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

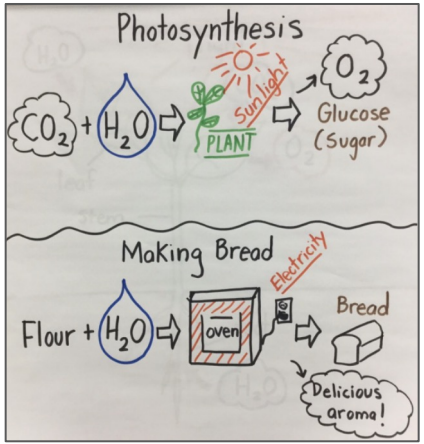
Bhavya[1], Jinjun Xiong[2], ChengXiang Zhai[1]

[1]University of Illinois at Urbana-Champaign, [2] University at Buffalo

IBM | ILLINOIS

Motivation

- Analogies play a major role in human cognition



Solar system	Structure of the atom
The sun	A nucleus
Planets orbit	Electrons orbit
Planets	Electrons
The spherical shape of the sun and planets	The spherical shape of nucleus and electrons
Fixed distance from the sun to the planets	Fixed distance from the nucleus to the electrons
Helium and hydrogen as component of the sun	Proton and neutron as components of the nucleus

- Automatically generating such analogies is a novel and challenging task as it often requires identifying relational similarities between concepts

Research Questions

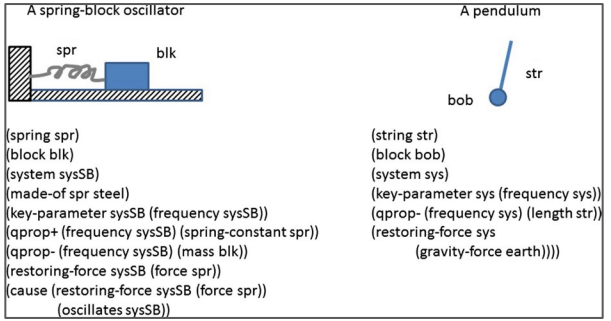
- Inspired by the recent success of prompting large pre-trained language models on various NLP tasks
  - Use textual prompts with unfilled slots and leverage the language models to fill those slots to obtain the output

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)**

- **RQ1:** How effective is a modern PLM such as InstructGPT (GPT-3 aligned to follow human instructions) 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?

Related Work

- *Extending SME to Handle Large-Scale Cognitive Modeling (Forbus et al., 2017)*
  - Structure Mapping Engine: Symbolic model that finds the mapping or connections between structured representations of source and target concepts and their attributes
  - Requires structured representation of concepts unlike our task setting



- *BERT is to NLP what Alex Net is to CV: Can Pre-Trained Language Models Identify Analogies? (Ushio et al., 2021)*
  - Prompts pre-trained language models to generate proportional analogies (e.g., ostrich:bird::lion:animal) unlike the analogies we aim to generate

Problem Formulation

- Analogous Concept Generation (ACG) or **No Source (NO_SRC)**
  - Given a target concept, 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.”

- Analogous Explanation Generation (AEG) or **With Source (WSRC)**
  - Given a target concept, and an analogous source concept, generate an explanation of how the two concepts are analogous.
  - For example, “Explain how Bohr’s atomic model is analogous to the solar system.”

Experiment Setup

- InstructGPT Model
  - GPT-3 model further trained to follow human instructions (Ouyang et al., 2022)
  - Three model sizes: Ada (350 M), Babbage (1.3B), Curie (6.7 B) , and Davinci (175 B)

- Datasets
  - **STD:** Ten standard science analogies previously used in another task (Turney et al, 2008)
    - Does not contain natural language explanations
  - **SAQA:** Science analogies from academic Q&A sites
    - Manually downloaded science analogies from sites like chegg.com
    - 148 English analogies about 109 high-school science concepts

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)**

Feasibility Analysis

RQ1: Investigate whether InstructGPT is capable of generating analogies by prompting

- Designed simple zero-shot prompts
  - For example, “Explain <target> using a well-known analogy.”

- Manually evaluated the following after identifying the source concepts in the generated analogies:
  - Exact matches of generated source concepts to those in the reference STD
  - “Valid” or meaningful sources

- All prompts generated valid analogies in most cases, suggesting the promise of InstructGPT for generating meaningful analogies

- A low number of exact matches (out of the high number of valid sources) show the promise of generating new (and possibly creative) analogies

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

Comparative Analysis of Prompts & Temperature

RQ2: Study how variations in prompts and temperature impact the generated analogies

- Designed paraphrastic prompts that systematically vary (e.g., Questions vs. Imperative Statements)
- Studied two temperature settings: Low (temp = 0), High (temp = 0.85)
- Automatically evaluated generated analogies against references in SAQA using measures like BLEURT (Sellam et al., 2020)
- Found that questions have significantly different and lower scores than statements; lesser sensitivity to synonyms, word order
- Lower temperature achieved higher scores on average, possibly due to more irrelevant words generated with higher temperature

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>?

Model Size Comparison

RQ3: How does the model size impact the quality of the generated analogies?

- Performance increases significantly with model size in both WSRC and NO_SRC settings
  - Larger models are better at generating analogy-like text for the given targets

- Performance in WSRC is higher than in NO_SRC
  - All models have some capacity to incorporate the source provided in the prompt

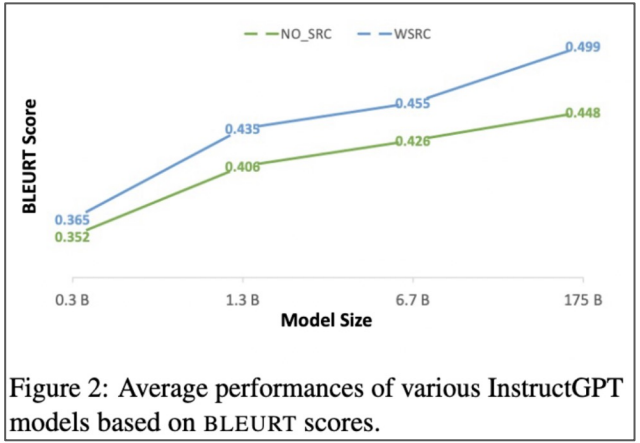


Figure 2: Average performances of various InstructGPT models based on BLEURT scores.

Human Evaluation

- For a more comprehensive analysis, conducted an Mturk study to annotate whether an analogy is meaningful or not. Annotated ~1.4k total analogies by 3 annotators each

- In the NO_SRC setting, the largest model has comparable performance to human-written analogies in the reference dataset

- In the WSRC setting, the performance of InstructGPT is much lower than human performance
    - WSRC might require more analogical reasoning from the models, especially for explaining analogies not seen during training

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>

Error Analysis

- No Analogy
  - Generated text is mostly a simple description of the target concept, an example, or a tautology
  - For example, “The b-lymphocytes are similar to the white blood cells.”

- Irrelevant to target
  - Little to none relevant information pertaining to the target
  - For example, computer “mouse” misidentified as a rodent

- Incorrect source or explanation
  - Incorrect or missing details about the source concept, or insufficient explanation 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.”

Conclusion

- Proposed and studied the novel task of generating analogies by prompting InstructGPT
- Showed that InstructGPT is effective on this task when precise prompts are used, thus offering a promising new way to generate analogies
- InstructGPT model is sensitive to variations in prompts (e.g., question vs. imperative-style), temperature, and spelling errors
- Quality of the generated analogies substantially increases with the model size, reaching human-level performance at the task of generating analogous source concepts
- Still much room for improvement, especially at the challenging task of explaining the analogical similarity between the given target and source concepts
- Future work includes developing better models for this task, including supervised models fine-tuned on our datasets; checking generalizability of our findings to other domains and larger datasets

Thank You!

Acknowledgment:

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

Github:

https://github.com/Bhaavya/InstructGPT-Analogies

Contact:

Bhavya, ChengXiang Zhai, Jinjun Xiong
{bhavya2,czhai}@illinois.edu}, jinjun@buffalo.edu

References

- Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending SME to handle large-scale cognitive modeling. *Cognitive Science*, *41*(5), 1152-1201.
- Ushio, A., Espinosa Anke, L., Schockaert, S., & Camacho-Collados, J. (2021). BERT is to NLP what AlexNet is to CV: Can Pre-Trained Language Models Identify Analogies?. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 3609–3624). Association for Computational Linguistics.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
- Turney, P. D. (2008). The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, *33*, 615-655.
- Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7881–7892). Association for Computational Linguistics.