# A Study of Pre-trained Language Models for Analogy Generation

**Anonymous ACL submission**

## Abstract

We propose a novel application of Pre-trained Language Models (PLMs) to generate analogies and study how to design effective prompts to prompt a PLM to generate a source concept analogous to a given target concept as well as to generate an explanation of the similarity between given pair of target concept and source concept. We found that it is feasible to prompt a GPT-3 PLM to generate meaningful analogies and the best prompts tend to be precise imperative statements especially with low temperature setting. We systematically analyzed the sensitivity of the GPT-3 model to prompt design and temperature and found that the model is particularly sensitive to certain variations (e.g., questions vs. imperative statements). We also investigated the suitability of using the existing reference-based metrics designed for evaluating natural language generation (NLG) to evaluate analogy generation and found that the recent BLEURT score is better than the others. We further propose a promising consensus measure based on diverse prompts and settings, which can be potentially used to both automatically evaluate the generated analogies in the absence of reference text (e.g., in novel domains) and rank a set of generated analogies to select analogies of different characteristics. Overall, our study shows that PLMs offer a promising new way to generate analogies in unrestricted domains, breaking the limitation of existing analogy generation methods in requiring structured representation.

## 1 Introduction

Pre-trained Language Models (PLMs) such as BERT and GPT 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 analogies.

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), e.g. understanding the Bohr's model of atoms using the solar system; and to make scientific innovations, e.g. the Wright brothers developed a steerable aircraft based on bicycles. As a result, computing analogies has been a long-standing goal of AI (Mitchell, 2021). This is a challenging problem because it requires computing structural similarities that are beyond the surface-level similarity. For example, at a surface level, the Bohr's atom model and the solar system do not share any attributes, but their relational similarities (i.e., atoms *orbit* around the nucleus just as planets around the sun) makes them analogous.

Much work has been done to compute such analogical similarities. However, existing approaches generally rely on structured representations, thus they can only work on limited domains where such representations already exist. Moreover, they cannot generate analogies in natural language. 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.

Inspired by the recent success in applying PLMs to many NLP tasks (see, 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, generate a source concept analogous to the target concept possibly with an explanation of the similarity of the two concepts. 2) Analogy Explanation Generation (AEG): Given a target concept and an analogous source concept, generate an explanation of their similarity.

We note the similarity of the two tasks defined above and other text generation problems, thus inspired by the recent success of using prompted PLMs for text generation, we hypothesize that analogy generation can also be solved by using a PLM with appropriately designed prompts.

Large pre-trained language models (e.g., GPT-3) have already been successful for other tasks like question answering with minimal further training for downstream tasks. Along this direction, prompting language models (Liu et al., 2021) is a promising emerging paradigm that uses textual prompts with unfilled slots, and directly leverages the language models to fill those slots and obtain the desired output. Our problem setup is similar to the use of PLMs for text generation (Li et al., 2021), but the task of analogy generation is challenging and different from the existing text generation tasks in multiple ways:

Firstly, analogical reasoning and explanation require deep knowledge of the attributes, functions and relations of both source and target concepts. Secondly, analogies are most useful when they explain target concepts using everyday-life scenarios. Both commonsense reasoning and creativity are required to generate plausible yet interesting analogical texts (e.g., electrical resistance is like cats blocking mice). It is unclear whether PLMs are capable of such tasks. Moreover, since this is a new problem, it is unclear how to evaluate the quality of the generated analogies.

In this paper, we address those challenges and study the following main research questions: RQ1) How effective is a modern PLM such as GTP-3 in generating meaningful analogies? RQ2) Are existing NLG evaluation metrics suitable for evaluating generated analogies and can they give meaningful results? RQ3) How sensitive are the generated analogies to prompt design and other hyperparameters? RQ4) Can we automatically assess analogies in the absence of reference dataset?

To study these questions, we design several experiments on analogies generated from the GPT-3 model. Firstly, to assess whether existing NLG metrics (e.g., BLEURT (Lin, 2004)) can be reliably used for evaluating the quality of generated analogies, we design two sanity tests. The tests check whether the metrics generally behave as expected, i.e. give higher scores to higher quality analogies. Using these tests, we select the best metric for automatically evaluating the GPT-3 generated analogies against a reference dataset of analogies of middle-school science concepts created from Chegg.com[1]. We also design and systematically vary prompt variants (e.g., imperative statements -> questions) and investigate the corresponding variations in generated text. Finally, since it is not always feasible to obtain reference datasets (e.g., for new domains or novel analogies), we design a scoring method based on consensus of generated analogies in various settings (e.g., prompt design, temperature), called Consensus Score, to automatically evaluate the analogies without reference text. We also investigate its effectiveness as an evaluation method based on its correlation with BLEURT.

Our study confirms that PLMs offer a promising general approach to generate analogies with properly designed prompts. Furthermore, the GPT-3 model is found to be sensitive to the prompt design and temperature 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. We also confirm the effectiveness of the recent BLEURT(Sellam et al., 2020) score for evaluating analogies. We show that our Consensus Score metric with diversity offers a promising way of reference-free evaluation of generated analogies. We will release all the GPT-3 generated analogies ($\approx$ 6k) for the community's further study of this novel problem.

## 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 between structured representations of source and target concepts. The recent deep learning-based approaches (Mikolov et al., 2013; Rossiello et al., 2019; Ushio et al., 2021), perform analogical reasoning on unstructured text, but are currently limited to simple word-level or proportional analogies, such as (Paris: France:: London: ?). However, none of the existing work has studied the problem of automatically generating complex analogies in natural language.

### 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).

[1] https://chegg.com/

Our work is closest to prompting for Question Answering (Khashabi et al., 2020) and Text Generation (Schick and Schütze, 2020; Li and Liang, 2021). Analogy generation is different from question answering since it also requires more creativity to construct new analogies that do not directly exist on the web. It is a challenging case of text generation that requires reasoning of relational similarities between two concepts or situations. None of the existing work has studied prompting PLMs for analogy generation.

## 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)**: In this setup, only the target concept is provided as the input. The goal is to generate a text that contains an analogous source concept or situation (e.g., solar system), along with some explanation to justify the analogy. Practically, this setting could be useful in finding unknown analogies and creating novel analogies.

2. Analogy Explanation Generation (AEG) or **With Source (WSRC)**: In this setup, in addition to the target, the source concept is also a part of input. The goal is to generate a text that should be an explanation of how the target and source are analogous. If successful, this setup could be useful in assisting users in creating better analogical explanations and seeing connections between two seemingly disparate concepts or situations.

## 4 Experiment Setup

In this section, we discuss the data sets and the GPT-3 PLM used in our experiments.

**Dataset:** As the task of analogy generation 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.

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

*Analogies from Chegg.com (CHG):* We searched for quiz questions that asked to create analogies on Chegg.com [2] by using search queries like 'create an analogy', 'analogy to explain', and downloaded the relevant questions and answers. After manually removing irrelevant data, 75 unique question-answer pairs were obtained. Next, we manually extracted the target and source concepts, and the full analogies from the answers (containing explanation of the analogical similarity). The final dataset has 148 analogies in English for 109 target concepts. The analogies are mostly about middle school science and few basic computer science concepts. The average length of analogies is 55.63 words.

**GPT-3 Davinci Instruct 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 popular GPT-3 model. With 175 billion parameters, the Davinci model of GPT-3 is the most powerful GPT-3 model. The Davinci "Instruct" model is further optimized to follow instructions better [3]. We leave the exploration of other PLMs to future work.

We used the Open AI API [4] to generate all analogies. Based on initial qualitative explorations, we noticed that setting a high number of maximum tokens worked better in generating more comprehensive analogies. Thus, we set it to 939. The default value of top_p = 1 was used. Other hyperparameters are described in Section 5.3.2.

## 5 Experiment Results

### 5.1 Feasibility Analysis

We first investigate whether PLMs are 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 11, Appendix A.

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 11) for the target con-

[2] https://chegg.com/. Terms of Use: https://www.chegg.com/termsofuse. We do not intend to distribute the dataset without their permission. We manually inspected the data and found no personal identifiers or offensive content.

[3] https://beta.openai.com/docs/engines/instruct-series-beta

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

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

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

cept "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 5.3.

To quantify the effectiveness of different prompts, we manually evaluated the source concepts mentioned in the generated analogies (if any). Table 2 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., available on the internet) or contains a meaningful justification. All prompts generated valid analogies in many cases, even if they didn't exactly match the reference source concept. This suggests that a concept could have several valid analogies and it might be infeasible to pre-specify all the valid analogies, making it challenging to accurately evaluate such generated analogies without relying on manual assessment of each result. We will explore automatic evaluation in Sections 5.2, 5.4.

Table 2: Comparison of prompts for STD analogies

| | 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 Suitability of existing evaluation metrics

Automatic evaluation of natural language generation is known to be challenging (e.g., in case of long-form question answering (Krishna et al., 2021)) as automatic metrics do not accurately reflect semantic similarity (Callison-Burch et al., 2006; Raffel et al., 2019). Evaluation of analogies is even more challenging especially when no source is provided (NO_SRC), because a target concept could have several valid analogies with seemingly different meanings (e.g., "artificial selection" vs. "sieve" from Section 5.1).

Given these challenges, we need to first investigate the suitability of existing evaluation metrics for generated analogies before we can trust any evaluation results using them. To this end, we design 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), ROGUE-L F1 (Lin, 2004).

*BLEURT (B)* is a recent machine learning-based metric that has been shown to capture semantic similarities between text. *ROUGE-L (R)*[5] measures longest matching subsequence of words. We use its F1-score. *METEOR (M)*[6] 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 (NO_ANLGY_RAND, NO_SRC_RAND,

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

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

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 3 shows the overall results of experiments on the CHG dataset. Each row shows the highest average scores given by a metric in various setups (performances of each prompt are in Section 5.3 and Appendix A).

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 generated analogies.

Moreover, this also indicates that the GPT-3 model is able to "understand" 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. However, 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.

## 5.3 Comparative analysis of prompts and temperature

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). Thus it is important to experiment with variations of both the prompts and the temperature parameter (with frequency_penalty, Section 5.3.2 ) and study how they impact the generated analogy.

Specifically, we mainly compare their average performances of the different hyperparameter combinations (tables 6, 7), and their correlations based on BLEURT scores (figures 1, 2).

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. In our case, a high Kendall's Tau would indicate that the two prompts generally perform better (or worse), with possibly different magnitudes, on the same individual test samples. 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.

### 5.3.1 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, such that they 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, i.e. we do not provide examples of analogies in the prompt mainly because the choice/number of examples used can significantly impact the results. We leave such explorations for future work.

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

**Results:** We discuss our major findings below.

*Questions are worse than statements*: The question prompts are as follows: P4 in NO_SRC and P5-P7 in WSRC. From Figure 2 top-left and bottom-right, as expected, those questions and statements that share the most words (e.g., P3 and P6) are more strongly correlated. Even so, from Tables 5 and 4, questions perform significantly worse, consistently. These effects could be an artifact of how the GPT-3 Instruct models are trained and should be investigated in the future.

*Impact of synonyms and word order:* Prompt performances vary based on synonyms and word order. For example, some synonymous prompt pairs (e.g,

Table 3: Testing results using OT and RPT. Bold font means the higher score between the random baseline and the non-random setup. 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 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>? |

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 GPT-3 is more robust to synonyms/word-order than to the prompt style (question/imperative statements) for this task. The overall winning prompts (P3 in NO_SRC, P2 in WSRC) contain some form of of the word "analogy" confirming that precise prompts are better.

### 5.3.2 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.

Table 6: Comparison of performances of different prompts and temperatures in NO_SRC. * and ** mean statistically significant 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** | 0.181** | 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** | 0.170** | 0.126** |
| P4$_{th}$ | 0.431** | 0.179** | 0.156 |
| P5$_{tl}$ | 0.451 | 0.188 | 0.154 |
| P5$_{th}$ | 0.449* | 0.183* | 0.163 |

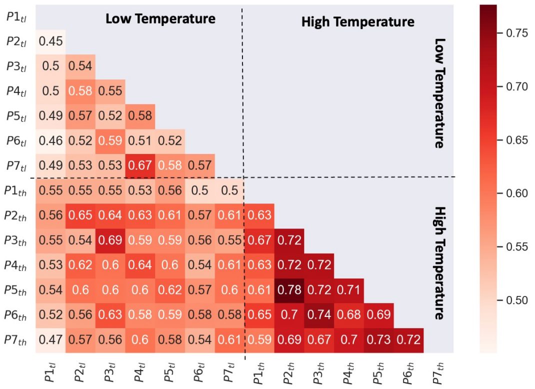

Figure 1: Kendall's Tau correlation between BLEURT scores of various prompts and temperatures in WSRC

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

Table 7: Comparison of performances of different prompts and temperatures in WSRC. * and ** mean statistically significant at p<0.1 and p<0.05 respectively based on a two-tailed t-test.

| | B | R | M |
|---|---|---|---|
| P1$_{tl}$ | 0.504 | 0.223 | 0.187** |
| P1$_{th}$ | 0.497** | 0.212** | 0.199 |
| P2$_{tl}$ | **0.515** | 0.217 | 0.203 |
| P2$_{th}$ | 0.502* | 0.210** | **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** | 0.206** | 0.2 |
| P5$_{tl}$ | 0.499* | 0.217 | 0.18** |
| P5$_{th}$ | 0.496** | 0.211** | 0.191* |
| P6$_{tl}$ | 0.500* | 0.216 | 0.176** |
| P6$_{th}$ | 0.494** | 0.212** | 0.183** |
| P7$_{tl}$ | 0.497** | 0.208** | 0.179** |
| P7$_{th}$ | 0.492** | 0.204** | 0.186** |

report the average performance of all 5 responses.

**Results:** Firstly, at high temperature, prompts are generally well-correlated with each other(lower right, figures 1, 2) 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 would 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.

### 5.4 Consensus Scoring

In the previous sections, we have analyzed generated analogies in the context of reference text. However, one remaining challenge is how to evaluate generated analogies in the absence of a reference dataset, which we anticipate to be a common scenario (e.g., for generation of novel analogies). Although some reference free-metrics have been developed, they are mostly task-specific, e.g. for dialogue generation (Mehri and Eskenazi, 2020) and thus, not usable. We design a reference-free way to rank generated text based on consensus that we call *Consensus Score* or $con\_score$.

Intuitively, if an analogy generated in one setting (e.g., combination of temperature and prompt) matches well with those generated in several other settings, it suggests that the generation is likely correct. Thus, given analogies generated in $K$ different settings, we estimate the correctness of a generated analogy $a_i$ in a given setting as follows:

$$con\_score(a_i) = \sum_{k \in K, k \neq i} \frac{1}{n} sim(a_i, a_k)$$

In other words, con_score is the average of pairwise similarities of (the embedding of) $a_i$, with (embeddings of) analogies generated in other settings. Embeddings were computed using Sentence-BERT(Reimers and Gurevych, 2019).

As an extension of this idea, instead of computing similarities with all the settings, we investigate selecting a subset. This is desirable because of two main reasons: Firstly, as in other consensus/ensembling approaches, some subsets of methods/settings could work better than all. Secondly, it is cheaper to generate text from less number of settings. The setting selections can be made in several ways. Firstly, we expect the settings with better overall performances to generate better analogies. Secondly, with a fixed number of settings, it is expected that consensus with the more diverse ones would increase the reliability. Thus, we do some exploration with these two criteria.

To investigate the goodness of the con_score, we compute their correlation with their BLEURT scores based on Kendall's Tau ($\tau$).

**Results:** In WSRC, a $\tau$ of 0.217 was obtained (p<0.0001), indicating moderate positive correlation (Botsch, 2011). In NO_SRC, $\tau$ of 0.07 (p<0.0001) was obtained, suggesting very weak positive correlation. This is expected because in NO_SRC, several settings can agree on a valid analogy that is different from the reference. But, in WSRC, there is less likelihood of multiple valid explanations of how two concepts are analogous.

We further explore if selecting only two prompts (with high or low temperatures) can improve the correlation in the WSRC setting. For this, we fix the best performing prompt (P2 with low temperature) and test its combinations with other prompts in both the temperatures. Table 9 shows the results.

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. |

Table 9: Kendall's Tau between BLEURT scores and con_score computed with P2 and another prompt in high and low temperature settings. p<0.0001 for all values. Highest values per row are bolded.

| temp. | P1 | P2 | P3 | P4 | P5 | P6 | P7 |
|---|---|---|---|---|---|---|---|
| low | **0.217** | - | 0.159 | 0.146 | 0.13 | 0.172 | 0.134 |
| high | **0.226** | 0.179 | 0.177 | 0.192 | 0.158 | 0.173 | 0.218 |

We can see that the overall correlation can indeed be improved upto 0.226 by selecting 2 prompts instead of all 14 settings. Secondly, combining with more diverse prompts, namely either high temperature prompts (e.g., $P1\_th$, $P7\_th$) or prompts that are less correlated with $P2_{tl}$ (e.g., $P1_{tl}$ – refer figure 1) achieves better results.

Overall, this suggests a reasonable way to assess analogies in the absence of reference data, at least in the WSRC setting. Moreover, our analysis gives an insight into selecting diverse prompts (e.g., for ensembling (Jiang et al., 2020)) based on temperatures and correlations of prompt performances.

**Characteristics of analogies based on consensus score** The previous section suggests that consensus can be used to identify better analogies. We now qualitatively look at some analogies with high and low consensus scores (based on all the 14 prompt+temperature combinations in the NO_SRC setting). We observed that the analogies with higher consensus scores were generally well-known analogies on the web (e.g., golgi apparatus ⟷ post office). On the other hand, while some analogies with low consensus scores were invalid or non-analogies, there were two interesting cases with low scores. Firstly, creative analogies (e.g., density wave ⟷ people walking through crowd) were found to be more diverse, as expected. Secondly, in case of ambiguous concepts (e.g., ram could be either computer RAM or push), analogies differed based on the word senses. Thus, both low and high consensus scores help identify analogies and target concepts with interesting characteristics. We show concepts having analogies with top highest and lowest average consensus in Table 10.

Table 10: Concepts with highest and lowest average consensus among their own respective analogies.

| Highest consensus | Lowest Consensus |
|---|---|
| tumor suppressor genes | universe |
| ligase | ram |
| lysosomes | transcription |
| motherboard | resonance hybrid |

## 6 Conclusion

In this study, we proposed and studied the novel task of generating analogies by prompting PLMs. Our experiments showed that the PLMs are 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, we found that the GPT-3 model is sensitive to those variations, particularly to the prompt style (question vs. imperative statements) and temperature. Additionally, we studied the automatic evaluation of generated analogies and confirmed that the recent BLEURT metric is more effective compared to others. We also designed a promising, diversity-based consensus score for evaluation.

A major limitation of our study is that we only used one domain (middle-school science) and one GPT-3 model. Its generalizability to other domains and with other models should be examined further in the future.

## 7 Ethical Considerations

The risks associated with using PLMs for analogy generation are similar to those of other NLG tasks, such as bias and misinformation. 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 model, especially Davinci. The OpenAI API charges $0.06 /1K tokens. Including early experiments, this study costed a total of about $240. Since we generated multiple runs in the high temperature settings with best_n=3, the cost rose sharply. Future work should investigate the capabilities of other smaller and more cost-effective models for this task.

## References

RE Botsch. 2011. Chapter 12: Significance and measures of association.

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*.

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.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

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

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

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*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

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.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*.

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*.

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*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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.

Timo Schick and Hinrich Schütze. 2020. Few-shot text generation with pattern-exploiting training. *arXiv preprint arXiv:2012.11926*.

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*.

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 Appendix

Table 11: 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 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 |

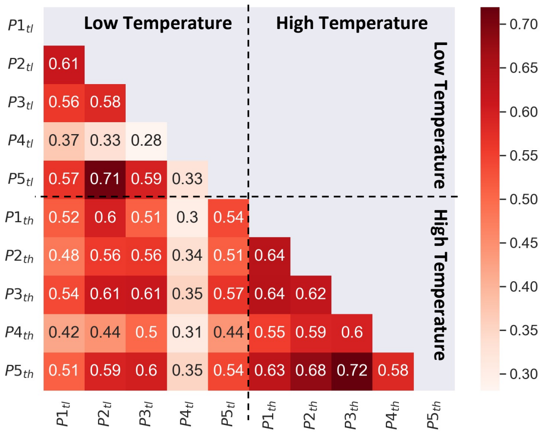

Figure 2: Kendall's Tau correlation between BLEURT scores of various prompts and temperature values in the NO_SRC setting

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 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 |

Table 17: 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 18: 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 |

Table 19: 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 |