Boosting Scientific Concepts Understanding: Can Analogy from Teacher Models Empower Student Models?

Siyu Yuan[♡*], **Cheng Jiayang**[♠*], **Lin Qiu**[◇] **Deqing Yang**[♡†]
[♡]School of Data Science, Fudan University
[♠]The Hong Kong University of Science and Technology
[◇]Shanghai Jiao Tong University
syyuan21@m.fudan.edu.cn, jchengaj@cse.ust.hk
lqiu@apex.sjtu.edu.cn, yangdeqing@fudan.edu.cn

Abstract

Analogical reasoning plays a critical role in human cognition, enabling us to understand new concepts by associating them with familiar ones. Previous research in the AI community has mainly focused on identifying and generating analogies and then examining their quality under human evaluation, which overlooks the practical application of these analogies in real-world settings. Inspired by the human education process, in this paper, we propose to investigate how analogies created by teacher language models (LMs) can assist student LMs in understanding scientific concepts, thereby aligning more closely with practical scenarios. Our results suggest that free-form analogies can indeed aid LMs in understanding concepts. Additionally, analogies generated by student LMs can improve their own performance on scientific question answering, demonstrating their capability to use analogies for self-learning new knowledge. Resources are available at https://github.com/siyuyuan/SCUA.

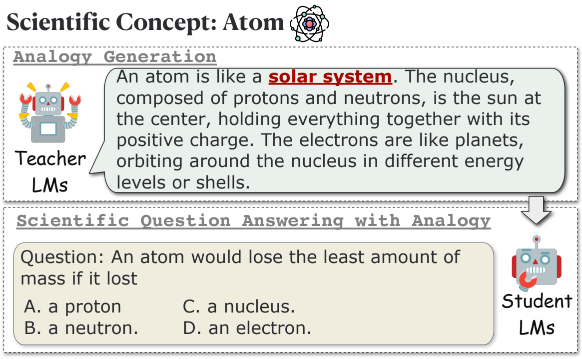

Figure 1: An example of the SCUA task. Given a scientific concept (*i.e.*, Atom), we ask teacher LMs to generate an analogy to explain the concept and then let student LMs answer the related scientific questions around this concept, both with and without the aid of the generated analogy.

1 Introduction

Analogy plays a crucial role in human cognition, facilitating the understanding of complex and unfamiliar concepts by relating them to familiar ones (Bunge, 1981; Glynn et al., 1989; Hofstadter, 2001; Bartha, 2013). For example, Figure 1 illustrates how using the solar system as an analogy can enhance understanding of the complex structure of atoms. Given its significant value across various fields, including creativity (Kang et al., 2022) and education (Richland and Simms, 2015; Thagard, 1992), the topic of analogy has been drawing significant research attention in the AI community.

Traditional research on analogy primarily focuses on evaluating (Allen and Hospedales, 2019; Schluter, 2018; Czinczoll et al., 2022; Chen et al., 2022) and enhancing (Ushio et al., 2021; Yuan et al., 2024) the analogical reasoning capabilities of language models (LMs) in word analogies (*e.g.*, "king is to man as queen is to woman"). Recent advancements in large language models (LLMs) (OpenAI, 2022, 2023) have shifted this focus from simple word analogies to exploring analogies between more complex situations such as systems (Yuan et al., 2023), processes (?Sultan and Shahaf, 2022; Ding et al., 2023; Sultan et al., 2024), paragraphs (Webb et al., 2022; Wijesiriwardene et al., 2023), and stories (Jiayang et al., 2023). However, these studies mainly examine whether LLMs can generate appropriate analogies under human evaluation without thoroughly assessing the practical functionality of the generated analogies in real-world scenarios.

In this paper, drawing on principles of human education, we propose the SCUA, *i.e.*, **S**cientific **C**oncept **U**nderstanding with **A**nalogy task, which aims to investigate whether analogies generated by teacher LMs can assist student LMs in understanding scientific concepts. Specifically, as shown in Figure 1, given a scientific concept, we initially prompt teacher LMs, (*e.g.*, GPT-4 (OpenAI,

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6026–6036
November 12-16, 2024

2023) and Claude (Anthropic, 2024)), to generate an analogy that explains the concept. Then, we collect related scientific questions around this concept from the database and let student LMs (*e.g.*, GPT-3.5 (OpenAI, 2022) and Vicuna (Chiang et al., 2023)) attempt to answer these questions, with and without the use of the generated analogy.

Under this setting, we conduct extensive experiments to evaluate strong and weak LMs with different analogy types. The main findings are as follows:

- Analogies indeed help LMs understand scientific concepts, improving their ability to answer scientific questions.
- Although word analogies generated by teacher LMs reveal the highest quality, the more sophisticated structured and free-text analogies bring higher improvements to the student LMs, suggesting that future work can focus on enhancing the quality of structured and free-text analogies.
- Analogy generated by student LMs can boost their own performance on scientific quizzes, illustrating their ability to leverage analogies for self-learning new knowledge.

2 Related Work

Analogical Reasoning Analogical reasoning has long interested the AI community (Davies, 1985; Gentner and Forbus, 2011; Mitchell, 2021). Traditional research has focused on word analogies, examining linear relationships between words (Gladkova et al., 2016; Schluter, 2018; Fournier et al., 2020; Ushio et al., 2021). With the development of LLMs (OpenAI, 2022, 2023; Team and Google, 2023), there has been a shift toward exploring analogies between situations, establishing mappings between concepts across two domains based on shared relational structures (Sultan and Shahaf, 2022; Ding et al., 2023; Jiayang et al., 2023; Sultan et al., 2024). Compared to these studies, our research is the first to explore how analogies generated by teacher LMs can aid student LMs in understanding scientific concepts, which is more aligned with real-world scenarios.

Explanation Generation With the rising capabilities of LLMs, prior research has adopted methods, *e.g.*, Chain of Thought (CoT) (Wei et al., 2022; Zhang et al., 2023), to generate a reasoning process before answering. Due to the relatively limited capabilities of smaller LMs, some studies employ knowledge distillation, which involves generating reasoning samples using larger LMs to instruct smaller models (Wang et al., 2023a; Hsieh et al., 2023; Wang et al., 2023b; Lin et al., 2023; He et al., 2024). Compared to these studies, our work is the first to explore explanations with analogical reasoning in understanding scientific concepts.

3 SCUA Task

3.1 Task Formulation

As illustrated in Figure 1, given a scientific concept C, we initially ask teacher LMs to generate analogies C_A to explain this concept. Then we will give a scientific question Q and m candidate answer $A = \{A_i\}_{i=1}^{m}$, which is related to the scientific concept C. The ultimate goal of student LMs is to make the correct choice $\mathcal{Y}$ for $\mathcal{X} = (Q, A, C_A)$.

3.2 Analogy Generation from Teacher LMs

Scientific Concept Extraction Current scientific question answering (QA) datasets rarely explicitly contain the related concepts for reference. Thus, given a scientific question, we adopt GPT-4 to extract one scientific concept related to this question. Next, we employ three annotators to evaluate and improve the quality of the concepts. Then, teacher LMs generate analogies for these concepts.[1]

Analogy Type In this paper, we select three types of analogies for generation:

- **Word Analogy**: We adopt the format from Chen et al. (2022) in generating word analogies ("A is to B as C is to D").
- **Structured Analogy**: Structured analogies originate from the Structure Mapping Theory (Gentner and Markman, 1997), which posits that analogies are formed by identifying common relational structures between two concepts. Thus, in addition to using one concept to explain another, we also ask the LMs to incorporate related concepts to demonstrate the analogy further.
- **Free-form Analogy**: These analogies utilize unstructured natural language to explain one concept through another. The popularity of this type is increasing with advancements in LLMs (Wijesiriwardene et al., 2023; Ye et al., 2024).

[1]The extraction prompt for GPT-4 is shown in Appendix C.3, and annotation details are shown in Appendix A.

Scientific Concept: Thermal Equilibrium	
Free-Form Analogy	Imagine a group of children, each holding a different number of balloons and standing in a room. Over time, they start trading balloons to balance out their amounts until each child is holding roughly the same number. Thermal equilibrium works similarly with temperature. If you place a hot object and a cold object close together, heat (like the balloons) will transfer from the hot object to the cold one until both...
Structure Analogy	1. Hot and cold objects correspond to weights on a scale. 2. Heat transfer corresponds to weight redistribution. 3. The point of equilibrium corresponds to the balance point on a scale. 4. The cessation of heat flow corresponds to the stillness of the scale.
Word Analogy	Thermal Equilibrium can be analogous to a Balancing Scale

Table 1: Examples of three types of analogy for a scientific concept.

Examples of these analogies are provided in Table 1, and the prompt templates for generation can be found in Appendix C.1.

3.3 Scientific QA for Student LMs

In the field of human education, a teacher typically introduces a concept to the class and often uses an analogy to clarify the concept (Thagard, 1992; Heywood, 2002; Gray and Holyoak, 2021). For example, when explaining the concept of a cell, drawing an analogy to an automobile factory enhances the understanding, *e.g.*, mitochondria are powerhouses. Such analogies help students grasp the concept of a cell, enabling them to correctly answer related questions on homework quizzes. To align with this, in SCUA task, given a concept with its analogy generated by teacher LMs, we ask the student LMs to answer questions related to the concept. The details of the prompt templates are available in Appendix C.2.

4 Experiment

4.1 Evaluation Protocol

Evaluation Models We choose GPT-4 (OpenAI, 2023), Claude-v3-Sonnet (Anthropic, 2024), Mixtral-8x7B (Mistral AI team, 2023) as **teacher LMs**, and GPT-3.5 (OpenAI, 2022), Gemini (Team and Google, 2023), Mistral-7B (Jiang et al., 2023), Llama3-8B (AI@Meta, 2024), Vicuna-13B and Vicuna-7B (Chiang et al., 2023) as **student LMs**.[2]

Evaluation Collection We evaluate the models on two datasets that feature various levels of question difficulty:

- **ARC Challenge** (Clark et al., 2018): This dataset includes 270 natural science questions that stumped both a retrieval-based and a word co-occurrence algorithm.
- **GPQA** (Rein et al., 2023): With 448 complex multiple-choice questions in biology, physics, and chemistry, it is designed by domain experts. PhD candidates can only achieve 65% accuracy.

[2]The detailed versions for openai models can be found in Appendix B.

Student LMs	Direct	CoT	Analogy (Teacher LMs)		
			GPT-4	Claude	Mixtral
ARC Dataset					
Gemini	88.88	**89.26**	**89.26**	85.56	85.18
GPT-3.5	83.33	84.44	**85.56**	80.37	84.07
Mistral-7B	68.52	70.74	**74.44**	72.59	70.74
LLama3-8B	75.19	77.04	78.89	**80.74**	78.52
Vicuna-13B	37.77	55.56	**63.83**	61.11	62.96
Vicuna-7B	25.55	34.44	**35.56**	34.44	33.42
GPQA Dataset					
Gemini	41.18	41.18	**46.41**	40.52	40.52
GPT-3.5	40.32	41.83	**43.79**	40.52	39.22
Mistral-7B	33.33	32.68	**35.87**	33.33	34.64
LLama3-8B	40.52	44.44	**46.38**	45.10	44.44
Vicuna-13B	26.80	30.72	30.72	**32.55**	30.72
Vicuna-7B	24.84	18.30	**27.45**	**27.45**	25.49

Table 2: Accuracy (%) of different student LMs under different strategies. The analogies generated by teacher LMs, *i.e.*, GPT-4, Claude-v3-Sonnet (**Claude**) and Mixtral-8x7B (**Mixtral**), are all free-form analogies.

Evaluation Metrics For all datasets, we report the accuracy of all questions. Moreover, we randomly sample 100 generated analogies from each dataset and employ three annotators to evaluate their accuracy, with with Fleiss's $\kappa = 0.96$ (Fleiss et al., 1981). The annotation details for quality evaluation of generated analogies are shown in Appendix A.

4.2 Result & Analysis

In the experiments, we expect to answer three research questions:

Teacher	Free-form		Structured		Word	
LMs	ARC	GPQA	ARC	GPQA	ARC	GPQA
GPT-4	100.0	94.0	100.0	98.0	100.0	100.0
Claude	97.0	82.0	100.0	85.0	100.0	100.0
Mixtral	92.0	79.0	95.0	80.0	100.0	100.0

Table 3: The accuracy (%) of three types of analogies generated by different teacher LMs. The results are evaluated by human annotators on 100 samples.

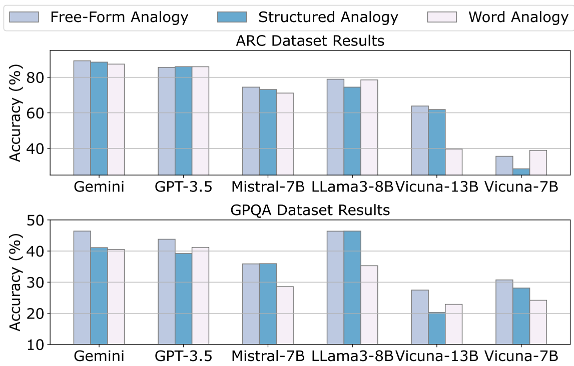

Figure 2: The performance of different student LMs under different types of analogies generated by GPT-4.

RQ1: Can Analogy from Teacher Models Empower Student Models? We adopt Zero-shot Prompting (**Direct**) and Chain-of-Thought Prompting (**CoT**) (Wei et al., 2022) as baselines.[3] The results in Table 2 indicate that: 1) Free-form analogies can indeed help student LMs understand scientific concepts better than Zero-shot and CoT Prompting, improving their ability to answer scientific questions. 2) The analogies generated by GPT-4 improve the ability of student LMs most significantly, indicating the potential of GPT-4 to assist weaker LMs in learning new knowledge. 3) For the GPQA dataset, characterized by specialized concepts and difficult scientific questions, Vicuna-7B and Vicuna-13B perform poorly with Zero-shot and CoT Prompting. However, with analogies, their performance is effectively enhanced. This finding inspires future work to explore using analogies to help the model learn new concepts.

RQ2: Which Type of Analogy Can Better Empower Student Models? Apart from free-form analogy, we also expect to examine two other analogy types, *i.e.*, structured analogy and word analogy, focusing on their effectiveness in aiding student LMs to grasp scientific concepts. As shown in

[3]The prompt templates of the two methods are shown in Appendix C.4.

Model	Direct	CoT	$\text{Analogy}_{\texttt{Self}}$	$\text{Analogy}_{\texttt{GPT-4}}$
Gemini	88.88	89.26	88.88	**89.26**
GPT-3.5	83.33	84.44	84.82	**85.56**
Mistral-7B	68.52	70.74	**77.04**	74.44
LLama3-8B	75.19	77.04	**80.37**	78.89
Vicuna-13B	37.77	55.56	55.93	**63.83**
Vicuna-7B	25.55	34.44	35.42	**35.56**

Table 4: Comparison of self-generated analogies (**$\text{Analogy}_{\texttt{Self}}$**) and GPT-4 generated analogies (**$\text{Analogy}_{\texttt{GPT-4}}$**) for the performance of student LMs on ARC dataset.

Table 3, advanced language models such as GPT-4 and Claude-v3-Sonnet and open-source models like Mixtral-8x7B are proficient in generating high-quality word analogies for scientific concepts. However, the generation quality significantly diminishes for free-form and structured analogies, especially in professional fields (*e.g.*, "wettability" and "contact angle hysteresis" in the GPQA dataset).

In comparison, Figure 2 reveals that compared to word analogy, free-form and structured analogies are more effective in helping models understand scientific concepts due to their more informative content. Future studies can consider strategies that initially have models generate high-quality word analogies, and then expand them into structured and free-form analogies to enhance their quality.

RQ3: How About Self-generated Analogy? In addition to using analogies generated by teacher LMs, we also ask student LMs to generate analogies to help themselves understand scientific concepts and answer related questions. As shown in Table 4, compared to CoT prompting, self-generated analogies can improve the model's understanding of scientific concepts and enhance its ability to answer related questions. Moreover, for some models, self-generated analogies outperform those generated by GPT-4, indicating their ability to use analogies to self-learn new knowledge.

5 Conclusion

In this paper, we propose the SCUA task, which simulates the human education process to explore how analogies created by teacher LMs can help student LMs understand scientific concepts. Our results suggest that free-form analogies indeed aid LMs in comprehending concepts and enhance their ability to answer related scientific questions accurately. Additionally, analogies generated by student

LMs can improve their own performance on scientific quizzes, demonstrating their capability to use analogies for self-learning new knowledge.

Limitations

First, this paper only considers scientific concepts. We do not cover concepts in other fields, such as historical events and social concepts. Second, some previous work (Saha et al., 2023) uses explanations generated by stronger LMs to help weaker LMs. However, we argue that models may have different strengths in different tasks. Therefore, we distinguish between teacher LMs and student LMs without fully evaluating their capabilities. Future work can explore this perspective. Additionally, our evaluation is limited to multiple-choice tasks. Investigating the performance on more complex tasks, such as RAG, would be beneficial.

Ethics Statement

We hereby acknowledge that all authors of this work are aware of the provided EMNLP Code of Ethics and honor the code of conduct.

Use of Human Annotations Evaluation on the generated analogies from stronger LMs in SCUA is implemented by three annotators recruited by our institution. The construction team remains anonymous to the authors. We ensure that the privacy rights of all annotators are respected throughout the annotation process. All annotators are compensated above the local minimum wage and consent to the use of SCUA for research purposes, as described in our paper. The annotation details are shown in Appendix A.

Risks The datasets we conduct in the experiment are sourced from publicly available sources, *i.e.*, ARC Challenge Set and GPQA. However, we cannot guarantee they are free of socially harmful or toxic language. Additionally, analogy evaluation relies on commonsense, and different individuals with diverse backgrounds may have varying perspectives. We use ChatGPT to correct grammatical errors in this paper.

Acknowledgement

We thank the anonymous reviewers for their valuable comments. This work is supported by the Chinese NSF Major Research Plan (No.92270121).

References

AI@Meta. 2024. Llama 3 model card.

Carl Allen and Timothy Hospedales. 2019. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, pages 223–231. PMLR.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Paul Bartha. 2013. Analogy and analogical reasoning.

Mario Bunge. 1981. Analogy between systems. *International Journal Of General System*, 7(4):221–223.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

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

Todd Davies. 1985. Analogy. *CSLI Informal Notes Series IN-CSLI-85-4, Center for the Study of Language and Information, Stanford.*

Zijian Ding, Arvind Srinivasan, Stephen MacNeil, and Joel Chan. 2023. Fluid transformers and creative analogies: Exploring large language models' capacity for augmenting cross-domain analogical creativity. *arXiv preprint arXiv:2302.12832*.

Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

Louis Fournier, Emmanuel Dupoux, and Ewan Dunbar. 2020. Analogies minus analogy test: measuring regularities in word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 365–375, Online. Association for Computational Linguistics.

Dedre Gentner and Kenneth D Forbus. 2011. Computational models of analogy. *Wiley interdisciplinary reviews: cognitive science*, 2(3):266–276.

Dedre Gentner and Arthur B Markman. 1997. Structure mapping in analogy and similarity. *American psychologist*, 52(1):45.

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

Shawn M Glynn, Bruce K Britton, Margaret Semrud-Clikeman, and K Denise Muth. 1989. Analogical reasoning and problem solving in science textbooks. *Handbook of creativity*, pages 383–398.

Maureen E Gray and Keith J Holyoak. 2021. Teaching by analogy: From theory to practice. *Mind, Brain, and Education*, 15(3):250–263.

Qianyu He, Jie Zeng, Qianxi He, Jiaqing Liang, and Yanghua Xiao. 2024. From complex to simple: Enhancing multi-constraint complex instruction following ability of large language models. *arXiv preprint arXiv:2404.15846*.

Dave Heywood. 2002. The place of analogies in science education. *Cambridge Journal of Education*, 32(2):233–247.

Douglas R Hofstadter. 2001. Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, pages 499–538.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Cheng Jiayang, Lin Qiu, Tsz Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. StoryAnalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11518–11537, Singapore. Association for Computational Linguistics.

Hyeonsu B. Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting scientific creativity with an analogical search engine. *ACM Trans. Comput.-Hum. Interact.* Just Accepted.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9114–9128, Singapore. Association for Computational Linguistics.

Mistral AI team. 2023. Mixtral of experts. Accessed: 2023-12-15.

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt-4 technical report.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Lindsey Engle Richland and Nina Simms. 2015. Analogy, higher order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(2):177–192.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. 2023. Can language models teach? teacher explanations improve student performance via personalization. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Oren Sultan, Yonatan Bitton, Ron Yosef, and Dafna Shahaf. 2024. Parallelparc: A scalable pipeline for generating natural-language analogies. *arXiv preprint arXiv:2403.01139*.

Oren Sultan and Dafna Shahaf. 2022. Life is a circus and we are the clowns: Automatically finding analogies between situations and processes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3547–3562, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gemini Team and Google. 2023. Gemini: A family of highly capable multimodal models.

Paul Thagard. 1992. Analogy, explanation, and education. *Journal of research in science teaching*, 29(6):537–544.

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? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.

PeiFeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2023a. PINTO: Faithful language reasoning using prompt-generated rationales. In *The Eleventh International Conference on Learning Representations*.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023b. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2022. Emergent analogical reasoning in large language models. *arXiv preprint arXiv:2212.09196*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549, Toronto, Canada. Association for Computational Linguistics.

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Tiyyala, Nicholas Andrews, and Daniel Khashabi. 2024. Analobench: Benchmarking the identification of abstract and long-context analogies. *arXiv preprint arXiv:2402.12370*.

Siyu Yuan, Jiangjie Chen, Xuyang Ge, Yanghua Xiao, and Deqing Yang. 2023. Beneath surface similarity: Large language models make reasonable scientific analogies after structure abduction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2446–2460, Singapore. Association for Computational Linguistics.

Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. ANALOGYKB: Unlocking analogical reasoning of language models with a million-scale knowledge base. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1249–1265, Bangkok, Thailand. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

A Crowd-sourcing Details

We have recruited a team of three undergraduates. To process conflicting annotations, we adopt a voting majority principle to determine the results. We pay each annotator $8/h, exceeding the local minimum wage. The screenshots of the instructions and interface for quality check of the extracted concepts and generated analogies annotation are shown in Figure 3 and Figure 4.

B Model Selection

For OpenAI models, we use `gpt-3.5-turbo-0613` and `gpt-4-0613`.[4] For Gemini-pro, we use Google Gemini-Pro APIs to obtain results. We set the temperature to 0 for all models.

C Prompt Template of SCUA

C.1 Analogy Generation

The prompt of the analogy generation from teacher LMs is given in List 1.

Listing 1: Instruction templates for teacher LMs to generate analogies.

```
Free-Form Analogy Generation:
Please use an analogy with no more than
100 words to explain the scientific
concept:
Example:
Concept: Thermal Equilibrium
Analogy: Imagine you're making a cup of
hot chocolate on a cold winter day. You
heat up the milk on the stove until it's
 steaming hot, then you pour it into
your favorite mug that's been sitting at
 room temperature. When you first pour
the hot milk into the mug, there's a big
 temperature difference between the two
- the milk is hot, and the mug is
relatively cold. But if you wait a few
minutes before taking a sip, you'll
notice that the mug has warmed up, and
the milk has cooled down a bit. This is
because heat has transferred from the
hot milk to the cooler mug until they've
 reached a point where they're the same
temperature. This is thermal equilibrium
.Just like the hot milk and the mug,
when two objects at different
temperatures come into contact, heat
will always flow from the hotter object
to the cooler one. This continues until
they reach thermal equilibrium, or the
same temperature. Once they're at the
same temperature, there's no more heat
flow because there's no temperature
difference to drive it.
Concept: {scientific_concept}
Analogy:
```

[4] https://platform.openai.com/docs/models

```
Structure Analogy Generation:
Given one scientific concept, you should
 use another concept as an analogy to
explain this concept. Moreover, you
should use other concepts that are
related to these two concept to explain
the analogy:
Example:
Concept: Thermal Equilibrium
Analogy: Thermal Equilibrium can be
analogous to a Balancing Scale
1. Hot and cold objects correspond to
weights on a scale: Just as a hot object
 and a cold object interact to reach
thermal equilibrium, weights on a scale
interact to reach a balanced state. The
hot object, like a heavier weight, has
an excess (of heat or weight) that it
transfers to the cold object or lighter
weight.
2. Heat transfer corresponds to weight
redistribution: In thermal equilibrium,
heat transfers from the hot object to
the cold object until they reach the
same temperature. Similarly, on a
balancing scale, weight redistributes
from the heavier side to the lighter
side until they reach the same level.
3. The point of equilibrium corresponds
to the balance point on a scale: In
thermal equilibrium, the point of
equilibrium is when both objects reach
the same temperature. On a balancing
scale, the balance point is reached when
 both sides of the scale are at the same
 level, indicating that the weights are
equal.
4. The cessation of heat flow
corresponds to the stillness of the
scale: Once thermal equilibrium is
reached, there is no more heat flow
because there's no temperature
difference to drive it. Similarly, once
a scale is balanced, there is no more
movement because there's no weight
difference to drive it.
Concept: {scientific_concept}
Analogy:


Word Analogy Generation:
Given one scientific concept, you should
 use another concept as an analogy to
explain this concept:
Example:
Concept: Thermal Equilibrium
Analogy: Thermal Equilibrium can be
analogous to a Balancing Scale
Concept: {scientific_concept}
Analogy:
```

C.2 Question Answering

The prompt of the question answering by student LMs is given in List 2.

Listing 2: Instruction templates for student LMs to answer questions based on analogies.

```
You need to select an answer for a
question.
This is the question:
{question}
{choices}
Since the question is difficult, we
asked a teacher to explain the concepts
in this question to you using analogies,
 which we hope can help you.
This is the explanation with analogies:
{analogy}
Please combine the explanation to better
 answer this question.
Answer:
```

C.3 Concept Extraction

The prompt of the concept extraction by GPT-4 is given in List 3.

Listing 3: Instruction templates for GPT-4 to extract scientific concepts.

```
Given a scientific question, you should
show the key scientific concept related
to this scientific question.
This is a scientific question:
{question}
The key scientific concept:
```

C.4 The Prompt Templates of Zero-shot and CoT Prompting

The prompt of Zero-shot and CoT Prompting is given in List 4.

Listing 4: Instruction templates for Zero-shot and CoT Prompting.

```
Zero-shot Prompting:
{question}
{Options}
Answer:

CoT Prompting:
{question}
{Options}
You need to give the reason first and
then choose the answer.
Answer:
```

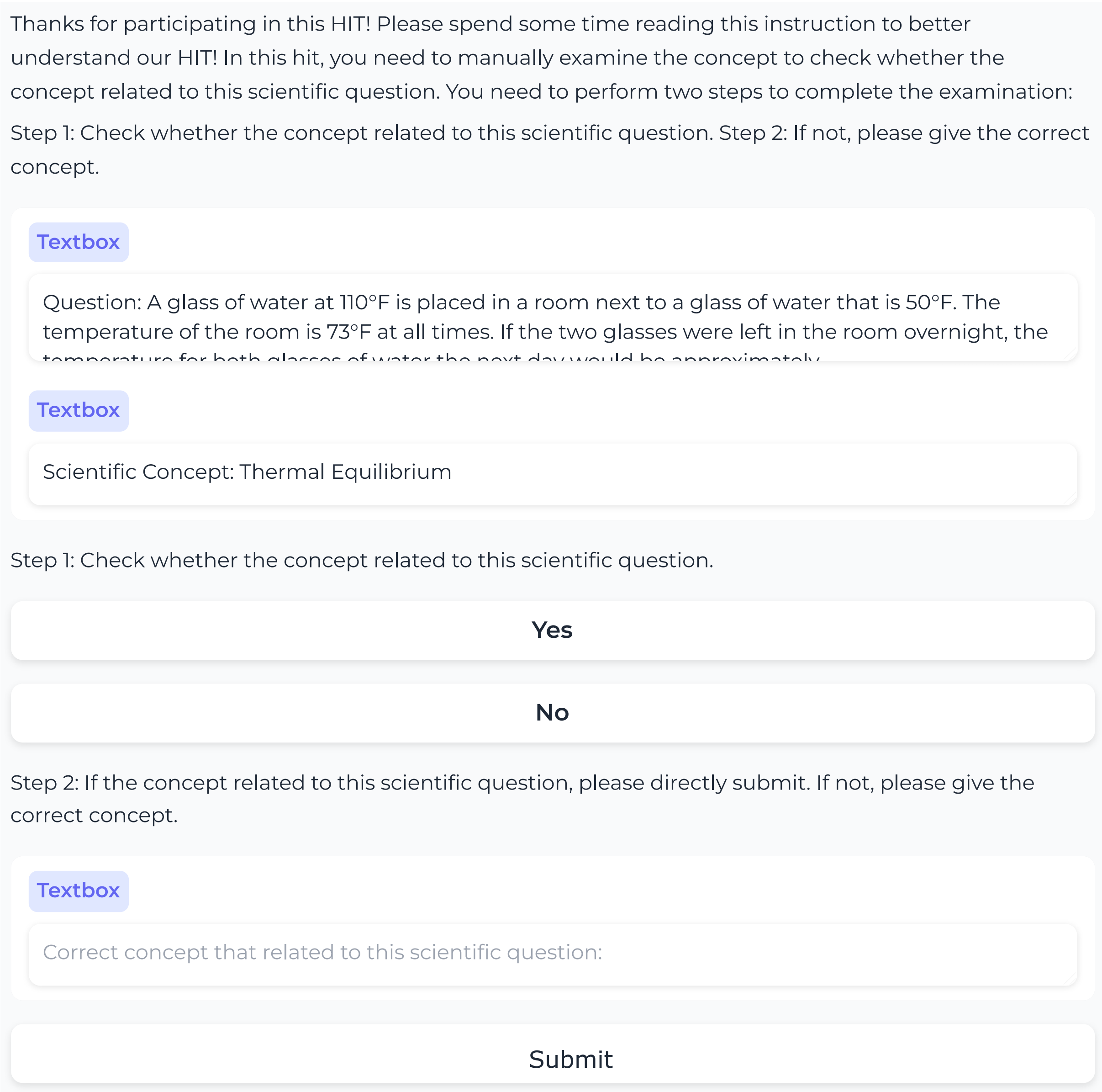

Figure 3: The screenshots of the instructions and interface for extracted concept annotation.

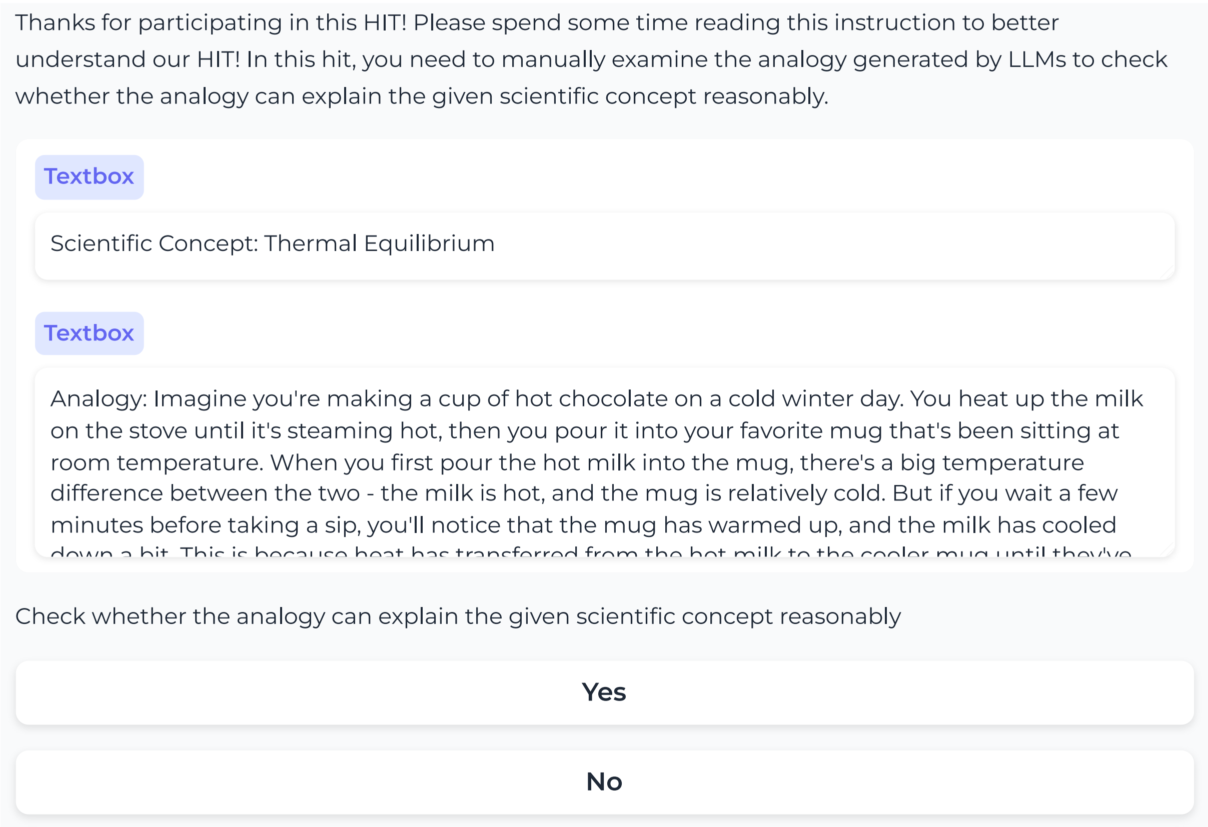

Figure 4: The screenshots of the instructions and interface for generated analogy annotation.