early experiments, analogy generation in this study costed a total of about $240. Since we conducted multiple runs with the same prompt account for randomness (e.g., in the high temperature setting), the costs rose sharply.

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. 2021. Scarecrow: A framework for scrutinizing machine text. *arXiv preprint arXiv:2107.01294*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Kenneth D Forbus, Ronald W Ferguson, Andrew Lovett, and Dedre Gentner. 2017. Extending sme to handle large-scale cognitive modeling. *Cognitive Science*, 41(5):1152–1201.

Sai Muralidhar Jayanthi, Danish Pruthi, and Graham Neubig. 2020. Neuspell: A neural spelling correction toolkit. *arXiv preprint arXiv:2010.11085*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Aniket Kittur, Lixiu Yu, Tom Hope, Joel Chan, Hila Lifshitz-Assaf, Karni Gilon, Felicia Ng, Robert E Kraut, and Dafna Shahaf. 2019. Scaling up analogical innovation with crowds and ai. *Proceedings of the National Academy of Sciences*, 116(6):1870–1877.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. *arXiv preprint arXiv:2103.06332*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, pages 228–231.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2105.10311*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Li Lucy and David Bamman. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Charbert: character-aware pre-trained language model. *arXiv preprint arXiv:2011.01513*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *arXiv preprint arXiv:2102.10717*.

Timothy J Newby, Peggy A Ertmer, and Donald A Stepich. 1995. Instructional analogies and the learning of concepts. *Educational Technology Research and Development*, 43(1):5–18.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019. Combating adversarial misspellings with robust word recognition. *arXiv preprint arXiv:1905.11268*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Gaetano Rossiello, Alfio Gliozzo, Robert Farrell, Nicolas R Fauceglia, and Michael Glass. 2019. Learning relational representations by analogy using hierarchical siamese networks. 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 3235–3245.

Keisuke Sakaguchi, Kevin Duh, Matt Post, and Benjamin Van Durme. 2017. Robsut wrod reocginiton via semi-character recurrent neural network. In *Thirty-first AAAI conference on artificial intelligence*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Peter D Turney. 2008. The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, 33:615–655.

Asahi Ushio, Luis Espinosa-Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. Bert is to nlp what alexnet is to cv: Can pre-trained language models identify analogies? *arXiv preprint arXiv:2105.04949*.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*.

A Hyperparameters

Based on initial explorations, where we varied the number of maximum tokens between 0 and 1000 in increments of 100, and then from 935-955 in increments of 1, we noticed that setting a high number of maximum tokens worked better in generating more comprehensive analogies that were not abruptly cut-off and there was little sensitivity to higher values around 950. So, we randomly chose one value in that range (939). The default value of top_p = 1 was used.

B Suitability of existing evaluation metrics

To first investigate the suitability of existing evaluation metrics for generated analogies before we can trust any evaluation results using them, we designed two testers to examine whether the existing metrics behave as expected: 1) **Ordering Tester OT**: This tester is to see if an evaluation metric can order a set of methods that have known orders between them correctly as expected. 2) **Random Perturbation Tester RPT** : This tester checks if an evaluation metric responds to a random perturbation to the ground truth data used for evaluation. A reasonable metric is expected to generate lower performance figures after perturbation.

We use those two testers to study the suitability of three popular and representative measures of automatic evaluation of generated text: BLEURT (Sellam et al., 2020), METEOR (Lavie and Agarwal, 2007), ROUGE-L (Lin, 2004).

BLEURT (B) is a recent machine learning-based metric that has been shown to capture semantic similarities between text. ROUGE-L (R)[11] measures longest matching subsequence of words. We use its F1-score. METEOR (M) [12] matches word stems and synonyms also.

Design of testers: We design an OT and a RPT based on the following baseline methods:

*No Analogy baseline (*NO_ANLGY*):* Here, the prompts instruct the model to generate an explanation or description of the target concept and do not ask for an analogy explicitly. Thus, we expect the generated text to be in a different "style" than analogies and the overall performance to be lower. However, the generation would still contain other relevant keywords describing the target. Thus, it is a good baseline to test if the metrics can distinguish between analogies and other descriptions.

Random baselines: For each of the three setups, we introduced random baselines

[11] https://pypi.org/project/rouge-score/

[12] https://www.nltk.org/api/nltk.translate.meteor_score.html

(NO_ANLGY_RAND, NO_SRC_RAND, and WSRC_RAND, respectively) where a generated string is evaluated against a random analogy (excluding the correct matching analogy) in the reference dataset (i.e., applying a random perturbation to the ground truth). These baselines preserve the "style" of the text but not the content. We expect these methods to perform worse than their non-random counterparts.

Additionally, NO_SRC setting is expected to perform worse than WSRC because in WSRC, the model has more information (i.e., the source concept) and thus has better chances of generating the correct analogical explanation. Thus, the expected order is NO_ANLGY < NO_SRC < WSRC.

Metric testing results: Table 11 shows the overall results of experiments on the SAQA dataset using the Davinci model. Each row shows the highest average scores given by a metric in various setups (performances of each prompt are in Section 5.2 and at the end of this section.).

We can see that all the three metrics order the setups as expected, i.e., random baselines are assigned a lower score than non-random setups, and scores for NO_ANLGY < NO_SRC < WSRC. This suggests that all the three metrics have "passed" our two testers and thus can be reasonably used to evaluate whether the automatically generated analogies are similar to those generated by humans. In other words, they should help assess whether the generated text is relevant to the target concept and discuss properties of the concept that could be explained using analogies (because they passed RPT), and written in an analogical style (because they passed OT).

Moreover, the results also indicates that the InstructGPT model is able to follow the prompts in the three settings to some extent and generate non-analogical descriptions, general analogies, and analogies containing the source concepts, in those settings respectively.

In terms of discernment power, all metrics have small gaps between the scores of random and non-random settings. Similar results were previously reported in (Krishna et al., 2021) for ROUGE scores on long-form question-answering. Out of the three metrics, the BLEURT score has the largest gaps in all the settings, both between the random and non-random baselines and also between settings. It is also shown to capture semantic similarity well (Sellam et al., 2020). Thus, we use it as the main metric in the rest of the experiments.

Table 12: Prompts for NO_ANLGY

Id	Prompt
P1	Explain <target>.
P2	What is <target>?
P3	Explain <target> in plain language to a second grader.

Table 13: Comparison of performances of different prompts and temperatures in NO_ANLGY.

	B	R	M
$P1_{tl}$	0.434	**0.183**	0.149
$P1_{th}$	0.432	0.18	**0.158**
$P2_{tl}$	0.43	0.175	0.129
$P2_{th}$	0.425	0.172	0.136
$P3_{tl}$	**0.445**	0.180	0.132
$P3_{th}$	0.444	0.179	0.144

Table 14: Comparison of performances of different prompts and temperatures in NO_SRC_RAND.

	B	R	M
$P1_{tl}$	**0.375**	**0.132**	0.103
$P1_{th}$	0.367	0.123	0.108
$P2_{tl}$	0.359	0.116	0.092
$P2_{th}$	0.366	0.127	0.105
$P3_{tl}$	0.362	0.124	0.099
$P3_{th}$	0.364	0.126	**0.109**
$P4_{tl}$	0.338	0.115	0.084
$P4_{th}$	0.348	0.121	0.1
$P5_{tl}$	0.358	0.121	0.097
$P5_{th}$	0.348	0.122	0.107

Table 15: Comparison of performances of different prompts and temperatures in WSRC_RAND.

	B	R	M
$P1_{tl}$	0.37	0.120	0.094
$P1_{th}$	0.363	**0.122**	0.107
$P2_{tl}$	**0.385**	0.117	0.096
$P2_{th}$	0.381	0.12	**0.109**
$P3_{tl}$	0.358	0.117	0.095
$P3_{th}$	0.359	0.115	0.1
$P4_{tl}$	0.367	0.113	0.096
$P4_{th}$	0.37	0.115	0.105
$P5_{tl}$	0.36	0.113	0.09
$P5_{th}$	0.356	0.117	0.094
$P6_{tl}$	0.346	0.111	0.086
$P6_{th}$	0.347	0.113	0.091
$P7_{tl}$	0.353	0.114	0.092
$P7_{th}$	0.352	0.109	0.093

Table 11: Testing results using OT and RPT. The higher score between the random baseline and the non-random setup is bolded. Highest score in a row in underlined.

	NO_ANLGY_RAND	NO_ANLGY	NO_SRC_RAND	NO_SRC	WSRC_RAND	WSRC
B	0.349	**0.445**	0.375	**0.462**	0.385	**0.515**
R	0.122	**0.183**	0.132	**0.196**	0.122	**0.229**
M	0.099	**0.158**	0.109	**0.171**	0.109	**0.208**

Table 16: Comparison of performances of different prompts and temperatures in NO_ANLGY_RAND.

	B	R	M
$P1_{tl}$	0.346	0.115	0.087
$P1_{th}$	**0.349**	**0.122**	**0.099**
$P2_{tl}$	0.322	0.116	0.077
$P2_{th}$	0.327	0.113	0.081
$P3_{tl}$	0.334	0.111	0.079
$P3_{th}$	0.336	0.11	0.081

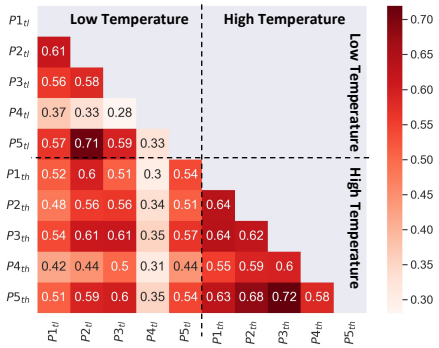

Figure 3: Kendall's Tau correlation between BLEURT scores of various prompts and temperatures in NO_SRC

C Experiments on STD dataset

Table 17: Prompts for STD analogies

Id	Prompt
P1	Explain <target> using an analogy.
P2	Explain <target> using a well-known analogy.
P3	What analogy is often used to explain <target>?
P4	Using a well-known analogy, explain <target>.
P5	Using an analogy, explain <target>.
P6	What is a well-known analogy to explain <target>?
P7	What is analogous to <target>?

Table 18: Most common analogies generated for each target concept in the STD dataset. #Pmt. means number of prompts that generated the shown analogy.

Target	Most common src.	# Pmt.
mind	computer	7
atom	solar system	6
heat transfer	fluid/water flow	4
sounds	wave	4
respiration	combustion	3
light	river	3
planet	rock	2
bacterial mutation	game of telephone	3
natural selection	sieve	2
gas molecules	balls	2

D Experiments on SAQA dataset

Table 19: Comparison of lengths of generated responses by question (Q) vs. statement (S) in the WSRC setting. Question versions of the prompts generate fewer words on average, than their statement counterparts.

Prompt Pair	Avg. Len. (S)	Avg. Len. (Q)
P2-P5	43.93	34.53
P3-P6	32.55	31.4
P4-P7	42.51	32.72

Table 20: Comparison of lengths of generated responses by low and high temperatures in the NO_SRC setting. High temperature generates consistently longer analogies. Same trend is observed in other settings also.

Prompt	Avg. Length (tl)	Avg. Length (th)
P1	39.74	47.62
P2	32.67	40.71
P3	40.06	46.62
P4	32.51	40.13
P5	36.53	38.50

E Mturk study details

For identifying qualified workers on Amazon Mechanical Turk, we designed a pre-screening test (Mturk Qualification) asking them to identify the meaningful analogy for a target concept (Figure 4, Appendix F). Further, we used the following additional qualifications: workers should have com-

pleted at least 5k tasks with >98% approval rate and be located in the US since the task requires proficiency in english (this way of filtering is not perfect but there is currently no good way to identify native english speakers via Mturk). We did not collect any other demographic or geographic information about the workers.

Those who passed these qualifications worked on a small test batch of analogies asking detailed questions about their quality (Figure 5, Appendix F). The questions consisted of both Likert-style or Binary choice questions and text inputs asking them to explain their choices. We manually assessed their responses, especially paying close attention to their reasoning to identify qualified workers for the main study.

For both the main study and the screening, a simple definition of the target from sites like Simple English Wikipedia [13] was provided to workers as reference and they were encouraged to refer to the internet to learn more about the shown concepts. We also provided several sample annotations as part of the instructions to guide workers. Moreover, we were available to answer clarification questions via a shared chatroom.

Annotators were paid at the rate of $50/hr. The rate was decided based on open discussions with them and is above the minimum wage. They were informed that the data generated would be used for research purposes. We consulted with our university ethics board and found that IRB was not required for this study.

F Human evaluation interface

[13] https://simple.wikipedia.org/wiki/Main_Page

NADH stands for "nicotinamide adenine dinucleotide (NAD) + hydrogen (H)." This chemical occurs naturally in the body and plays a role in the chemical process that generates energy.

Which of the following options contains a good analogy to explain NADH? You may refer to the internet to learn more about NADH.

- ○ NADH plays a role in the chemical process that generates energy.
- ○ I always drink the energy drink NADH while playing video games with my friends.
- ○ Nadh is an important molecule in the body. Nadh is kind of like a battery that stores energy until it is needed.
- ○ NADH is like a blood vessel. It helps to carry oxygen and nutrients to your cells.
- ○ The word "nadh" is derived from the an Egyptian word. The analogy used to explain nadh is that between a population and a tribe.

Figure 4: Pre-screening question for identifying qualified workers.

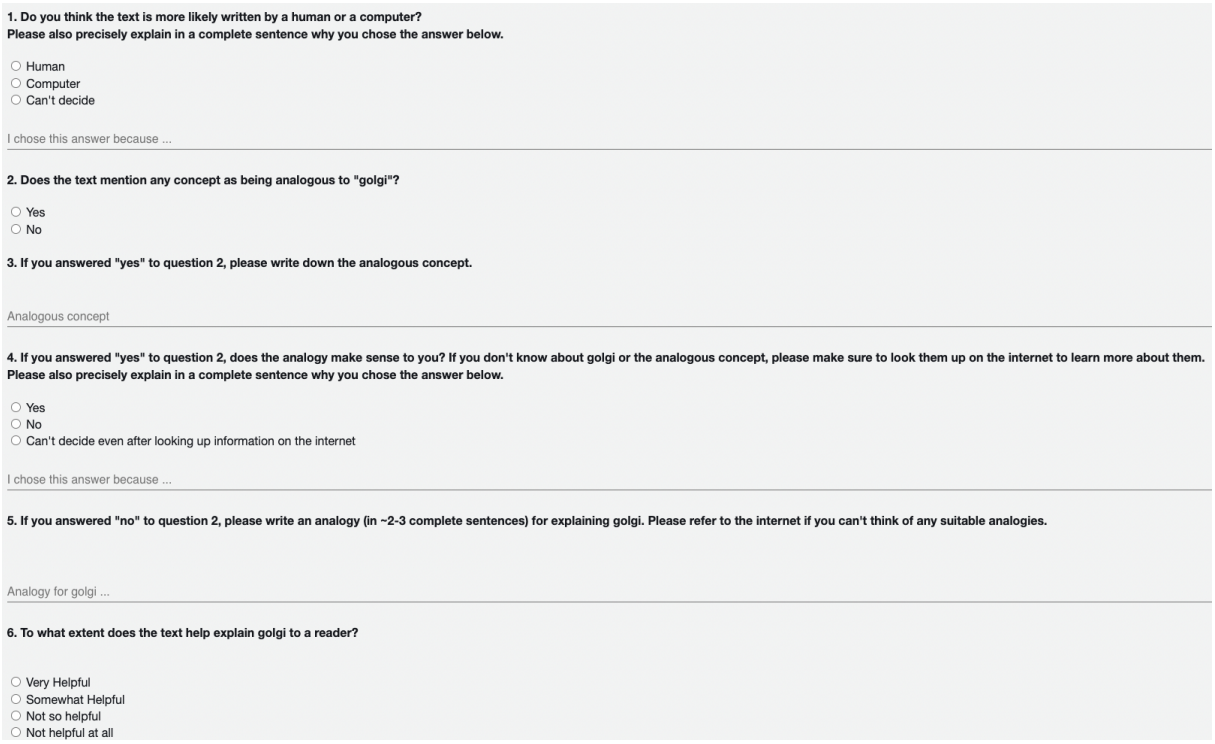

Figure 5: Sample interface for screening qualified workers.

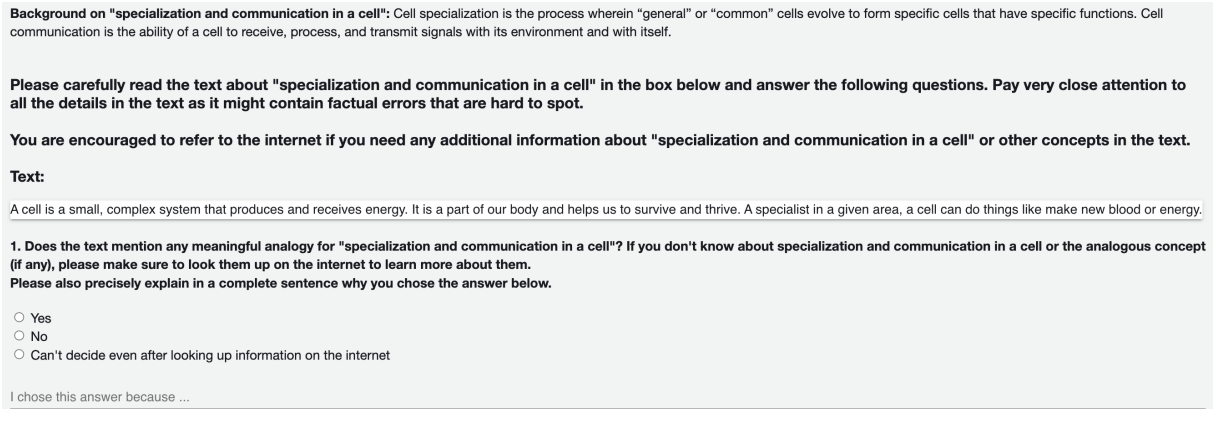

Figure 6: Sample interface for human evaluation of the analogies.