Long-Form Analogy Evaluation Challenge

Bhavya Bhavya[1], Chris Palaguachi[1], Yang Zhou[1], Suma Bhat[1], and ChengXiang Zhai[1]

[1]University of Illinois at Urbana-Champaign
{bhavya2, cwp5, yz96, spbhat2, czhai}@illinois.edu

Abstract

Given the practical applications of analogies, recent work has studied analogy generation to explain concepts. However, not all generated analogies are of high quality and it is unclear how to measure the quality of this new kind of generated text. To address this challenge, we propose a shared task on automatically evaluating the quality of generated analogies based on seven comprehensive criteria. For this, we will set up a leaderboard based on our dataset annotated with manual ratings along the seven criteria, and provide a baseline solution leveraging GPT-4. We hope that this task would advance the progress in development of new evaluation metrics and methods for analogy generation in natural language, particularly for education.

1 Introduction

Analogies are integral to several practical applications. In education, they help explain complex concepts by mapping them to more familiar ones (Glynn et al., 1989; Thagard, 1992) (e.g., "earth rotates on its axis like an ice skater doing a pirouette"). They also inspire creativity by connecting seemingly disparate concepts (Hey et al., 2008).

Since manually creating good analogies can be challenging and require domain expertise (Goldwater et al., 2021), recently, large language models (LLMs) like GPT-3 (Brown et al., 2020) have been used to aid with all such applications (Bhavya et al., 2022, 2023; Kim et al., 2023). They have shown great promise in generating long-form analogies (i.e., natural language analogies, typically a few paragraphs long, that describe the similarities between concepts) that are meaningful, novel (Bhavya et al., 2022, 2023) and useful for science writers (Kim et al., 2023).

However, not all automatically generated analogies are accurate or useful. Poor analogies can have negative consequences, such as, leading to misunderstanding or misconceptions (Kaufman et al., 1996). This effect can be particularly concerning when such analogies are used in educational contexts, where clarity and accuracy are crucial. Thus, evaluating the quality of generated analogies is important to identify good analogies. Although a human evaluation of all generated analogies would be ideal, it is impossible to scale up. Thus, there is a need for automatic evaluation metrics. Moreover, there is a need to develop evaluation metrics for this new type of generated text to measure the progress of analogy generation methods.

While several automatic evaluation metrics have been developed to evaluate generated text (Sai et al., 2022), they are not directly applicable to evaluate analogies. Limited work has been done on automatically evaluating generated analogies using reference-based metrics (e.g., BLEURT (Sellam et al., 2020)) and reference-free metrics (e.g., novelty estimation based on similarity to a reference corpus of analogies) (Bhavya et al., 2022, 2023). Such metrics have mostly been found to be inadequate. Moreover, it is unclear as to what precisely makes a good generated analogy since its goodness depends on multiple factors (e.g., accuracy, strength of analogical connections).

To address these challenges, we propose a new shared task for developing evaluation metrics that measure the quality of generated analogies. Specifically, we identify seven major criteria for evaluating their quality based on existing literature and our pre-pilot experiments, namely, target concept comprehensiveness, accessibility, source and target concept accuracy, mapping soundness, coherence, and repetition. Based on these evaluation criteria, we will create a dataset of manually rated analogies that are generated by models like GPT-4 in domains like science. This dataset will be used to assess the performance of automatic evaluation metrics submitted to our task.

Since LLMs have recently shown great promise in evaluating generated text (Li et al., 2024), we

Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges, pages 1–16
September 23 –27, 2024.

will provide a baseline method that prompts GPT-4 for evaluation in a reference-free setting. We've found this method to be reasonably accurate based on pre-pilot experiments. But, we encourage participants to develop metrics using smaller language models and other types of models too (e.g., fact verification models for accuracy).

Similar to shared tasks on evaluation metrics for other NLG tasks (e.g., machine translation (Blain et al., 2023)), we expect our proposed task to accelerate research in both evaluation metric and text generation methods, particularly in the context of long-form analogies. More broadly, the insights from the task would also be useful for evaluating other kinds of generated long and creative text (e.g., stories). With the advent of LLMs, generation of various kinds of text has become feasible and useful for many practical applications. Therefore, we believe that this is a timely novel shared task.

2 Task Description

Given a generated analogy to explain a target concept, the overall task is to rate its quality based on defined criteria. A leader board competition would be set up to evaluate the submissions on our task and dataset. In this section, we describe the criteria we plan to use for evaluation of analogies, our datasets of human ratings and evaluation metrics to quantitatively evaluate the automatic ratings submitted to the task, and our proposed schedule.

2.1 Analogy Evaluation Criteria

Few recent work have studied evaluation of automatically generated analogies (Kim et al., 2023; Bhavya et al., 2022, 2023). Inspired by these and prior work (e.g., (Forbus and Gentner, 1989), (Glynn et al., 1989)), and further refinement based on our pre-pilot experiments (Section 3), we select seven criteria for a holistic evaluation of analogies.

Our selected criteria include measures for three main components of long-form analogies, namely, target concept, source concept, and mapping. Target is the more unfamiliar concept, and the source is the more familiar one used to explain the target. The mapping is the set of relationships or similarities between the source and the target.

For example, consider the following analogy: *"The heart is like a pump in the body's circulatory system. The pump moves fluid through a system, just as the heart moves blood through the body."* In this analogy, "the heart" is the target concept and "the pump" is the source concept. The mapping is "the pump ... the body."

We describe each of the seven criteria below. All criteria will be rated on an Ordinal scale.

Target concept comprehensiveness: Whether the analogy covers the most important details to explain the target concept.

Accessibility: Whether the analogy is familiar and easily understandable by learner.

Source Accuracy and Target Accuracy: Truthfulness of stated facts pertaining to the two analogous concepts. Instead of a single measure of overall accuracy, analyzing its two components separately is useful for applications like education, where one of them (e.g., target accuracy) is more critical.

Mapping soundness: Whether the connection between source and target is logically sound or farfetched.

Coherence: Whether the analogy is cohesive.

Repetition: Whether the same sentence is repeated or same source concept is repeated for another target concept within the analogy.

2.2 Analogy Ratings Dataset

We plan to create an annotated dataset with human ratings to quantitatively evaluate the automatic evaluation metric submissions as described below.

Analogy Collection: To enable creation of diverse and representative data, we will include analogies that vary on the following two dimensions.

Target concept domain: Given the popularity of analogies in teaching STEM subjects(Cao et al., 2023; Glynn et al., 1989), we will focus on science and computer science domains. Depending on budget and feasibility of recruiting suitable raters, we will include other domains, such as, economics and political science. For the science domain, we will leverage existing datasets of generated analogies (Bhavya et al., 2022; Kim et al., 2023). Within a particular domain, we will consider rating analogies about target concepts of varying grade-level difficulty (e.g., beginner, intermediate, and advanced) because we expect the quality of generated analogies to differ based on them.

Generation method: Another interesting variable that impacts the quality of generated analogies is the model used for generation. For example, larger models typically generate better analogies (Bhavya et al., 2022; Kim et al., 2023). Following such work, we mainly plan to leverage the GPT-family of models, including GPT-3, GPT3.5 and GPT-4 (Brown et al., 2020; Ouyang et al., 2022; Achiam

et al., 2023).

The style of generated analogy also differs based on the model and prompt used while generation. For example, GPT-3-generated analogies in one prior dataset (Bhavya et al., 2022) generally contain a single analogical comparison. While, prompts designed in another work (Kim et al., 2023) generate analogies containing several comparisons (aka "sub-analogies"). For instance, in the following analogy, *"Stratosphere is like the sky because ... Troposphere is like the earth.."*, "stratosphere" is compared to "sky", and "troposhere" to "earth". We do not plan to do an extensive exploration of prompt design, but will mostly leverage prompts from prior research.

Rating procedure: For rating analogies based on our evaluation criteria, we plan to recruit human annotators on Upwork[1], a free-lancing platform that has been used in similar prior work (Kim et al., 2023; Ouyang et al., 2022). Annotator requirements include English proficiency and prior teaching experience in the particular domain. The final set of qualified raters (up to 20 per domain) would be selected based on their performance on rating a small test batch. Each sample would be rated by three raters. We will follow other best practices for annotation and reporting (van der Lee et al., 2021; Howcroft et al., 2020), including detailed task instructions, as shown in Appendix A.1. Each rater would be paid an hourly wage of about $25-$35.

Dataset statistics: Our data would consist of validation and test sets only and no training set. To enable calibration of automatic metrics, we will use a validation set for evaluating submissions on the leader board. After the competition is over, submissions will be evaluated on a blind test set.

We plan to collect at least 1k manually rated analogies. The final number of rated analogies would mainly depend on budget and time constraints. 50% of this data would be released as the validation set and the remaining 50% would be the test set.

Evaluation of analogies would be done in a reference-free setting. This is mainly because there are many equally plausible analogies relevant for a given concept and building an exhaustive reference corpus of analogies for all concepts in the dataset is impossible. Thus, we will not release any such resources. However, participants would be free to use any external knowledge (e.g., web data).

[1]https://www.upwork.com/

2.3 Evaluation of automatic metrics

To evaluate the submitted automatic evaluation metrics, we will compare them with human ratings on each of the seven evaluation criteria using the following statistics.

Kendall's tau-b: It is commonly used to compare the rank order of automatic evaluation metrics with human ratings (Kendall, 1945; Sellam et al., 2020).

Kendall's tau-b after outlier removal: We will also measure Kendall's tau after removing outliers to avoid spurious correlations (Mathur et al., 2020).

Pairwise accuracy: To mitigate short-comings of kendall's tau in case of several ties, this metric uses pairwise accuracy, which rewards metrics for both predicting correct pair rankings and correctly predicting ties, and a tie calibration method that allows for comparing metrics that do and do not predict ties (Deutsch et al., 2023).

Krippendorff's alpha: Agreement after accounting for chance-agreements (Krippendorff, 2011).

Mean Squared Error: This measures the average difference between squared values of human and automatic ratings (James, 2013).

2.4 Baseline method

Recently, prompting LLMs like GPT-4 has shown great potential in automatically evaluating generated text based on several criteria like accuracy, coherence, and engagement in both reference-free and reference-based settings (Liu et al., 2023; Chhun et al., 2024; Li et al., 2024; Wang et al., 2023). Our pre-pilot experiments (Section 3) show reasonable results of this method on our task as well. Accordingly, we will design suitable prompts for automatic evaluation with GPT-4 based on our criteria. But, we encourage participants to leverage smaller and other kinds of models as well. For fairness, we will separately report the performances of different types of models (e.g., based on LLM size, use of external resources, etc.).

2.5 Schedule

We propose the following schedule:

September, 2024: The shared task is announced at the INLG conference. Validation data is available on the shared task website and participants can sign up for the task.

December 1st, 2024: Leaderboard based on our test sets are open for the shared task. Participants can submit their solutions and view their updated ranking on the online leaderboard based on perfor-

mance on the validation set.
April 1st, 2025: Submissions are closed. Organizers conduct automatic evaluation of all submissions on the blind test set.
June 1st, 2025: Organizers will submit participant reports and overall challenge reports to INLG 2025 and present their findings.

3 Pre-pilot study

To understand the task feasibility and guide the task design, we conducted a pre-pilot study. Below, we describe the initial evaluation criteria and manual rating datasets used in this study, the results of prompting GPT-4 for automatic evaluation, and qualitative discussions to refine these criteria and finalize the ones reported in Section 2.1.

3.1 Evaluation Criteria

In addition to source and target accuracy defined in Section 2.1, we analyzed the following four criteria, guided by prior research, for the pre-pilot study.
Meaningfulness: Whether it is an accurate and coherent analogy (Bhavya et al., 2022, 2023).
Novelty: How unique is the generated text (Bhavya et al., 2023). It could be important for creative writing applications (Kim et al., 2023).
Usefulness: Overall utility of the analogy for explaining concepts, since it is one of the most important use-cases of analogies (Glynn et al., 1989).
Structural mapping consistency: It is defined by the following two constraints from Structural Mapping Engine framework (Forbus and Gentner, 1989). 1:1 constraint means that one attribute of the source concept should be connected to at most one attribute of the target and vice versa. The parallel connectivity constraint states that if two concepts are connected, then so must their attributes.

3.2 Datasets

We use the following three datasets for this study.

3.2.1 Meaningfulness and Novelty Datasets

For meaningfulness and novelty, we use datasets from previous work (Bhavya et al., 2022, 2023). In particular, one work (Bhavya et al., 2022) asked crowd-workers to rate 1608 science analogies on a binary scale for meaningfulness. Of these, 1543 are generated by GPT-3 models of various sizes (ranging from 0.3B to 175B) and 65 are human-generated ones scraped from online websites like chegg.com. We call this dataset as **BAM** for Binary Analogy Meaningfulness.

In another work (Bhavya et al., 2022), crowd-workers were asked to rate 347 GPT-3-generated science analogies on both meaningfulness and novelty on a scale of 1-4. We call this dataset as **OAMN** for Ordinal Analogy Meaningfulness and Novelty. Three annotators rated each analogy in both cases.

Table 1: Krippendorff's alpha (α) between human annotator (ann.) and GPT-4 on automatically and human generated analogies in BAM.

	Auto-generated	Human-generated
All ann.	0.49	0.22
GPT-4 v. ann.	0.56 ± 0.009	0.35 ± 0.045

Table 2: Krippendorff's alpha (α) and Kendall's tau (τ) between human annotator (ann.) and GPT-4 on OANM.

	Meaningfulness		Novelty	
	α	τ	α	τ
All ann.	0.247	-	0.4	-
GPT-4 v. ann.	0.46 ± 0.02	0.48 ± 0.02	0.33 ± 0.003	0.33 ± 0.001

3.2.2 Multi-Aspect Analogy Annotation for Education (MANAED)

For the remaining four criteria, we manually rate a 50 analogies about 7 target concepts released by another work (Kim et al., 2023).[2] Two researchers, a graduate student in Educational Psychology and an undergraduate in Computer Science, rate each analogy on a scale of 1-4 for all criteria. Source and target accuracy were rated at the sub-analogy level (refer Section 2.2, Generation method).

3.3 Experiments

Using the above datasets, we study the feasibility of prompting GPT-4 for automatic analogy evaluation, and the suitability of our evaluation criteria based on the quantitative and qualitative results.
Methodology: We leverage prompt templates from recent work on prompting GPT-4 for text evaluation (Liu et al., 2023), and conduct light prompt-tuning, including the use of suitable instructions and examples for our task. The best performing prompts for each criteria are shown in Appendix A.2.

We quantitatively compare GPT-4 (gpt4-0125-preview) ratings with average human ratings based on Krippendorff's alpha and Kendall's tau. As an upper limit, we also report the inter-annotator agreements and correlations (if applicable). Further, qualitative discussions and analysis of manual

[2] Although they release manual ratings by science writers on some criteria, those are not usable because ratings cannot be mapped to their corresponding analogies.

Table 3: Krippendorff's alpha (α) and Kendall's tau (τ) between human annotators (ann.) and GPT-4 on MANAED

	Structural Consistency		Usefulness		Source Accuracy		Target Accuracy	
	α	τ	α	τ	α	τ	α	τ
All ann.	0.6	58	0.62	0.56	0.51	0.48	0.49	0.48
GPT-4 v. ann.	0.23 ± 0.05	0.2 ± 0.05	0.29 ± 0.07	0.25 ± 0.06	0.37 ± 0.001	0.33 ± 0.001	0.31 ± 0.01	0.3 ± 0.02

ratings were conducted to refine criteria.

Results: From Tables 1, 2 and 3, on all the six criteria, GPT-4 generally achieves fair to moderate agreements and correlations (Landis and Koch, 1977; Schober et al., 2018), suggesting its feasibility to use as a baseline method.

On meaningfulness, from Tables 1 and 2, we observe that GPT-4's agreement and correlation with human ratings is comparable to that among humans. Due to this already strong performance of GPT-4, we discard this criteria for the main task.

Results for novelty and other remaining criteria are in Tables 2 and 3, respectively. For these criteria, there is a gap between GPT-4 and human performance, suggesting room for research.

After discussions, we discard novelty because it depends on training and reference dataset. For instance, an analogy can be considered not novel (or novel) depending on whether the model that generates it has seen it during training (or not).

Further, by analyzing annotator disagreements, we identified usefulness to be highly subjective because it spans multiple aspects. So, we identify the following three salient aspects, aligned with prior research (Glynn et al., 1989), that impact utility of long-form analogies for education, in addition to our other included criteria: "target comprehensiveness", "accessibility", and "mapping soundness".

Additionally, the two structural mapping constraints are decoupled and adapted for LLM-generated analogies. In this way, we finalize "repetition", corresponding to 1:1 constraint, and "coherence", corresponding to parallel connectivity.

4 Related Work

Prior work has studied the modeling and generation of various forms of analogies (Mitchell, 2021), such as, analogies between structured representations of concepts (Forbus et al., 2017), relational and proportional analogies (e.g., king:queen::man:woman) (Ushio et al., 2021; Yuan et al., 2023; Chen et al., 2022), analogies relating longer text, such as, two sentences or stories (Jiayang et al., 2023; Wijesiriwardene et al., 2023; Sultan et al., 2024), and more recently, *long-form* analogies that explain the relation between concepts using natural language (Seals and Shalin, 2023; Bhavya et al., 2022, 2023; Kim et al., 2023; Cao et al., 2023). We aim to evaluate long-form analogies that are typically a few paragraphs long.

Human evaluation of generated text, although ideal, is highly resource extensive. Accordingly, several automatic metrics have been developed for evaluating generated text (Sai et al., 2022), and shared tasks have been established to drive such efforts (Blain et al., 2023). We build upon recent work on holistic evaluation of other types of figurative and creative text (Chhun et al., 2022; He et al., 2023), because it enables a fine-grained evaluation. However, for automatic evaluation of generated long-form analogies, there has been very limited work (Bhavya et al., 2022, 2023; Kim et al., 2023). We compile and refine seven major evaluation criteria based on these and prior work on analogical modeling and reasoning (Falkenhainer et al., 1989; Glynn et al., 1989), aim to extend their datasets both in the number of samples and ratings based on our criteria, and call for development of suitable automatic evaluation metrics.

5 Conclusion

We propose a new shared task for development of automatic metrics to evaluate generated long-form analogies, which describe the analogical relation between concepts in natural language, on seven comprehensive criteria. The submissions would be evaluated based on their agreement with human ratings on our datasets. With this shared task, we hope to accelerate the progress in evaluation metrics and generation methods for long-form analogies.

6 Acknowledgment

This material is based upon work supported by the National Science Foundation and the Institute of Education Sciences under Grant #2229612. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation or the U.S. Department of Education.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bhavya Bhavya, Jinjun Xiong, and Chengxiang Zhai. 2022. Analogy generation by prompting large language models: A case study of instructgpt. *arXiv preprint arXiv:2210.04186*.

Bhavya Bhavya, Jinjun Xiong, and Chengxiang Zhai. 2023. Cam: A large language model-based creative analogy mining framework. In *Proceedings of the ACM Web Conference 2023*, pages 3903–3914.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M Guerreiro, Diptesh Kanojia, José GC de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, et al. 2023. Findings of the wmt 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chen Cao, Zijian Ding, Gyeong-Geon Lee, Jiajun Jiao, Jionghao Lin, and Xiaoming Zhai. 2023. Elucidating stem concepts through generative ai: A multi-modal exploration of analogical reasoning. *arXiv preprint arXiv:2308.10454*.

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. *arXiv preprint arXiv:2203.08480*.

Cyril Chhun, Pierre Colombo, Chloé Clavel, and Fabian M Suchanek. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. *arXiv preprint arXiv:2208.11646*.

Cyril Chhun, Fabian M Suchanek, and Chloé Clavel. 2024. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *arXiv preprint arXiv:2405.13769*.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Brian Falkenhainer, Kenneth D Forbus, and Dedre Gentner. 1989. The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, 41(1):1–63.

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.

Kenneth D Forbus and Dedre Gentner. 1989. Structural evaluation of analogies: What counts. In *Proceedings of the eleventh annual Conference of the Cognitive Science Society*, volume 34, pages 341–348.

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.

Micah B Goldwater, Dedre Gentner, Nicole D LaDue, and Julie C Libarkin. 2021. Analogy generation in science experts and novices. *Cognitive Science*, 45(9):e13036.

Qianyu He, Yikai Zhang, Jiaqing Liang, Yuncheng Huang, Yanghua Xiao, and Yunwen Chen. 2023. Hauser: Towards holistic and automatic evaluation of simile generation. *arXiv preprint arXiv:2306.07554*.

Jonathan Hey, Julie Linsey, Alice M Agogino, and Kristin L Wood. 2008. Analogies and metaphors in creative design. *International Journal of Engineering Education*, 24(2):283.

David M Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *13th International Conference on Natural Language Generation 2020*, pages 169–182. Association for Computational Linguistics.

G James. 2013. An introduction to statistical learning.

Cheng Jiayang, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, et al. 2023. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. *arXiv preprint arXiv:2310.12874*.

David R Kaufman, Vimla L Patel, and Sheldon A Magder. 1996. The explanatory role of spontaneously generated analogies in reasoning about physiological concepts. *International Journal of Science Education*, 18(3):369–386.

Maurice G Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging large language models to support extended metaphor creation for science writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pages 115–135.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging large language models for nlg evaluation: A survey. *arXiv preprint arXiv:2401.07103*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv:2006.06264*.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

SM Seals and Valerie L Shalin. 2023. Long-form analogies generated by chatgpt lack human-like psycholinguistic properties. *arXiv preprint arXiv:2306.04537*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

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

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? *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.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

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.

Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2023. Analogykb: Unlocking analogical reasoning of language models with a million-scale knowledge base. *arXiv preprint arXiv:2305.05994*.

A Appendix

A.1 Sample instructions for manually rating analogies

Task Overview:

By connecting abstract or unfamiliar concepts (called the target) to more familiar ones (called the source), analogies play a huge role in education as they help with understanding concepts, problem-solving, increasing learners' interest and motivation.

For example, "The heart is like a pump in the body's circulatory system.The pump moves fluid through a system, just as the heart moves blood through the body."

In this analogy, the heart is the target concept and the pump is the source concept. The mapping is the set of relationships or correspondences between the source and the target. In the example above, the mapping is: The pump moves fluid through a system, just as the heart moves blood through the body.

Your task is to rate analogies based on seven criteria defined below.

Target concept comprehensiveness/scope: Whether the analogy covers the most important details to explain the target concept

1 - Does not cover anything; not suitable for anyone

2 - Covers sufficient details for elementary school

students and beginners
3 - Covers sufficient details for middle school students and intermediate learners
4 - Covers sufficient details for high school students and advanced learners
Examples:
1- Does not cover anything; not suitable for anyone:
Target concept: Photosynthesis, Analogy: "Photosynthesis is like a tree eating sunshine."
This analogy is too simplistic and doesn't cover any important details about photosynthesis. It doesn't explain the process, components involved, or the purpose of photosynthesis.
2 - Covers sufficient details for elementary school students and beginners:
Target concept: The water cycle, Analogy: "The water cycle is like a never-ending merry-go-round. Water from puddles, lakes, and oceans gets warmed by the sun and turns into vapor that rises into the sky. It forms clouds, and when the clouds get heavy, the water falls back to Earth as rain or snow, starting the ride all over again." This analogy covers basic components of the water cycle (evaporation, condensation, precipitation).
3- Covers sufficient details for middle school students and intermediate learners:
Target concept: The immune system, Analogy: "The immune system is like a well-organized army protecting a country. It has scouts (white blood cells) that patrol the body looking for invaders (pathogens). When they spot an enemy, they alert the command center (lymph nodes) which then sends out specialized troops (antibodies) to fight the specific invader. The army also keeps records of past battles (memory cells) to respond more quickly if the same invader returns."
This analogy covers more complex aspects of the immune system, including different types of cells and their functions, making it suitable for intermediate learners.
4- Covers sufficient details for high school students and advanced learners:
Target concept: DNA replication, Analogy: "DNA replication is like a highly efficient book-copying process in a specialized library. The original DNA double helix is the master book, which is carefully unzipped (by helicase enzymes) into two single strands. Each strand serves as a template for creating a new complementary strand. Skilled workers (DNA polymerase) move along each template, reading the sequence and adding corresponding nucleotides to build the new strands. They work in a specific direction (5' to 3'), creating a continuous leading strand and a fragmented lagging strand (Okazaki fragments). Proofreaders (exonuclease function) check for errors, and librarians (ligase enzymes) connect the fragments. The result is two identical copies of the original DNA book, each containing one old and one new strand."
This analogy covers detailed aspects of DNA replication, including enzyme names, directionality, and specific processes like the formation of Okazaki fragments. It's suitable for advanced learners or high school students studying biology.
Accessibility:
Whether the analogy is familiar and easily understandable by learner
1 - Easily understandable by elementary school students and beginners
2 - Easily understandable by middle school students and intermediate learners
3 - Easily understandable by high school students and advanced learners
Examples
1 - (Elementary school/Beginners):
Target concept: The water cycle, Analogy: "The water cycle is like a merry-go-round. Water goes up into the sky, forms clouds, falls as rain, and then goes back up again, just like how you go up and down on a merry-go-round."
This analogy uses a merry-go-round, which is a simple, familiar concept for young children.
2 - (Middle school/Intermediate):
Target concept: Photosynthesis, Analogy: "Photosynthesis is like a plant's kitchen. The leaves are the chef, sunlight is the stove, water and carbon dioxide are the ingredients, and glucose is the meal the plant makes for itself."
This analogy uses the concept of a kitchen, which is familiar to most people but requires a slightly more abstract understanding than the merry-go-round example. It introduces more specific terms (like "ingredients" and "glucose") and requires understanding the idea of transforming ingredients into a meal.
3 - (High school/Advanced):
Target concept: DNA replication, Analogy: "DNA replication is like creating a backup of an important computer file. The original DNA strand serves as a template, much like the original file, while enzymes act as the copying software, creating an exact duplicate to ensure the genetic information is

preserved and can be passed on."
This analogy uses the concept of computer file backup, which is more technologically advanced and less universally familiar than the previous examples.

Mapping soundness:

Whether the similarity between source and target is logically sound

1- No mapping; connections are illogical or nonexistent. Includes cases where the text is an example, definition, or a tautology.

2 - Weak mapping; similarities are far-fetched or barely logical

3 - Moderate mapping; similarities are mostly logical with minor logical stretches

4 - Strong mapping; similarity between source and target concept is logically sound

Examples

1- No mapping; connections are illogical or nonexistent:

Analogy: "Life is like a banana because both have letters."

This is illogical as the presence of letters in words is not a meaningful connection between life and bananas.

2- Weak mapping; similarities are far-fetched or barely logical:

Analogy: "A computer is like a sandwich because both can be found on a desk."

While it's possible to find both on a desk, this similarity is superficial and doesn't relate to their core functions or properties.

3 - Moderate mapping; similarities are mostly logical with minor logical stretches:

Analogy: "A library is like a beehive because both are organized systems where individuals work together to store and access valuable resources."

This analogy has logical similarities in terms of organization and collective effort, but stretches the comparison between books/information and honey/pollen.

4- Strong mapping; similarity between source and target concept is logically sound:

Analogy: "The human brain is like a computer because both process information, store memories, and can be overloaded with too much input."

This analogy draws clear, logical parallels between the functions of a brain and a computer, with sound similarities in their core operations.

Coherence:

Whether the whole analogy is cohesive

1 - All the sentences or sub-analogies are disconnected

2 - Most of the sentences or sub-analogies are disconnected

3 - Some of the sentences or sub-analogies are disconnected

4 - None of the sentences or sub-analogies are disconnected

Examples

1 - All the sentences or sub-analogies are disconnected:

Target concept: Learning, Analogy: "Learning is like climbing a mountain. Fish swim in schools. Knowledge weighs nothing. Time is money."

2 - Most of the sentences or sub-analogies are disconnected:

Target concept: Career growth, Analogy: "Career growth is like tending a garden. You need to plant seeds of opportunity. Success doesn't happen overnight. A rolling stone gathers no moss."

3 - Some of the sentences or sub-analogies are disconnected:

Target concept: Problem-solving, Analogy: "Problem-solving is like untangling a knot. You need patience and persistence to work through the complications. Sometimes you need to approach it from a different angle. Every cloud has a silver lining."

4 - None of the sentences or sub-analogies are disconnected:

Target concept: The internet, Analogy: "The internet is like a vast ocean of information. Websites are islands, each with their own unique landscape and inhabitants. Search engines are the ships that navigate these waters, helping users chart a course to their desired destination. Social media platforms are bustling ports where people from all over this digital world gather to exchange ideas and experiences."

Repetition:

Whether the same sentence is repeated or same source concept is repeated for another target concept within the analogy

1 - All the sentences or source concepts are repeated

2 - Most of the sentences or source concepts are repeated

3 - Some of the sentences or source concepts are repeated

4 - None of the sentences or source concepts are repeated

Examples

1 - All the sentences or source concepts are

repeated:
Target: The Atom, Analogy: "The atom is like the solar system. The nucleus is like the solar system. Electrons are like the solar system. Protons are like the solar system. Neutrons are like the solar system."
2 - Most of the sentences or source concepts are repeated:
Target: The Human Body, Analogy: "The human body is like a machine. The brain is like a machine. The heart is like a pump. The lungs are like bellows. The digestive system is like a machine."
3 - Some of the sentences or source concepts are repeated:
Target: The Solar System Analogy: "The Solar System is like a family. The Sun is like a parent. Planets are like children. Moons are like children. Asteroids are like extended family members. Comets are like distant relatives."
4 - None of the sentences or source concepts are repeated:
Target: Cell Structure, Analogy: "A cell is like a city. The nucleus is like the city hall containing DNA blueprints. Mitochondria are like power plants generating energy. The cell membrane is like the city walls controlling what enters and exits. Ribosomes are like factories producing proteins."
Target Accuracy:
Truthfulness of all facts pertaining to target concept.
N/A - Target missing
1 - None of the facts stated about the target are accurate
2 - Some of the facts stated about the target are accurate
3 - Most of the facts stated about the target are accurate
4 - All of the facts stated about the target are accurate
Examples
N/A - Target missing:
Target: Photosynthesis, Analogy: "A refrigerator keeps food cold to prevent spoilage."
Analogy is not about photosynthesis
1 - None of the facts stated about the target are accurate: Target: Photosynthesis, Analogy: "Photosynthesis is like a furnace burning wood to generate heat and ash."
This analogy is completely inaccurate about the energy conversion and processes involved in photosynthesis.
2 - Some of the facts stated about the target are accurate:
Target: Photosynthesis, Analogy: "Photosynthesis is like a factory where plants produce packaged goods by absorbing water and heat from the soil."
Plants produce energy, not packaged goods. While plants do absorb water and use energy, the source of energy is sunlight, not heat from the soil.
3 - Most of the facts stated about the target are accurate:
Target: Photosynthesis, Analogy: "Photosynthesis is like a solar-powered factory. The leaves act as solar panels, capturing sunlight energy. The process occurs in special organelles called mitochondria, and the green pigment responsible for absorbing light is called chlorophyll."
There is one significant inaccuracy: the process occurs in chloroplasts, not mitochondria.
4 - All of the facts stated about the target are accurate:
Target: Photosynthesis, Analogy: "Photosynthesis is like a solar-powered factory. Plants use sunlight energy to convert carbon dioxide and water into glucose and oxygen. This process takes place in chloroplasts, where the green pigment chlorophyll absorbs sunlight to drive the chemical reactions."
This analogy accurately describes the inputs, outputs, energy source, and location of the photosynthesis process.
Source Accuracy:
Truthfulness of all facts pertaining to source concept.
N/A - Source missing
1 - None of the facts stated about the source are accurate
2 - Some of the facts stated about the source are accurate
3 - Most of the facts stated about the source are accurate
4 - All of the facts stated about the source are accurate
Examples
N/A - Source missing:
Target: Lightning, Analogy: "Lightning is like a big spark."
Lighting is an example of a big spark, they are not different concepts.
1 - None of the facts stated about the source are accurate:
"The solar system is like a beehive, where the queen bee (the Sun) stays stationary in the center while worker bees (planets) fly in concentric circular paths around her at the same speed."

This analogy contains no accurate facts about beehives. Queen bees don't stay stationary in the center, worker bees don't fly in concentric circles around the queen, and they certainly don't all move at the same speed.
2 - Some of the facts stated about the source are accurate:
Target: Solar system, Analogy: "The solar system is like a classroom, where the teacher (the Sun) stands at the front, and students (planets) sit in rows, getting colder as they sit further back. Each student spins in their chair while moving around the classroom."
Some facts are accurate: teachers often stand at the front, and students do sit in rows. However, students don't typically spin in their chairs or move around the classroom, and the temperature doesn't necessarily decrease as you move further back.
3 - Most of the facts stated about the source are accurate:
Target: Solar system, Analogy: "The solar system is like a playground merry-go-round, where the center pole (the Sun) remains fixed while children (planets) spin around it. The kids closer to the center (inner planets) complete their revolutions faster than those at the edge (outer planets). Some children have backpacks (moons) attached to them."
All facts about the merry-go-round are correct except that kids closer to the center do not complete revolutions faster. All riders complete one revolution in the same amount of time, regardless of their position.
4 - All of the facts stated about the source are accurate:
Target: Solar system, Analogy: "The solar system is like a clock, with the central point (Sun) remaining stationary while the hands (planets) move around it at different speeds. Each hand (planet) follows a predictable path, completing full revolutions in varying amounts of time."
This analogy uses entirely accurate facts about the clock.

A.2 Prompt Templates for Pre-pilot Study

```
You will be given one piece of text written to explain a target concept.

Your task is to rate the text on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Meaningful analogy (1 or 0) - Whether the given text is a meaningful analogy or not. Some examples of text that is not a meaningful analogy include the following cases:
The text is not actually an analogy. It could be a definition, example, tautology, etc.
The text contains little to no relevant information pertaining to the target concept.
Important details about the analogous concepts are either incorrect or missing, or the provided explanation was insufficient, making the analogy completely wrong or weak at best.
The text is completely incoherent or gramatically incorrect.

Evaluation Steps:

1. Read the given text carefully.
2. Assign a 0 or 1 score for the meaningful analogy criteria.

Examples:

Text: Cytoplasm is like a school secretary with the difference that cytoplasm is in a liquid form and school secretary is in a dry form.

Evaluation Form:

- Meaningful analogy: 0

Text: Macrophages are similar to guards in that they are both responsible for protecting the body from harm. Macrophages are the first line of defense against infection, while guards are responsible for protecting people and property.

Evaluation Form:

- Meaningful analogy: 1

==========
Target: '{{Target}}'

Text:

{{Document}}


Evaluation Form:

- Meaningful analogy:
```

Figure 1: Prompt template used for BAM

```
You will be given one piece of text written to explain a target concept.

Your task is to rate the text on two metrics.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Meaningful analogy (1-4) - Whether the given text is a meaningful (i.e., valid and correct) analogy, where,
1 means Strongly Disagree that text contains meaningful analogy,
2 means Somewhat Disagree that text contains meaningful analogy,
3 means Somewhat Agree that text contains meaningful analogy,
4 means Strongly Agree that text contains meaningful analogy.
Some examples of text that is not a meaningful analogy include the following cases:
The text is not actually an analogy. It could be a definition, example, tautology, etc.
The text contains little to no relevant information pertaining to the target concept.
Important details about the analogous concepts are either incorrect or missing, or the provided explanation was insufficient, making the analogy completely wrong or weak at best.
The text is completely incoherent or gramatically incorrect.

Novelty (1-4) - How novel is the text, i.e., can similar text be found online?
1 means the same text (potentially paraphrased) is found on the web,
2 means similar text is found on the web,
3 means no similar text is found online but text is straightforward to infer from the content found online,
4 means no remotely similar text is found online and text is not straightforward to infer from the content found online.


Evaluation Steps:

1. Read the given text carefully.
2. Assign a score on a scale of 1-4 for the meaningful analogy criteria.
3. Assign a score on a scale of 1-4 for the novelty criteria.

Examples:

 Text: DNA replication can be thought of as a photocopier. The DNA molecule is like the original document, and each strand of DNA is like one copy of the document. During replication, the two strands are separated, and new copies of each strand are created.

 Evaluation Form:

  - Meaningful analogy: 4
  - Novelty: 1

 Text: Breathing mechanism of frogs can be analogy to bellows of blacksmith. Just like bellows, the frog's lungs are inflated and deflated by muscles that run along either side of its ribcage. When the frog inhales, the muscles contract, pushing air into the lungs. When it exhales, the muscles relax and air is forced out.

 Evaluation Form:

  - Meaningful analogy: 4
  - Novelty: 4

 Text: In computing, an operating system kernel is the core of a computer operating system. It is responsible for managing hardware and software resources and providing common services for application programs. The kernel performs its tasks in cooperation with device drivers, which are modules that load into the kernel to provide specific functions, such as access to the disk drive or network card.

 Evaluation Form:

  - Meaningful analogy: 1
  - Novelty: 1

=========================
Target: '{{Target}}'

Text:

 {{Document}}


Evaluation Form:
```

Figure 2: Prompt template used for OAMN

You will be given one analogy written to explain a target concept.

Your task is to rate the analogy on four metrics.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Target Accuracy (1-4) - The accuracy of facts about the target concept. Penalize factually incorrect text about the target concept.

Source Accuracy (1-4) - The accuracy of facts about the source concept. Penalize factually incorrect text about the source concept. If a separate source concept is not found (e.g., source concept is missing or the target concept is compared to itself), set this score to -1.

Mapping Consistency (1-4) - Structural consistency of the mapping between source and target concepts. Penalize if the source concepts of the sub-analogies are disconneted (i.e., do not coherently consitute a single concept). Also, penalize if 1:1 mapping is not found in the sub-analogies (i.e., if the same source or target concept is used in multiple sub-analogies).

Usefulness (1-4) - The usefulness of the analogy for explaining the concept.

Evaluation Steps:

1. Read the analogy carefully and identify all the sub-analogies.
2. Read each sub-analogy and identify the target and source concept (the concept being compared to the target).
3. For each sub-analogy, write it and assign a score for its target accuracy on a scale of 1 to 4, where 1 is the lowest and 4 is the highest based on the Evaluation Criteria.
4. For each sub-analogy, write it and assign a score for its source accuracy on a scale of 1 to 4, where 1 is the lowest and 4 is the highest, or set it to -1 based on the Evaluation Criteria .
5. Assign a score for the overall mapping consistency on a scale of 1 to 4, where 1 is the lowest and 4 is the highest as per the Evaluation Criteria.
6. Assign a score for the overall usefulness on a scale of 1 to 4, where 1 is the lowest and 4 is the highest as per the Evaluation Criteria.

Example:

Analogy Text:
The atmosphere is like a hug because it is warm and comforting. The thermosphere is like the top of a mountain because it is the highest point. The mesosphere is like the middle of a journey because it is the middle point. The troposphere is like the bottom of the ocean because it is the lowest point.

Evaluation Form:

- Sub-analogy 1: The atmosphere is like a hug because it is warm and comforting.
  - Source Accuracy: 4
  - Target Accuracy: 2
- Sub-analogy 2: The thermosphere is like the top of a mountain because it is the highest point.
  - Source Accuracy: 4
  - Target Accuracy: 1
- Sub-analogy 3: The mesosphere is like the middle of a journey because it is the middle point.
  - Source Accuracy: 4
  - Target Accuracy: 4
- Sub-analogy 4: The troposphere is like the bottom of the ocean because it is the lowest point.
  - Source Accuracy: 4
  - Target Accuracy: 4
- Mapping Consistency: 2
- Usefulness: 3

==========

Target: '{{Target}}'

Analogy Text:

{{Document}}

Evaluation Form:

- Sub-analogy 1:

Figure 3: Best performing prompt template for structural consistency on MANAED

You will be given one analogy written to explain a target concept.

Your task is to rate the analogy on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Usefulness (1-4) - The usefulness of the analogy for explaining the concept.

Evaluation Steps:

1. Read the analogy carefully.
2. Assign a score for the overall usefulness on a scale of 1 to 4, where 1 is the lowest and 4 is the highest as per the Evaluation Criteria.

Example:

Analogy Text:
The atmosphere is like a hug because it is warm and comforting. The thermosphere is like the top of a mountain because it is the highest point. The mesosphere is like the middle of a journey because it is the middle point. The troposphere is like the bottom of the ocean because it is the lowest point.

Evaluation Form:

- Usefulness: 3

==========

Analogy Text:

{{Document}}

Evaluation Form:

- Usefulness:

Figure 4: Best performing prompt template for usefulness on MANAED

You will be given one analogy written to explain a target concept.

Your task is to rate the analogy on four metrics.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Source Accuracy {-1, 1-4} - The accuracy of facts about the source concept. Penalize factually incorrect text about the source concept. If a separate source concept is not found (e.g., source concept is missing or the target concept is compared to itself), set this score to -1.

Target Accuracy (1-4) - The accuracy of facts about the target concept. Penalize factually incorrect text about the target concept.

Evaluation Steps:

1. Read the analogy carefully.
2. Identify all facts related to the source concept (the concept being compared to the target).
3. Assign a score for its source accuracy on a scale of 1 to 4, where 1 is the lowest and 4 is the highest, or set it to -1 based on the Evaluation Criteria.
4. Read each sub-analogy and identify all facts related to the target concept.
5. Assign a score for the target accuracy on a scale of 1 to 4, where 1 is the lowest and 4 is the highest.

Examples:

Analogy Text: The atmosphere is like a blanket because it surrounds and protects us.

Evaluation Form:
- Source Accuracy (blanket): 4
- Target Accuracy (atmosphere): 4

Analogy Text: System software is like the sugar for a cake because it helps to sweeten the final product.

Evaluation Form:
- Source Accuracy (sugar): 4
- Target Accuracy (system software): 1

Analogy Text: The moons are the cousins because they orbit the planets and are much smaller than the planets.

Evaluation Form:
- Source Accuracy (cousins): 1
- Target Accuracy (moons): 4

==========

Target: '{{Target}}'

Analogy Text:

{{Document}}

Evaluation Form:

Figure 5: Best performing prompt template for source and target accuracy on MANAED