

Beyond Accuracy: Improving LLMs' Science Communication Skills

Mattan Yeroushalmi^{1,*}, Maxim Bragilovski^{2,*}, and Nir Grinberg²

¹Open University of Israel ²Ben-Gurion University of the Negev
yematta@post.openu.ac.il maximbr@post.bgu.ac.il nirgrn@bgu.ac.il

Abstract

Science as a body of knowledge is important for human learning, yet beyond factual accuracy, little research has focused on the capacity of Large Language Models (LLMs) to communicate science effectively to the public. Here, we develop and validate a computational framework for measuring science communication skills of LLMs. Using this framework, we evaluate different alignment approaches: fine-tuning using top-ranked human explanations, distilling synthetic explanations from a teacher model, and preference optimization. Our results show that annotators overwhelmingly prefer LLM-generated science explanations over top-ranked organic answers, with this preference primarily attributed to better structure and readability. We find that distillation with synthetic answers achieves the most significant and stable improvement, allowing a smaller open-source model to achieve statistical parity with the proprietary teacher model in blind side-by-side comparisons.

1 Introduction

Our collective ability to benefit from accumulated knowledge depends on the capacity to comprehend it. Large Language Models (LLMs) have rapidly become central to information seeking (Caramanion, 2024), often serving as the final stop in this process. For example, following the introduction of AI-generated answers in Google Search, there has been a significant drop in traffic to original content providers (Chapekis and Lieb, 2025). Of course, a prerequisite for information to be useful is accuracy, and indeed, a considerable body of research focuses on making LLMs more factual and hallucination-free (Sahoo et al., 2024). However, as decades of science communication research show, factuality alone does not guarantee comprehension (McNamara et al., 1996; Bullock et al.,

Example question: “What is CRISPR?”

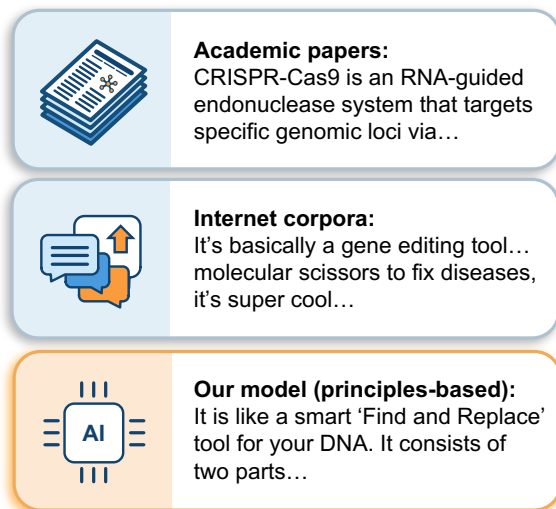


Figure 1: Illustrative example of a scientific question and available answers. Academic texts (top) are precise but dense; top-ranked human explanations (middle) may lack depth. Our model (bottom) aims for accessibility and structure among other science communication principles deemed important for learning.

2019) or accessibility across people of different backgrounds (Scharrer et al., 2021).

Despite significant strides in improving factuality and reasoning capabilities of LLMs, relatively little research has focused on the communicative structure of LLM-generated scientific explanations. While current models are pre-trained on vast amounts of scientific text, from academic articles to Internet answers, these sources do not inherently produce effective science communicators (see Figure 1 for an illustrative example). Alignment techniques, such as Reinforcement Learning from Human Feedback (RLHF; Ouyang et al. (2022b)), optimize for broad notions of helpfulness rather than specific pedagogical objectives. As a result, critical traits for science communication, such as having a clear logical flow, using analogies,

*These authors contributed equally.

and avoiding jargon, emerge only implicitly, if at all. This leaves a gap in LLM training to produce scientific explanations that are not just accurate or preferred by people, but also adhere to science communication best practices, which are ultimately important for human learning. A key question is what supervision signals and training procedures will be most effective at enabling LLMs to acquire science communication skills.

Improving LLMs’ capacity to explain science is crucial for enabling society to “stand on the shoulders of giants”, especially as more people turn to LLMs for answers. A critical challenge in this era of information abundance is no longer mere access to information, but the delivery of clear, accurate, and pedagogically effective explanations (Kozyreva et al., 2020). Unlike traditional search engines that retrieve information and rely on the user to process it, LLMs synthesize and distill information into complete answers (Ai et al., 2023). Moreover, LLMs are uniquely positioned to persuade people to correct their misbeliefs (Costello et al., 2024). This fundamental shift transfers the cognitive burden of synthesis from the user to the model, offering new opportunities to deliver explanations that enhance public understanding of science. In other words, delivering high-quality explanations is important not only for user satisfaction but for facilitating better-informed personal and collective decisions, and advancing science learning across the population (Su et al., 2015; Scheufele, 2013).

However, training LLMs to deliver quality scientific explanations presents several challenges. First, what constitutes a “good” explanation is a complex and multi-faceted construct (Baram-Tsabari and Lewenstein, 2012; Zemla et al., 2017). For example, an explanation may fully answer the user’s question but provide no insight into the underlying mechanism, or it may provide the information but fail to connect it to everyday life and actionable decision-making. Second, the field lacks reliable training signals. Currently, there are no corpora for training LLMs for effective science communication to the general public. Training on scientific Web forums by and for laypeople (e.g., Reddit’s Explain Like I’m Five) may steer the model towards simpler answers, but not necessarily deeper understanding, as social “upvotes” are noisy proxies for scientific quality. Third, while inference-time prompting can potentially close some of these gaps, it suffers from instability (Camburu et al., 2020), in-

struction drift (Liu et al., 2024), and a fundamental inability to shift model priors (Min et al., 2022), all while incurring higher costs and latency. Finally, there is currently no standard computational framework for evaluating the science communication skills of LLMs, and it is unclear how well expert-curated principles for science communication align with the preferences of non-experts.

In this study, we bridge the gap between theoretical science communication and natural language alignment by examining methodologies to instill these skills into open-source LLMs. We contrast three distinct alignment approaches: fine-tuning using top-ranked human explanations, distilling synthetic explanations from a teacher model, and further refinement via Direct Preference Optimization (DPO). To evaluate these approaches, we operationalize the Baram-Tsabari and Lewenstein (2012) framework into a validated computational evaluation suite, conduct blind side-by-side human preference evaluation, and verify that our alignment techniques do not compromise underlying factual accuracy. We also examine the alignment between expert-curated principles and human preferences. Finally, we benchmark our models against both inference-time prompting with the teacher model and human-authored baselines to isolate the most effective technique for instilling science communication skills in LLMs.

Therefore, our contributions are the following:

- A theory-grounded and validated computational framework for measuring science communication skills of LLMs.
- A systematic evaluation of alignment techniques for improving science communication skills in open-source LLMs.
- Empirical evidence challenging the reliance on organic web data: synthetic distillation significantly outperforms fine-tuning on top-ranked human answers, enabling a smaller open-source model to effectively internalize science communication skills into model weights and achieve statistical parity with the teacher model without sacrificing on factual accuracy.
- An analysis of the alignment between expert-curated science communication principles and the actual preferences of active science learners (evaluated among STEM undergraduates), revealing a significant preference for structural scaffolding over other rhetorical devices.

2 Related Work

Current literature on LLM scientific explanations can be organized into three lines of work: (i) science communication literature defining qualities of effective explanations, (ii) evaluation frameworks that measure explanation quality beyond factual accuracy, and (iii) NLP methods that improve scientific QA via training strategies.

Science Communication Literature: Prior work in science education and communication has examined what makes scientific explanations effective for the general public, repeatedly emphasizing clarity, coherent structure, and accessibility (McNeill and Krajcik, 2008; De Andrade et al., 2019). Although there is no single universally accepted definition of a “good” explanation, several dimensions recur across the literature. Baram-Tsabari and Lewenstein (2012) highlight clarity, structure, and stylistic features such as humor and narrative, as well as explanatory devices such as analogy and metaphor, while Rakedzon and Baram-Tsabari (2017) extend this perspective with a rubric emphasizing coherence, rhetorical devices, and accessibility. Evidence also suggests that structured explanations and simplified language improve comprehension (McNeill et al., 2006), and gradually dividing deeper into the underlying mechanism supports deeper understanding, motivating the role of scaffolding in effective explanations (Lombrozo, 2006).

Analogies and metaphors further support the development of a mental model of the scientific concept (Gentner, 1983; Niebert et al., 2012), while jargon reduces understanding and perceived credibility among non-experts (Bullock et al., 2019; Shulman et al., 2020). Prior work also showed that narrative elements and humor can increase engagement with scientific content, supporting their use as rhetorical devices in explanations (Dahlstrom, 2014). From the Baram-Tsabari and Lewenstein (2012) framework, we focus on six dimensions that can be reliably operationalized by both humans and LLMs (see Section 4.2): three from the *basic* cluster (language and accessibility: *Jargon*, *Readability*; knowledge organization: *Scaffolding*) and three from the *intermediate* cluster (rhetorical devices: *Analogy*, *Metaphor*, *Humor*).

Evaluation frameworks: Most evaluation work on scientific LLM outputs centers on factuality, using benchmarks such as FEVER (Thorne et al.,

2018), FactCC (Kryściński et al., 2020), SummaC (Laban et al., 2022), and domain QA sets like PubMedQA (Jin et al., 2019) and BioASQ (Tsatsaronis et al., 2015). Complementary metrics such as BERTScore (Zhang et al., 2020) and SciBERT-based similarity (Beltagy et al., 2019) capture semantic relevance, but cover communicative and pedagogical quality only weakly. Engagement signals (likes, shares, upvotes) are also weak proxies for quality, as they are shaped by social and algorithmic dynamics rather than rigor or correctness (Avram et al., 2020; Milli et al., 2025; Sîrbu et al., 2019).

Recent frameworks increasingly target communicative and audience-oriented properties, including rubrics for naturalness/conciseness/citation quality (Hashemi et al., 2024), hierarchical explanation schemes (Galvan-Sosa et al., 2025), and science-focused evaluators of readability and cohesion (D’Souza et al., 2025). In parallel, LLM-as-a-judge methods score rhetorical and creative attributes such as humor and metaphor (Goes et al., 2022; DiStefano et al., 2025; Olson et al., 2024). We extend this direction by developing additional LLM-based metrics that capture science-communication skills of LLMs based on science communication literature (Baram-Tsabari and Lewenstein, 2012).

Improving scientific QA via training strategies:

Scientific QA has progressed from domain-specific pretraining to alignment and reasoning-oriented methods. Domain-adapted models such as SciBERT (Beltagy et al., 2019) and later instruction-tuned LLMs (Singhal et al., 2023) show strong performance, while chain-of-thought prompting and its scientific variants (e.g., Sci-CoT, T-SciQ) improve reasoning and transfer (Wei et al., 2022; Ma et al., 2023; Wang et al., 2024). Beyond prompting, SFT and RLHF remain standard alignment methods (Ouyang et al., 2022a), and DPO provides a scalable alternative (Rafailov et al., 2023). In this work, we apply SFT and DPO to train an open-source LLM for scientific explanations and evaluate it with explicit science-communication metrics.

3 Data

Training LLMs to produce high-quality scientific explanations requires suitable training corpora. While organic sources such as Reddit discussions or Wikipedia articles provide authentic examples of science communication, our preliminary exper-

iments revealed that these sources do not consistently exhibit the structural and rhetorical qualities associated with effective science communication (see Section 5.1 for empirical support). This motivated us to use synthetically generated answers that explicitly incorporate established science communication principles.

QA-Pairs Dataset. We began with a HuggingFace dataset¹ containing 131,441 question–answer pairs from *r/askscience*², in compliance with their permitted use for non-commercial academic research. From the training split, we retained only pairs where the answer received at least 50 upvotes and removed questions containing URLs or meta-content. The resulting dataset, **QA-Pairs**, contains 8,901 question–answer pairs for training, 91 pairs for validation, and 91 pairs for testing. For each question, we include three answers: (i) the original human-written answer from Reddit; (ii) a GPT-3.5-generated answer³ (using a fixed instructional prefix; see Appendix A.4); and (iii) a Llama-2-7B-generated answer⁴. The Llama-2-7B answers serve as *rejected* candidates for synthetic preference optimization, paired against GPT-3.5 answers as *chosen*. The choice of Llama-2-7B as the rejected source was deliberate: we hypothesized that using a weaker model would maximize the quality gap, providing a clearer preference signal for DPO without requiring human annotation. Interestingly, as detailed in Appendix A.6, a non-negligible fraction of Llama-2-7B test-set questions (25.3%) had higher rubric scores than GPT-3.5, indicating that these preference labels are noisy, but still a useful proxy.

Pref-Human contains 1,445 human preference judgments collected in the first round of Experiment B (Section 5.2), where annotators compared model-generated answers in pairwise side-by-side comparisons.

4 Methods

In this section, we describe the different approaches used for model training and the computation measures developed and validated for quantifying science communication skills of LLMs.

¹<https://huggingface.co/datasets/dhmeltzer/ask-science-qg>

²<https://www.reddit.com/r/askscience>

³`gpt-3.5-turbo-0125`: <https://platform.openai.com/docs/models/gpt-3.5-turbo>

⁴`meta-llama/Llama-2-7b-chat-hf`: <https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

⁵<https://huggingface.co/mattany/datasets>

Dataset	Type	Train	Source
QA-Pairs	SFT/DPO	8,901	Human, GPT-3.5, and Llama-2 answers
Pref-Human	DPO	1,445	Human preferences

Table 1: Training datasets (train split sizes shown; 91 held out each for validation and test). QA-Pairs provides three answers per question: human-written (Reddit), GPT-3.5-generated (using a fixed instructional prompt; see Appendix A.4), and Llama-2-generated; the latter two form synthetic preference pairs. Pref-Human contains human preference judgments. Both datasets are publicly available⁵.

4.1 Model Training

We trained all models⁶ using a similar set of standard hyperparameters, selected empirically on the validation set, and constrained by available GPU memory, as described below.

Human SFT. We fine-tuned Llama-3.1-8B⁷, an Unsloth 4-bit quantized version of Meta’s Llama model⁸, on the human-written answers from QA-Pairs using QLoRA adapters (rank $r=32$, $\alpha=16$). Training used TRL’s *SFTTrainer* with cross-entropy loss for 3 epochs (batch size 16, learning rate 2×10^{-5} , 500 warmup steps).

Synthetic SFT. We fine-tuned the same base model on the GPT-3.5-generated answers from QA-Pairs with identical hyperparameters.

Synth-SFT 70B. To test whether our approach scales to larger models, we applied the same synthetic SFT procedure to Llama-3.3-70B⁹, an Unsloth 4-bit quantized version of Meta’s Llama model¹⁰. Training used matching LoRA and optimizer settings but 2 epochs (vs. 3 for the 8B model) due to faster convergence.

Human DPO. Starting from the Human SFT checkpoint, we applied DPO using the human preference dataset (Pref-Human). Training used LoRA adapters (rank $r=8$, $\alpha=16$, dropout 0.05) with sigmoid DPO loss ($\beta=0.1$) for 2 epochs (batch size 4, gradient accumulation 8, learning rate 1×10^{-5} , cosine schedule).

⁶Models and datasets are available at <https://huggingface.co/mattany>; code is available at <https://github.com/mattany/beyond-accuracy>.

⁷`unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit`: <https://huggingface.co/unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit>

⁸<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

⁹`unsloth/Llama-3.3-70B-Instruct-bnb-4bit`: <https://huggingface.co/unsloth/Llama-3.3-70B-Instruct-bnb-4bit>

¹⁰<https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct-bnb-4bit>

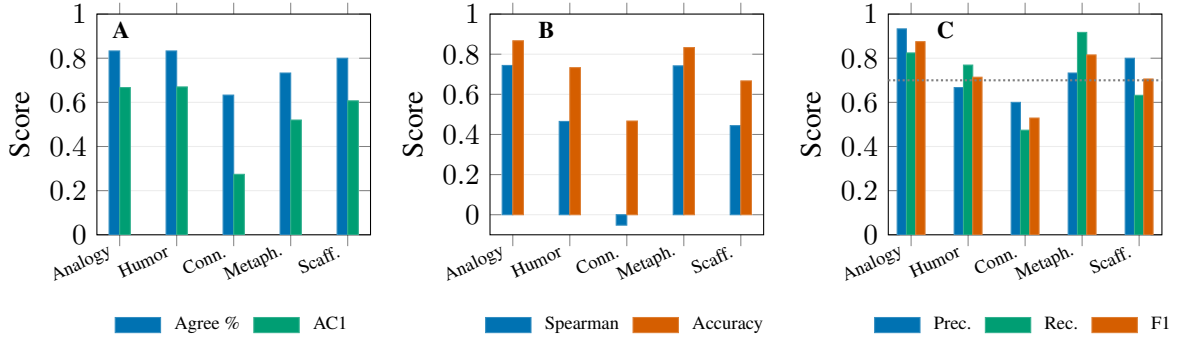


Figure 2: LLM judge validation on 30 questions with 3 annotators. (a) Inter-annotator agreement measured as agreement percentage and Gwet AC1. (b) Spearman and accuracy measured against the human majority. (c) Precision, recall, and F1 measured against the human majority.

Synthetic DPO. Starting from the Synthetic SFT checkpoint, we applied DPO using the synthetic preference pairs from QA-Pairs (GPT-3.5 as *chosen*, Llama-2-7B as *rejected*). Training used TRL’s *DPOTrainer* for 2 epochs (batch size 4, gradient accumulation 8, learning rate 1×10^{-5} , cosine schedule).

4.2 Model Evaluation

Evaluation Rubrics. We implemented evaluation in DeepEval (Confident AI, 2025), combining algorithmic metrics with G-Eval scoring (Liu et al., 2023). Because these presence-oriented rubric scores capture presence rather than usage quality, we complement them with a blinded human-preference study (§5.2) and a regression linking rubric dimensions to preference (§5.2.1). Following Baram-Tsabari and Lewenstein (2012), we use six dimensions in two clusters: *basic accessibility/structure* (readability, jargon, scaffolding) and *intermediate rhetorical devices* (analogy, metaphor, humor). We score scaffolding and the rhetorical-device dimensions (analogy, metaphor, humor) with explicit G-Eval prompts (Appendix A.5), while readability and jargon use algorithmic metrics. *Analogy* is mapped to $[0, 1]$ (absence/presence), and for agreement with binary human labels we binarize LLM scores at 0.5 (chosen empirically because scores cluster near 0 or 1). Readability aggregates normalized Flesch Reading Ease, Flesch–Kincaid, ARI, and Dale–Chall. Jargon uses our Python refactoring of the Automatic Jargon Identifier (Rakedzon et al., 2017); for aggregation we clip to 0.65–1.0 (an empirical floor just below the lowest observed score, since the metric

is compressed near the top of its range) and map linearly to $[0, 1]$.

Aggregation. After each dimension is normalized to $[0, 1]$ as above, the overall rubric score is their weighted average, $\sum_i w_i s_i$ with $\sum_i w_i = 1$. The six dimensions receive equal weight $1/6$; readability’s $1/6$ is split evenly across its four sub-metrics ($1/24$ each). This weighted score is the x-axis reported in Figure 3.

Judge Validation. Two of the authors annotated 30 model-generated answers for each rhetorical device; disagreements were resolved by a third author acting as a tie-breaker. We measured primary-annotator reliability using Gwet’s AC1 (Figure 2a). For human–LLM agreement, we computed Spearman correlation between continuous LLM scores and the binary human majority label, plus accuracy, precision, recall, and F1 after binarizing LLM scores at a 0.5 threshold (Figures 2b–2c).

Inter-annotator reliability (Figure 2a) was higher for analogy and humor ($AC1 \approx 0.67$) than for scaffolding and metaphor ($AC1 \approx 0.61$ and 0.52 , respectively). Human–LLM correlation (Figure 2b) was strongest for analogy and metaphor (Spearman ≈ 0.74), but weaker for humor and scaffolding (Spearman ≈ 0.45). Classification metrics (Figure 2c) reveal that humor shows lower precision (0.67), indicating the LLM over-predicts, while scaffolding shows lower recall (0.63), indicating the LLM under-predicts. The remaining metrics achieve $F1 \geq 0.70$. Because identifying rhetorical devices is inherently subjective, perfect inter-rater agreement is unrealistic even for human annotators; nonetheless the judge tracks the human majority well across dimensions ($F1 \geq 0.70$ apart from humor), and we further corroborate the rubric with

the blind human-preference study (§5.2).

It is important to note the scope of this validation: the rubric is validated as a measure of device *presence* (a graded $[0, 1]$ score, binarized only for this agreement analysis), not of whether a device is used well or appropriately, so a higher rubric score is not in and of itself evidence of better communication. We use it accordingly, as a scalable, interpretable instrument for measuring the prevalence of each device across many models (§5.1); whether those measured features actually relate to communication quality is a separate question we address directly through human preference (§5.2.1). We also evaluated *Connection*, capturing whether the explanation links the scientific concept to everyday experience; however, it exhibited lower reliability ($AC1 \approx 0.27$) and near-zero human–LLM correlation (Spearman ≈ -0.05), and was therefore excluded.

5 Results

In this section, we report on two sets of results. First, we present the performance of different models on the validated science communication rubrics (§5.1). Then, we report results from the analysis of human preferences (§5.2), including an analysis of the alignment between rubric adherence and annotator preferences (§5.2.1).

5.1 Rubric-Based Evaluation

We compare training data sources (synthetic vs. human), alignment methods (SFT vs. DPO), and inference modes (prompted vs. unprompted) to determine which combinations best align with our rubric metrics, and whether these gains generalize across model scales without degrading factual accuracy. We evaluate on the held-out test set, comparing the teacher model (GPT-3.5), the base Llama-3.1-8B, and our four fine-tuned variants (Synthetic/Human SFT and DPO), each in both prompted and unprompted modes.

Results. Figure 3 shows rubric scores for all evaluated models, ranging from GPT-5 (9.5%) to the prompted Synthetic DPO (57.3%). A clear finding from the figure is that inference-time prompting makes the use of rhetorical devices such as scaffolding, analogies, and metaphors significantly more common than in organic human answers (with a weighted score of 21.8%). Conversely, unprompted base models, including recent frontier models, utilize these communication devices less frequently than humans.

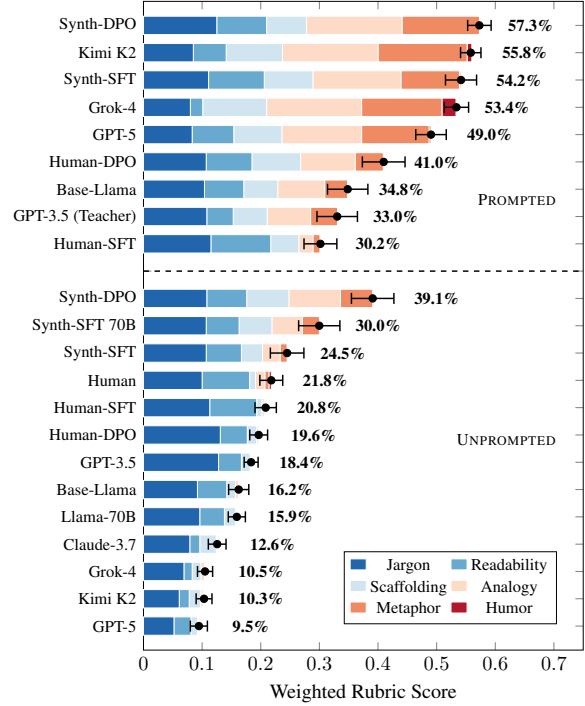


Figure 3: Science communication rubric scores across models.

When examining training data sources, fine-tuning using top-ranked human answers (Human-SFT) fails to significantly increase rubric adherence above the raw human baseline (20.8% vs. 21.8%). In contrast, training on synthetic answers successfully increases adherence: unprompted Synthetic SFT (24.5%) demonstrates a significant improvement over the human baseline, approaching the prompted Teacher’s score (33.0%), while unprompted Synthetic DPO (39.1%) actually exceeds it (all with $p < 0.01$).

Interestingly, our results demonstrate that a small 8B model can effectively internalize science communication skills. The use of rhetorical devices in our synthetically trained, unprompted models is more aligned with organic human answers than the larger, unprompted frontier models, including Llama-70B (15.9%), Claude-3.7 (12.6%), Grok-4 (10.5%), Kimi K2 (10.3%), and GPT-5 (9.5%). Furthermore, when prompted, our 8B synthetic models (54.2% to 57.3%) perform on par with these much larger prompted frontier models (49.0% to 55.8%).

Importantly, training did not degrade factuality. Tested in three separate ways (see Appendix A.3), we find no significant degradation in factuality between our trained models and the base models.

Next, we extend beyond rubric adherence and test the models against human preferences.

5.2 Human Preference Evaluation

We conducted a manual human evaluation of explanation preferences using blind side-by-side comparison task.

Subjects. Participants ($N = 81$) were recruited from a large undergraduate course via an institutional mailing list. Participation was voluntary, with up to three bonus points contingent on passing attention checks. The protocol was approved by an Ethics Committee at Ben-Gurion University (approval number CIS2025_001.1). The study was run on Qualtrics; participants provided informed consent, could skip questions or withdraw at any time, and all data were anonymized before analysis. Participants had a mean age of 24.79 years ($SD = 1.90$, ranging from 20 to 28), comprising 53.1% female, 45.3% male, and 1.6% identifying as other or preferring not to say. While this sample captures the preferences of relatively educated young adults in STEM – a highly relevant subpopulation of active science learners – we make no claims about generalization to other subgroups or the general public.

Task. Participants completed two sessions on different days. In each session, participants evaluated 25 scientific questions through side-by-side comparisons of two explanations, blinded to their origin, choosing the preferred explanation or *I do not know* when a clear preference could not be determined. Each session included 3 author-written attention checks with an obviously preferable answer. We excluded sessions that failed any attention check or were completed in under 20 minutes, given that reading alone takes approximately 15 minutes. Answer order, question order, and left-right positioning were randomized and counterbalanced.

Model comparisons. Due to the high number of possible pairwise comparisons, we focused our analysis on two questions of primary interest: (1) do humans prefer AI-generated explanations over human-written answers, and (2) which training method (SFT vs. DPO, synthetic vs. human data) yields preferred outputs? We omitted Human-DPO because of the limited resources and because the effect of DPO was already tested via the Synthetic SFT vs. Synthetic DPO comparison.

Results. Table 2 summarizes the results of our blind pairwise comparison of human preferences,

Category	Model A	Model B	Win %	Result
Human-AI	Teacher [†]	Human answers	72.5% (166/229)	WIN ***
Intern-alization	Synthetic SFT	Teacher [†]	50.2% (139/277)	TIE ns
	Synthetic SFT	Base [†]	59.0% (102/173)	WIN *
Pref. opt.	Synthetic SFT [†]	Synthetic DPO [†]	63.4% (203/320)	WIN ***
	Teacher	Synthetic DPO	56.1% (185/330)	WIN *
Prompting Effects	Synthetic SFT [†]	Teacher [†]	18.3% (51/279)	LOSS ***
	Synthetic SFT [†]	Human SFT [†]	46.1% (83/180)	TIE ns

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 2: Pairwise comparison of human preferences over model outputs. Win rates indicate how often Model A was preferred. Dagger (†) denotes the use of instruction system prompt; unmarked models are unprompted.

reporting the win rate of model A against model B along with statistical significance levels. First, the *Human-AI* comparison shows that annotators overwhelmingly preferred the Teacher model’s explanations over the original human answers (72.5% win rate, $p < 0.001$). While we cannot rule out a causal link between answer length, formatting, or formality and this preference, the results in Section 5.2.1 indicate that annotators’ preferences remain largely unchanged when controlling for these variables or when analyzing AI-generated answers only. We also found that our unprompted Synthetic SFT model *effectively internalizes* science communication skills: it achieves statistical parity with the prompted Teacher (50.2%, $p = 0.952$) and significantly outperforms the prompted base model (Llama-3.1-8B, 59.0%, $p = 0.018$). This confirms that fine-tuning on synthetic data successfully distills the necessary capabilities into the model weights, enabling high-performance inference without the need for complex prompting.

However, we observed that further alignment interventions, specifically preference optimization and inference-time prompting, degraded performance. In *preference optimization*, we found that optimizing for model-derived preferences negatively impacted preferences: the prompted Synthetic SFT model was significantly preferred over the prompted Synthetic DPO model (63.4%, $p < 0.001$). This indicates that the DPO process, while increasing rubric adherence (Figure 3), may overuse rhetorical elements that reduce naturalness. Fi-

nally, the *prompting effects* comparisons show that adding an instructional prompt to the Synthetic SFT model results in a significant loss against the Teacher (18.3% win rate, $p < 0.001$), suggesting a potential “prompt interference” where instructions conflict with the model’s internalized fine-tuning.

5.2.1 Rubric-Preference Alignment

To better understand which features are associated with annotator preferences, we fit a logistic regression predicting human choice from metric score differences, controlling for answer length (log-transformed) and sentence-level formality to address potential verbosity- and style-based confounds. Formality is calculated using Heylighen–Dewaele F-measure (Heylighen and Dewaele, 2002) and rescaled to $[0, 1]$. Table 3 shows the results of this regression.

Two rubrics have significant positive coefficients. *Scaffolding* shows the most reliably estimated effect, the only factor significant at $p < 0.001$ ($\beta = 0.67$; OR=1.94): explanations that build concepts progressively from simple to complex are preferred. *Readability* also predicts preference ($\beta = 0.84$, $p < 0.05$; OR=2.32), indicating that evaluators favor more accessible language over complex prose. Counter-intuitively, *Metaphor* shows a significant negative coefficient ($\beta = -0.58$, $p < 0.01$; OR=0.56): explanations containing metaphors are less preferred. Indeed, when one explanation contains a metaphor and the other does not ($N=325$), evaluators chose the metaphor-free option 60% of the time ($p < 0.001$). However, this effect disappears when excluding Prompted-DPO comparisons ($\beta = -0.27$, n.s.), indicating that the penalty reflects DPO over-optimization of metaphor rather than a general aversion to figurative language.

We also confirmed DPO over-optimization directly: using our validated metaphor judge, we found that DPO systems employ metaphors in 63% of answers versus 29% for all other systems (a $2.2\times$ increase; Welch $t=13.0$, $p < 0.001$), rising to 82% for prompted-DPO. Moreover, in 141 of 300 DPO answers the DPO output scored higher on the aggregate rubric yet lost the human vote, and 71% of these also contained more metaphor, a concrete “rubric-up, preference-down” pattern. For instance, answering *why does stretching feel so good?*, the prompted-DPO opened with “*Imagine your muscles as a group of sleepy giants...*”, which received high rubric score (0.73 vs. 0.21), yet annotators preferred a simpler, literal explanation

	Full (N=800)		No P-DPO (N=600)		No Human (N=700)	
Metaphor	-0.58**	(0.22)	-0.27	(0.26)	-0.61**	(0.23)
Analogy	-0.01	(0.20)	-0.05	(0.22)	-0.15	(0.21)
Scaffolding	.67***	(0.20)	.63**	(0.23)	.64**	(0.21)
Jargon	.93	(0.68)	1.04	(0.76)	.89	(0.73)
Readability	.84*	(0.37)	.83*	(0.41)	1.16**	(0.41)
Ans. length	-0.37*	(0.15)	-0.28	(0.15)	-0.60**	(0.19)
Formality	2.31*	(1.16)	3.27**	(1.26)	.98	(1.35)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Table 3: Logistic regression predicting human preference from metric score differences. All models control for log answer length and sentence-level formality (Heylighen–Dewaele F-measure, rescaled to $[0, 1]$). “No P-DPO” excludes prompted-DPO comparisons; “No Human” restricts the set to model-generated answers only, mitigating formatting differences in the human-authored baseline.

(further examples in Appendix A.1).

The remaining rhetorical dimensions – *analogy*, *jargon*, and *humor* – are not significant predictors of preference in any specification. Overall, this regression enabled us to identify the rhetorical devices associated with human preference.

Notably, annotators actually preferred shorter answers ($\beta = -0.37$, $p < 0.05$), directly countering verbosity-bias concerns. A restricted regression on model-generated answers only (“No Human”, $N=700$), which mitigates formatting differences in the human-authored baseline, shows that most coefficients’ sign, magnitude, and significance remain stable, with two factors intensifying (readability and shorter answer length). Controlling for formality does not attenuate the substantive effects: *scaffolding* ($\beta = 0.63$ – 0.67) and *readability* remain significant. Formality itself is significant only where human answers are present (“Full”, $\beta = 2.31$; “No P-DPO”, $\beta = 3.27$) and non-significant among model-generated answers (“No Human”, $\beta = 0.98$); its large coefficient reflects the narrow spread of formality differences ($SD \approx 0.06$ on the $[0, 1]$ scale), so its realized effect is small. Thus surface style plays a minor role but does not account for the structural and accessibility effects; restricting to the human-vs-AI comparisons, scaffolding remains the dominant significant driver ($\beta = 2.15$, $p < 0.05$).

6 Discussion

Our findings show that annotators prefer model-generated scientific explanations over top-ranked human answers. This highlights a critical deficiency in organic supervision, suggesting that even “high-quality” web answers are not ideal, and that

there is potential to improve LLM-generated science explanations with synthetic data. Further analysis suggested that the performance gap is primarily structural: *scaffolding* emerged as the strongest predictor of human preference, indicating that gradual development of ideas is the most important trait for explanations. Furthermore, the ability of our unprompted SFT model to match the teacher model (Table 2) demonstrates that these “soft” communication skills are not emergent properties of model scale, but capabilities that can be effectively internalized by smaller models through high-quality synthetic supervision. This internalization also removes the specialized inference-time prompt and enables local deployment without API costs; Appendix A.8 provides an illustrative comparison of prompt overhead, throughput, and hardware requirements. The fact that more recent frontier models do not show increased use of science communication skills suggests that there is a persistent gap, which is not addressed by current alignment techniques of frontier models. Importantly, the gains we achieved by aligning open-source models did not come at the expense of factual accuracy, as confirmed in three separate tests (see Appendix A.3).

Both preference optimization (DPO) and inference-time prompting improved rubric scores yet were rated unfavorably by annotators, revealing a trade-off between rubric adherence and human preference driven by the overuse of discourse structures, especially metaphors. Comparing unprompted DPO variants isolates the cause: DPO on our synthetic pairs inflated metaphor use 4.6 times over its unprompted synthetic SFT base (0.07 to 0.33), whereas DPO on genuine human preferences left its unprompted human SFT base unchanged at 0.01. Because the synthetic DPO model uses metaphors more than *both* the chosen (GPT-3.5) and rejected (Llama-2) answers it was trained on, this reflects reward-driven over-optimization rather than imitation of the preference data. Future work could address this in several ways: (i) grounding the preference signal in human judgments rather than a synthetic proxy (human-preference DPO showed no metaphor inflation); (ii) selecting checkpoints by held-out *human preference* rather than rubric adherence (the two diverge under DPO); (iii) drift-limiting objectives such as KL-regularized or IPO training (Azar et al., 2024); and (iv) Process Reward Models (Lightman et al., 2024) that penalize rubric over-use. Finally, the absence of humor in LLM-generated explanations is noteworthy, even

when considering its rare use in top-ranked human results. This aligns with prior work suggesting that the surprise involved in humor is incongruent with LLMs’ objective of minimizing next-token surprise (Hessel et al., 2023). However, the importance of humor as a rhetorical device for effective science communication (Baram-Tsabari and Lewenstein, 2012) should motivate future work on this topic.

7 Conclusion

In this study, we examined methodologies for aligning open-source LLMs with established science communication principles, contrasting the efficacy of organic human supervision against synthetic distillation and preference optimization. We identify synthetic supervision as the most effective approach for this domain: our unprompted SFT model significantly outperformed both the base Llama 3.1 model and top-ranked human answers, while achieving statistical parity with the much larger teacher model. This confirms that science communication skills can be effectively internalized via fine-tuning without the need for complex inference-time prompting.

8 Limitations

Our results should be interpreted in light of the following limitations. Our experiments focus on a single domain, namely English *r/askscience* questions, and therefore may not fully represent the breadth of scientific topics, discourse norms, or reader backgrounds encountered in other settings. Nonetheless, the proposed training pipeline is not domain-specific and can, in principle, be applied to other datasets. A further consideration is our reliance on GPT-3.5 as the teacher, which could imprint teacher-specific stylistic biases onto the fine-tuned models. Repeating the synthetic SFT pipeline with a newer teacher (GPT-5) and with a non-OpenAI teacher (Kimi K2.6) using the same Llama-3.1-8B base model yields strong students in both cases. As detailed in Appendix A.7, GPT-5 matches GPT-3.5 unprompted (24.3% vs. 24.5%, $p=0.25$) with n.s. stylistic dimensions, while Kimi is higher overall (34.1%/59.3%) but with more analogy and metaphor, so teacher identity still shapes style even as the pipeline generalizes beyond GPT-3.5 and beyond OpenAI models. Relatedly, our fine-tuning experiments were confined to the Llama family, trained at two widely sepa-

rated sizes (8B and 70B) rather than a full size-range sweep. While the prompting effect replicates across model families in our evaluation (Figure 3) and our headline distillation is itself cross-family (a GPT-3.5 teacher into a Llama student), cross-family *training* (e.g., Qwen or Mistral students) and a denser size sweep (e.g., 1B, 3B, 13B) remain natural extensions for future work.

Another limitation stems from the sample of people providing annotations. Our human annotators were relatively educated young adults in STEM, not a representative sample of the general population nor representative of school-age learners who may benefit the most from clearer scientific explanations. As such, our claims about human preferences should be understood as reflecting this specific subpopulation of active science learners, and future work should expand beyond this population. We also cannot rule out self-selection of participants into the study. Future work could also explore tailoring explanations to different levels of knowledge or even personalizing them for specific individuals.

Our metrics also impose constraints. The LLM-based rubric metrics show only moderate agreement with human annotators for some dimensions (scaffolding $AC1 \approx 0.61$, metaphor $AC1 \approx 0.52$), and the rhetorical metrics are graded presence scores in $[0, 1]$ (binarized only for human-agreement validation), capturing presence rather than usage quality. Moreover, we measured preferences rather than learning outcomes, and make no claims about the downstream impacts on comprehension or retention. Conducting rigorous behavioral experiments on human learning involves complex cognitive variables that are beyond the scope of the current study. Finally, future work could also further examine the causal link between human preference, answer length, and formatting.

9 Ethical considerations

The study protocol was reviewed and approved by the Human Subjects Research Committee of Ben-Gurion University. Participants were recruited from a large undergraduate course via an institutional course mailing list. Participation was voluntary, participants received up to three bonus points on their final grade as compensation, subject to passing attention checks embedded within two blocks of questions. The study was administered online via Qualtrics. Informed consent was

obtained from all participants, explicitly disclosing that they will be evaluating LLM-generated content, which might contain inaccuracies. Participants performed blind side-by-side comparisons and could skip questions or withdraw at any time. All data were anonymized prior to storage and analysis.

References

- Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, Shen Gao, J. Guo, Xiangan He, Yanyan Lan, Chenliang Li, Yiqun Liu, Ziyu Lyu, Weizhi Ma, Jun Ma, Z. Ren, Pengjie Ren, Zhiqiang Wang, Min Wang, Jirong Wen, Lei Wu, Xin Xin, Jun Xu, Dawei Yin, Peng Zhang, Fan Zhang, Wei na Zhang, M. Zhang, and Xiaofei Zhu. 2023. [Information retrieval meets large language models: A strategic report from chinese ir community](#). *AI Open*, 4:80–90.
- Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer. 2020. [Exposure to social engagement metrics increases vulnerability to misinformation](#). *Misinformation Review*.
- Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.
- Ayelet Baram-Tsabari and Bruce V. Lewenstein. 2012. [An instrument for assessing scientists’ written skills in public communication of science](#). *Science Communication*, 34(1):56–76.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [SciBERT: A pretrained language model for scientific text](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.
- Olivia M Bullock, Dana Colón Amill, Hillary C Shulman, and Graham N Dixon. 2019. [Jargon as a barrier to effective science communication: Evidence from metacognition](#). *Public Understanding of Science*, 28(7):845–853.
- Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. [Make up your mind! adversarial generation of inconsistent natural language explanations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

- Kevin Matthe Caramancion. 2024. Large language models vs. search engines: evaluating user preferences across varied information retrieval scenarios. *arXiv preprint arXiv:2401.05761*.
- Athena Chapekis and Anna Lieb. 2025. [Do people click on links in google ai summaries?](#) Pew Research Center. Published July 22, 2025; accessed September 23, 2025.
- Confident AI. 2025. [Deepeval: The open-source llm evaluation framework](#). <https://deepeval.com/>. Accessed March 1, 2025.
- Thomas H Costello, Gordon Pennycook, and David G Rand. 2024. [Durably reducing conspiracy beliefs through dialogues with ai](#). *Science*, 385(6714):eadq1814.
- Michael F Dahlstrom. 2014. Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences*, 111(Supplement 4):13614–13620.
- Vanessa De Andrade, Sofia Freire, and Mónica Baptista. 2019. [Constructing scientific explanations: A system of analysis for students’ explanations](#). *Research in Science Education*, 49(3):787–807.
- Paul V. DiStefano, John D. Patterson, and Roger E. Beaty. 2025. [Automatic scoring of metaphor creativity with large language models](#). *Creativity Research Journal*, 37(4):555–569.
- Jennifer D’Souza, Hamed Babaei Giglou, and Quentin Münch. 2025. [YESciEval: Robust LLM-as-a-judge for scientific question answering](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13749–13783, Vienna, Austria. Association for Computational Linguistics.
- Diana Galvan-Sosa, Gabrielle Gaudeau, Pride Kavumba, Yunmeng Li, Hongyi Gu, Zheng Yuan, Keisuke Sakaguchi, and Paula Buttery. 2025. [Rubrik’s cube: Testing a new rubric for evaluating explanations on the CUBE dataset](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23800–23839, Vienna, Austria. Association for Computational Linguistics.
- Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.
- Fabricio Goes, Zisen Zhou, Piotr Sawicki, Marek Grzes, and Daniel G Brown. 2022. Crowd score: A method for the evaluation of jokes using large language model ai voters as judges. *arXiv preprint arXiv:2212.11214*.
- Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. [LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.
- Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. [Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.
- Francis Heylighen and Jean-Marc Dewaele. 2002. [Variation in the contextuality of language: An empirical measure](#). *Foundations of Science*, 7(3):293–340.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. [Pubmedqa: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2567–2577. Association for Computational Linguistics.
- Anastasia Kozyreva, Stephan Lewandowsky, and Ralph Hertwig. 2020. [Citizens versus the internet: Confronting digital challenges with cognitive tools](#). *Psychological Science in the Public Interest*, 21(3):103–156.
- Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. [Evaluating the factual consistency of abstractive text summarization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9332–9346. Association for Computational Linguistics.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. [SummaC: Re-visiting NLI-based models for inconsistency detection in summarization](#). *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252. Association for Computational Linguistics.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranajpe, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.

- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-Eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470.
- Yuhan Ma, Chenyou Fan, and Haiqi Jiang. 2023. Sci-cot: Leveraging large language models for enhanced knowledge distillation in small models for scientific qa. In *2023 9th International Conference on Computer and Communications (ICCC)*, pages 2394–2398. IEEE.
- Danielle S. McNamara, Eileen Kintsch, Nancy B. Songer, and Walter Kintsch. 1996. [Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text](#). *Cognition and Instruction*, 14(1):1–43.
- Katherine L. McNeill and Joseph Krajcik. 2008. [Scientific explanations: Characterizing and evaluating the effects of teachers’ instructional practices on student learning](#). *Journal of Research in Science Teaching*, 45(1):53–78.
- Katherine L. McNeill, Daniel J. Lizotte, Joseph S. Krajcik, and Ronald W. Marx. 2006. [Supporting students’ construction of scientific explanations by fading scaffolds in instructional materials](#). *The Journal of the Learning Sciences*, 15(2):153–191.
- Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. 2025. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus*, 4(3):pgaf062.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Kai Niebert, Sabine Marsch, and David F. Treagust. 2012. [Understanding needs embodiment: A theory-guided reanalysis of the role of metaphors and analogies in understanding science](#). *Science Education*, 96(5):849–877.
- Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Shao-yen Tseng, and Vasudev Lal. 2024. Steering large language models to evaluate and amplify creativity. *arXiv preprint arXiv:2412.06060*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022a. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022b. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Tzipora Rakedzon and Ayelet Baram-Tsabari. 2017. [To make a long story short: A rubric for assessing graduate students’ academic and popular science writing skills](#). *Assessing Writing*, 32:28–42.
- Tzipora Rakedzon, Elad Segev, Niv Chapnik, Roni Yosef, and Ayelet Baram-Tsabari. 2017. [Automatic jargon identifier for scientists engaging with the public and science communication educators](#). *PLOS ONE*, 12(8):e0181742.
- Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. [A comprehensive survey of hallucination in large language, image, video and audio foundation models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.
- Lisa Scharrer, Rainer Bromme, and Marc Stadler. 2021. [Information easiness affects non-experts’ evaluation of scientific claims about which they hold prior beliefs](#). *Frontiers in Psychology*, 12:678313.
- Dietram A Scheufele. 2013. [Communicating science in social settings](#). *Proceedings of the National Academy of Sciences*, 110(supplement_3):14040–14047.
- Hillary C Shulman, Graham N Dixon, Olivia M Bullock, and Dana Colón Amill. 2020. [The effects of jargon on processing fluency, self-perception, and scientific engagement](#). *Journal of Language and Social Psychology*, 39(5-6):579–597.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, Said Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather J. Cole-Lewis, Stephen J. Pfohl, P. A. Payne, Martin G. Seneviratne, Paul Gamble, Chris Kelly, Nathanael Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Greg S. Corrado,

Yossi Matias, Katherine Hui-Ling Chou, Juraj Gotweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joëlle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. [Large language models encode clinical knowledge](#). *Nature*, 620(7972):172–180.

Alina Sirbu, Dino Pedreschi, Fosca Giannotti, and János Kertész. 2019. Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. *PloS one*, 14(3):e0213246.

Leona Yi-Fan Su, Heather Akin, Dominique Brossard, Dietram A Scheufele, and Michael A Xenos. 2015. Science news consumption patterns and their implications for public understanding of science. *Journalism & Mass Communication Quarterly*, 92(3):597–616.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Stasinios Petridis, Dimitris Polychronopoulos, et al. 2015. [An overview of the bioasq large-scale biomedical semantic indexing and question answering competition](#). *BMC Bioinformatics*, 16:138.

Lei Wang, Yilang Hu, Jiabang He, Xingdong Xu, Ning Liu, Hui-juan Liu, and Hengtao Shen. 2024. [T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19162–19170.

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

Jeffrey C. Zemla, Steven A. Sloman, Christos Bechli-vanidis, and David A. Lagnado. 2017. [Evaluating everyday explanations](#). *Psychonomic Bulletin & Review*, 24(5):1488–1500.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

A Appendix

A.1 Additional Examples of Metaphor Over-optimization in DPO

Complementing Section 5.2.1, we show full answer pairs from the pairwise human evaluation in which the prompted-DPO answer scored higher on the

aggregate rubric, driven by sustained, often stacked metaphors, yet lost the human vote to a plainer, literal explanation. Of the 141 such “rubric-up, preference-down” DPO cases, 71% had a higher metaphor score than the preferred answer. Here the metaphor score is the G-eval judge’s graded presence rating in $[0, 1]$ (0 = no metaphor, 1 = a clear, active metaphor); the binary presence rate reported in the main text is this score thresholded, so a “higher metaphor score” means the answer was judged to contain more prominent metaphorical content. Rubric scores are reported as DPO vs. preferred (0–1 scale); in every pair below the preferred (winning) answer contained no active metaphor.

Q: “What happens to water when it freezes and can’t expand?” (Rubric: DPO 0.72 vs. preferred 0.28.)

Prompted-DPO (rejected). Imagine water molecules as tiny dancers on a stage. When water freezes, these dancers start to slow down and arrange themselves in a more structured way, forming a solid ice crystal. This change in behavior is due to the decrease in temperature, which gives the dancers less energy to move around freely. Now, here’s the interesting part: as water freezes, it expands. This expansion is like the dancers spreading out their arms and legs, taking up more space on the stage. However, in some cases, like in a sealed container, this expansion is restricted. Picture the dancers trying to spread out but being pushed back by the walls of the container. This leads to a buildup of pressure, which can cause the container to crack or even burst. When water can’t expand when it freezes, it’s like having a group of dancers stuck in a small room with no space to move. This situation is known as “frozen water expanding in a confined space.” In nature, this phenomenon is commonly observed in lakes and rivers during extremely cold temperatures. The ice forms on the surface, but the water beneath remains liquid and expands, exerting pressure on the ice above. This pressure can lead to the formation of cracks and eventually the breaking of the ice.

Preferred (winning). When water freezes, it undergoes a phase transition from a liquid to a solid state due to the low temperature. Normally, as water freezes, it expands because the molecules form a specific arrangement that requires more space than in the liquid state. This expansion is why ice floats in water. However, if water is contained in a completely sealed environment where it cannot expand as it freezes, unique phenomena occur. The pressure in the sealed container increases significantly as the ice forms, exerting force on the container walls. The pressure continues to rise until a new equilibrium is reached where the expansion from freezing is balanced by the increased pressure. This phenomenon is quite fascinating as it showcases the incredible forces at play when water transitions from a liquid to a solid without the ability to expand freely.

Q: “As someone gets more obese, do their skin cells stretch or replicate to keep up with the increasing surface area?” (Rubric: DPO 0.68 vs. preferred 0.21.)

Prompted-DPO (rejected). Imagine your skin as a team of tiny workers, each responsible for maintaining the outer layer of your body. As you gain weight, your skin needs to adapt to the increasing surface area. Now, let’s delve into the fascinating world of skin cells. When you gain weight, your skin doesn’t just stretch like a rubber band; instead, it undergoes a process called hyperplasia. This means that your skin cells start replicating at a faster rate to keep up with the expanding surface area. It’s like a construction crew working overtime to build

new houses to accommodate a growing city. As your skin cells multiply, they work tirelessly to maintain the integrity of your skin. However, this rapid cell division can sometimes lead to issues like skin sagging or stretch marks. Picture your skin as a busy highway with cars zooming by; as more cars (cells) are added, the road (skin) needs to expand to accommodate them. While your skin cells do their best to keep up, sometimes the road gets a bit worn out, leading to visible signs of strain. So, while your skin cells do replicate to match the increasing surface area, the constant battle to keep up with weight gain can sometimes result in noticeable changes to your skin's appearance.

Preferred (winning). As a person becomes more obese, their skin doesn't mainly stretch but the fat cells beneath the skin increase in size. This leads to an increase in overall body volume and surface area that the skin needs to cover. The skin itself, made up of cells tightly packed together, can stretch to accommodate some changes in body size but has limitations. The stretch marks that can occur during rapid weight gain are a result of the skin being pushed beyond its limits without enough time for gradual stretching. Skin cells, actively renewing themselves throughout life, do replicate to some extent to support the growth of the individual. However, the rate at which skin cells replicate is relatively slow compared to the rapid increase in fat volume, so the skin may struggle to keep up. This imbalance can result in skin looking loose or sagging in individuals who undergo significant weight gain. It's crucial to remember that maintaining a healthy weight can help the skin remain in better condition and minimize these effects.

Q: "Can film exist in a format that isn't a series of still frames?" (Rubric: DPO 0.61 vs. preferred 0.17.)

DPO (rejected). Imagine a film reel as a series of still frames strung together like beads on a necklace. Each frame captures a moment in time, creating the illusion of movement when played back in rapid succession. Now, let's venture beyond this traditional concept. In the realm of digital filmmaking, we can think of a film as a continuous flow of data, akin to a video stream. Just as a video player can render a continuous stream of pixels to create the illusion of movement, a film can exist as a dynamic, flowing entity rather than a series of static frames. This digital format allows for smoother transitions and more fluid motion, blurring the lines between individual frames. In essence, while traditional film relies on a series of still images, digital film can be seen as a seamless tapestry of moving images, akin to a flowing river rather than a series of stepping stones. This shift in perspective opens up new possibilities for storytelling and visual representation in the world of film.

Preferred (winning). Film traditionally captures a series of still frames, either on celluloid film or digitally. Each frame is a snapshot of a moment in time. However, there are experimental and art forms that can create moving images without utilizing traditional still frames. Some examples include techniques like light painting, where light is manipulated to create a moving image in a single long exposure, or zoetropes and phenakistoscopes which use spinning objects to create the illusion of motion. These techniques do not involve capturing a series of still frames in the same sense as traditional film, yet they still create the illusion of movement.

Q: "Why does the flame of a cigarette lighter aid visibility in a dark room, but the flame of a blowtorch has no effect?" (Rubric: DPO 0.64 vs. preferred 0.20.)

Prompted-DPO (rejected). Imagine you are in a dark room, and you light a cigarette lighter. The flame from the lighter is like a tiny spotlight shining in the darkness. When light from the flame hits objects in the room, it bounces back to your eyes, making them visible. This bouncing back of light

is called reflection. Now, let's compare this to a blowtorch flame. A blowtorch flame is like a bright spotlight shining in a well-lit room. In a well-lit room, there is already plenty of light bouncing around, so the additional light from the blowtorch doesn't make a significant difference. In contrast, in a dark room, the light from the cigarette lighter is like a beacon that helps illuminate the surroundings, making objects more visible. To understand this better, think of a dark room as a canvas waiting for colors to be painted on it. The flame from the cigarette lighter is like adding a few strokes of paint to the canvas, making it visible. On the other hand, the blowtorch flame is like adding more paint to a canvas that is already fully painted. The extra paint doesn't make a significant difference in the overall picture.

Preferred (winning). To understand this, it's important to delve into the concept of luminous intensity. When you light a cigarette lighter in a dark room, the flame acts as a source of light due to its relatively high luminous intensity, making it glow and aiding visibility in the surrounding darkness. This is akin to how a candle can partially illuminate a dark room. In contrast, the flame of a blowtorch, while being incredibly hot and powerful, has a much lower luminous intensity compared to the cigarette lighter. The blowtorch's flame emits most of its energy in the form of heat and isn't able to produce enough visible light to significantly improve visibility in a dark setting. In essence, the key difference lies in the ability of each flame to emit visible light. The cigarette lighter flame, with its higher luminous intensity, can contribute to lighting up the area around it, hence aiding visibility in a dark room. On the other hand, the blowtorch flame, with its focus on heat production rather than light emission, doesn't have the same effect on visibility despite its ferocious intensity.

Q: "Could a woman deform her breasts if she sleeps on her stomach every night?" (Rubric: DPO 0.66 vs. preferred 0.24.)

Prompted-DPO (rejected). Imagine your breasts as delicate flowers that need gentle care to bloom beautifully. When you sleep on your stomach, it's like placing a heavy book on top of those flowers every night. The constant pressure can cause the tissues around your breasts to stretch and lose their natural shape. Over time, this repeated strain can lead to a deformation, making your breasts appear less perky and more saggy. To understand this better, think of your breasts as a balloon filled with air. When you sleep on your stomach, it's like squeezing that balloon every night, causing it to lose its round shape. The skin and tissues around your breasts are like the balloon material, which can stretch and lose its elasticity due to the constant pressure. So, while sleeping on your stomach might not cause immediate harm, the cumulative effect over months or years can lead to noticeable changes in the appearance of your breasts. In essence, just as a tree needs proper support to grow straight and strong, your breasts need a comfortable sleeping position to maintain their natural shape. So, consider switching to a side-sleeping position or using a supportive pillow to give your breasts the rest they deserve. By doing so, you can help keep those delicate flowers blooming beautifully for years to come.

Preferred (winning). Sleeping on your stomach every night will not deform your breasts. Breasts are composed of glandular and fatty tissue, not muscle or bone, so they are quite flexible. Sleeping in different positions, including on your stomach, will not change the shape or structure of your breasts. Breasts can change shape or size due to factors like pregnancy, weight gain or loss, hormonal fluctuations, or aging. However, the position you sleep in is unlikely to have a significant impact on the shape of your breasts. It's more important to wear a properly fitting bra during the day to provide adequate support to prevent any discomfort or issues related to breast health. So, feel free to sleep in your preferred position without

worrying about it having any negative effects on your breast shape.

Q: “Are vitamins in cereal ‘sprayed on’? Do you miss some of them if you don’t drink the leftover milk?” (Rubric: DPO 0.62 vs. preferred 0.21.)

Prompted-DPO (rejected). Imagine vitamins in cereal as tiny superheroes that come to save the day by making your body strong and healthy. When you pour milk over your cereal, it’s like sending out a rescue mission to capture these superheroes. The milk acts as a special agent that helps the vitamins get absorbed by your body. Now, when you leave some milk behind, it’s like leaving some superheroes stranded. If you don’t drink that leftover milk, you might miss out on some of these superheroes that were meant to save the day. When you pour milk over your cereal, it helps break down the vitamins into smaller pieces that your body can easily use. Think of it like a chef breaking down a big piece of meat into smaller, bite-sized pieces for a delicious meal. If you don’t drink the leftover milk, some of these broken-down vitamins might not get absorbed by your body. It’s not that you’re missing out on a huge amount, but it’s like leaving a few superheroes behind in the cereal bowl. So, while it’s not a catastrophe if you don’t drink the leftover milk, it’s always a good idea to make sure you get as many superheroes as possible to keep your body strong and healthy.

Preferred (winning). Vitamins in cereals are not exactly “sprayed on,” but rather they are often added during the manufacturing process. These vitamins are included to fortify the cereal and make it a more nutritious choice for consumers. The vitamins can be sprayed on the cereal as a coating or mixed directly into the cereal flakes or pieces before they are packaged. This is done to ensure that the cereals provide essential nutrients like vitamins and minerals that might be lacking in the original ingredients or lost during processing. Some of the added vitamins can be water-soluble, meaning they can dissolve in water, like the milk in your cereal. If you pour milk over your cereal and then drink the leftover milk after eating the cereal, you may indeed consume some of the vitamins that have dissolved into the milk. This is especially true for certain vitamins like vitamin C and some B vitamins. However, not all vitamins will dissolve into the milk, and some may be left behind in the cereal bowl. So, while you may miss out on some vitamins if you don’t consume the leftover milk, the majority of the added vitamins would have been consumed along with the cereal itself.

These qualitative patterns reinforce the quantitative finding (Section 5.2.1) that DPO over-optimizes metaphor at the expense of human preference.

A.2 Across-Model Examples: Rubric Scores and Human Preference

The over-optimization examples above are drawn from the DPO variant. To show how the rubric relates to human preference more broadly, we give representative pairs from the pairwise human evaluation spanning *other* models (the GPT-3.5 teacher, the human-written Reddit answers, and the Human-SFT and Synthetic-SFT Llama-3.1-8B students), each labeled with its aggregate rubric score (0–1) and with the human-preferred answer marked. The pattern is not monotone: a higher rubric score often coincides with the preferred answer (*caffeine*),

but not always, because the rubric measures device *presence* rather than usage quality (§4.2), so a rhetorically dense answer can still lose to a plainer, more precise one (*blue light, chromosomes*).

Q: “Why do coffee and tea plants produce caffeine?” (Rubric: GPT-3.5 teacher 0.61 vs. Human 0.19; **human-preferred: GPT-3.5 teacher.**)

Here higher rubric and human preference agree: the structured, analogy-driven teacher answer was both scored higher and chosen.

GPT-3.5 teacher (preferred). Think of caffeine as a defense mechanism for coffee and tea plants, like a natural alarm system. These plants don’t have the luxury of running away from threats like animals do. Instead, they have developed caffeine as a way to fend off pests. . . . Just like how a loud siren would scare off a burglar, the bitterness and toxicity of caffeine in these plants would deter pests from coming back for more. . . . Therefore, over time, coffee and tea plants have evolved to produce caffeine as a survival strategy in the wild.

Human / Reddit (rejected). caffeine is produced by plants as a defense against insects. you find caffeine in the parts of plants most vulnerable to herbivory: coffee beans which develop into seedlings, tea leaves, etc. . . . many naturally occurring poisons have been co-opted for alternative uses by humans because we are simply too big to be killed by them.

Q: “If an object is ‘blue’ because it only reflects blue light, why isn’t it colorless under red light?” (Rubric: GPT-3.5 teacher 0.59 vs. Human 0.20; **human-preferred: Human.**)

Here the rubric and preference diverge: the higher-scoring teacher answer stacks metaphors (“picky eater,” “blue-themed party”) but is looser, and annotators preferred the shorter, physically precise human answer.

GPT-3.5 teacher (rejected). Let’s think of colors on objects like a picky eater that only wants to reflect specific types of light. . . . imagine you’re at a blue-themed party, and suddenly someone shows up in a bright red outfit. The blue object will still play hard to get, only reflecting the blue light. . . . It’s like trying to change someone’s favorite color – they just won’t budge!

Human / Reddit (preferred). most materials aren’t perfect absorbers of light outside the range that makes up the main color. a blue object will still reflect a small amount of red light. if an object would absorb 100% of the light falling onto it, it would not appear to be invisible, but instead would appear perfectly black.

Q: “How did we get extra chromosomes?”

(Rubric: Synthetic-SFT (prompted) 0.55 vs. Human-SFT (prompted) 0.31; **human-preferred: Human-SFT.**) A within-student example: the Synthetic-SFT answer scores higher on the rubric (analogy 1.0, readability 0.77) via a “deck of cards” analogy, yet annotators preferred the plainer, more directly explanatory Human-SFT answer.

Synthetic-SFT, prompted (rejected). Let’s start with the basics. Chromosomes are like the instruction manuals for our cells. . . . imagine these chromosomes as a deck of cards. Each pair of chromosomes is like a pair of identical twins. . . . When a sperm and egg cell come together during fertilization, they combine to form a zygote, which is like shuffling the deck and creating a new set of cards. . . . the extra chromosomes are like the duplicate set of instructions that come together to create a new individual.

Human-SFT, prompted (preferred). the human body has 23 pairs of chromosomes, for a total of 46. . . . the reason we have 23 pairs is because of a process called meiosis, which is how we make sperm and eggs. meiosis is a special type of cell division that reduces the number of chromosomes in a cell by half. this is important because when a sperm and an egg combine, they need to have the same number of chromosomes. . .

Model	MC2	Δ	Wilcoxon p	Cohen's d
Base LLaMA3.1	.461	—	—	—
Synthetic SFT	.469	.009	.247	.064
Synthetic DPO	.469	.008	.270	.053
Human SFT	.471	.010	.308	.075
Human DPO	.461	.001	.206	.059

Table 4: TruthfulQA MC2 scores across model variants. All differences from the base model are not statistically significant (Wilcoxon signed-rank test, $p > 0.05$) with negligible effect sizes ($|d| < 0.1$).

so, meiosis is a way of making sure that the sperm and egg have the right number of chromosomes.

Together these cases illustrate why we treat the rubric as a measure of rhetorical-device *presence* and rely on the human-preference study (§5.2.1) to establish which of those features actually track communication quality.

A.3 Factuality and Truthfulness Checks

We assessed truthfulness in three complementary ways: TruthfulQA, participant ratings on a five-point accuracy scale, and a claim-level factuality analysis using GPT-5.2 with web search. Together, these provide converging evidence of robust factuality, though none constitutes exhaustive verification of domain-specific scientific accuracy.

TruthfulQA. To verify that fine-tuning for science communication does not compromise factual accuracy, we evaluated all model variants on TruthfulQA (Lin et al., 2022), a benchmark of 817 questions designed to measure whether models generate false statements that mimic human misconceptions. Table 4 shows MC2 scores, which measure the probability mass assigned to correct answers (higher is better), along with statistical tests measuring differences relative to the base model.

Participant accuracy ratings. As part of the human preference study, participants also rated answer accuracy on a five-point scale. These ratings were generally credible: 42.3% of responses were rated *Mostly accurate*, 41.0% were rated *Mixed*, and 16.7% fell into the remaining categories.

Claim-level factuality. To verify that fine-tuning did not degrade factual accuracy, we decomposed answers into atomic claims and verified them using GPT-5.2 with web search access. As shown in Table 5, the majority of claims were supported across all models, with the SFT student even slightly improving over the base model.

Model	Micro	Macro
GPT-3.5	81.8	80.0
SFT	75.9	75.8
Base	74.5	73.7

Table 5: Claim-level factuality analysis (micro = over claims; macro = per question).

A.4 Answer Generation Instructional Prefix

The following system prompt was used to generate synthetic answers with GPT-3.5, with max_tokens set to 256:

You are tasked with writing high-quality
 → scientific answers, given these criteria:
 1. The explanation should have a structured flow
 → from simple to complex concepts.
 2. Establish clear connections between various
 → parts of the explanation.
 3. Assume the reader has minimal prior knowledge.
 4. Usage of didactic tools such as examples,
 → metaphors, analogy, and humor is encouraged.
 5. If possible, try to paint mental images that
 → will stay with the reader. e.g. "Consider
 → each computer as a node and the Internet as a
 → web."
 6. Avoid domain specific jargon and unfamiliar
 → concepts.
 7. Ensure the language is unambiguous, concise,
 → and with clearly defined terminology.
 8. Use of paragraphs will be preferred over
 → bullet points and lists.
 The answers should be around two to three
 → paragraphs long.

A.5 LLM Judge Prompts (G-Eval)

For reproducibility, we include the full G-Eval evaluation instructions used for the metric versions reported in this paper.

Analogy.

STEP 1: An ANALOGY explicitly compares TWO
 → DIFFERENT DOMAINS to explain a concept. It
 → must have:
 - A SOURCE: a familiar, concrete situation
 → (e.g., "a library", "traffic on a highway",
 → "a factory assembly line")
 - A TARGET: the scientific/abstract concept
 → being explained
 - EXPLICIT comparison language such as: "like",
 → "similar to", "just as", "think of X as Y",
 → "imagine", "consider", "analogous to",
 → "comparable to"

Examples of analogies:

- "Think of DNA like a recipe book" (source:
 → recipe book, target: DNA)
- "The immune system works like an army
 → defending a castle" (source: army/castle,
 → target: immune system)
- "Electrons orbit the nucleus similar to how
 → planets orbit the sun" (source: solar
 → system, target: atom)

STEP 2: The following are NOT analogies:

- Simple examples without cross-domain
 - comparison ("For example, water boils at 100°C")
- Definitions ("Photosynthesis is the process by which plants convert sunlight to energy")
- Within-domain comparisons ("Viruses are smaller than bacteria")
- Metaphors without explicit comparison markers
 - ("The brain is a computer" - this is a metaphor, not an analogy)

STEP 3: Return a score of 10 if at least one

- clear analogy (as defined above) is present.

 Return 0 otherwise. Do not use intermediate

- scores.

Humor.

STEP 1: Look for INTENTIONAL JOKES.

- Does the text contain any of these INTENTIONAL
 - comedic elements?
 - A joke with a clear punchline or comedic setup
 - A pun or deliberate wordplay
 - A pop culture reference used for comedic effect
 - (not just as an example)
 - A clearly sarcastic or mocking statement that
 - subverts expectations

EXAMPLES of intentional jokes (Score 10):

- "Schrodinger's cat walks into a bar... and doesn't" (punchline joke)
- "holy proofreading batman" (pop culture catchphrase used for comedic effect)
- "you would not pass go, nor would you collect \$200, you would go directly to nobel prize" (Monopoly reference twisted for humor)

If you found an intentional joke → Score 10.

- Otherwise continue to Step 2.

STEP 2: Look for ABSURDIST or DEADPAN humor.

- Does the text contain any of these?
 - An unexpected non-sequitur at the end that
 - breaks the serious tone
 - A brand name, product, or random word dropped
 - without explanation for comedic effect
 - A jarring tonal shift that seems intentionally
 - funny
 - Self-deprecating or self-aware comments that
 - acknowledge absurdity

EXAMPLES of absurdist/deadpan humor (Score 10):

- A serious scientific answer that ends with
 - "...and partially because american apparel" (non-sequitur brand drop)
- "I'd be lying if I said I've seen it" in a
 - technical explanation (self-aware honesty)
- "at which point you have big problems^(tm)"
 - (fake trademark = comedic aside)
- "now, i must go masturbate before the shakes
 - set in" (absurd self-deprecating callback)

If you found absurdist or deadpan humor → Score

- 10. Otherwise continue to Step 3.

STEP 3: Look for IRONY, SARCASM, or DARK HUMOR.

Does the text contain any of these?

- Saying the opposite of what you mean with clear
 - comedic intent
- Dark understatement: making something serious
 - sound trivially unimportant
- Violent, morbid, or extreme language used
 - casually for shock-comedy value

EXAMPLES of irony/dark humor (Score 10):

- "The surgery went well, apart from the patient
 - dying" (dark understatement)
- "It's perfectly safe, unless you count the
 - possibility of horrible death" (dark humor)
- "telescopes were played out" (ironic dismissal
 - of something obviously important)

If you found clear irony or dark humor → Score

- 10. Otherwise continue to Step 4.

STEP 4: CRITICAL - Distinguish HUMOR from CASUAL

- TONE.

- The following are NOT humor - they are just
 - informal/casual writing:
 - Candid admissions: "I would recommend staying
 - away from Yahoo Answers" (just honest advice)
 - Casual language: "[pretty good summary here]"
 - (informal tone, not sarcasm)
 - Enthusiastic phrases: "great question!"
 - (friendly, not funny)
 - Mild exaggeration without comedic intent:
 - "scrounge up enough firepower" (vivid language)
 - Informal asides that don't subvert
 - expectations: "which is why it's not used much on earth"
 - Rhetorical questions: "good question"
 - (engaging, not joking)

KEY TEST: Would a reader actually laugh or smile?

- Casual/informal tone = reader thinks "that's
 - friendly" → Score 0
- Intentional humor = reader thinks "that's
 - funny" → Score 10

FINAL DECISION:

- If you found intentional humor in Steps 1-3
 - that passes the laugh test → Score 10
- If you only found casual tone, informal
 - language, or vivid explanations → Score 0

Do not use scores between 0 and 10.

Metaphor.

1. DEFINITION

A METAPHOR is ACTIVE / LIVE figurative language

- in which a concept is framed, evaluated, or described using another
 - domain in a NON-LITERAL way.

For this task, ONLY score ACTIVE metaphors. Conventional, idiomatic, or dead metaphors must

- be scored 0,

 UNLESS they are reactivated via a clear domain

- shift.

Metaphors may be imagistic OR abstract. Metaphors may appear WITH or WITHOUT explicit

- comparison words

such as "like", "as", or "similar to".

Score 10 if an active metaphor is present.

Score 0 otherwise.

Do not use intermediate scores.

2. CORE METAPHOR TEST (WITH EXAMPLES)

Ask:

Does the passage DEPEND on a NON-LITERAL

↳ conceptual framing

that changes how the target concept is

↳ understood,

explained, or evaluated?

If YES → continue evaluation.

If NO → NOT a metaphor (Score 0).

Examples:

- "The theory cannibalizes its own assumptions."

↳ → metaphor

- "The theory has internal inconsistencies." →

↳ NOT metaphor

3. NOVELTY & DOMAIN-SHIFT TEST (CRITICAL, WITH EXAMPLES)

Novelty does NOT require original wording.

A metaphor may be ACTIVE if novelty arises from:

- the phrasing itself, OR

- a DOMAIN SHIFT, where a familiar expression is

↳ applied

to a domain where it is not conventionally

↳ used.

HOWEVER:

- If an expression is idiomatic and used freely

↳ across domains,

it should be treated as FROZEN and scored 0.

Positive examples (Score 10):

- "She put him in the friend zone so hard that a

↳ force field

seemed to form around him."

(physics → social interaction)

Negative examples (Score 0):

- "That policy change was the final nail in the

↳ coffin."

(domain-agnostic idiom)

4. FROZEN / IDIOMATIC METAPHORS (WITH EXAMPLES)

Treat expressions as FROZEN (Score 0) if they

↳ are:

- idiomatic and widely recognized

- processed without invoking their source domain

- commonly used across many topics

Examples (Score 0):

- "tip of the iceberg"

- "double-edged sword"

- "final nail in the coffin"

These do NOT become active merely by appearing in technical, scientific, or abstract contexts.

5. LITERAL FALLBACK TEST (WITH EXAMPLES)

Ask:

Can the sentence be interpreted fully literally, with no loss of meaning or intent?

If YES → NOT a metaphor (Score 0).

Examples:

- "The committee rejected the proposal." → NOT

↳ metaphor

- "The proposal was smothered before it could

↳ take shape." → metaphor

6. ABSTRACT METAPHORS (ALLOWED, WITH EXAMPLES)

Metaphors do NOT need to be visual or imagistic.

Abstract metaphors SHOULD be scored 10 if:

- a non-literal source domain is imported

- the framing alters how the concept is

↳ understood

- the meaning cannot be preserved under a literal

↳ reading

Examples (Score 10):

- "The policy devours the resources it was meant

↳ to protect."

- "The algorithm starves certain outcomes of

↳ attention."

Abstract, academic, or technical language alone does NOT qualify as metaphor.

7. PERSONIFICATION TEST (WITH EXAMPLES)

For personification, ask:

Is the attributed action, intention, or behavior clearly impossible or unexpected for the domain?

If YES → metaphor (Score 10).

If NO → NOT metaphor (Score 0).

Examples:

- "White blood cells ambushing invaders." →

↳ metaphor

- "The market responded to new data." → NOT

↳ metaphor

8. METAPHORICAL ANALOGIES (WITH EXAMPLES)

Explicit comparison language ("like", "similar

↳ to",

"think of X as Y", etc.) does NOT exclude a

↳ metaphor.

An analogy is ALSO a metaphor if it introduces novel, expressive, or domain-shifting figurative

↳ framing,

rather than neutral explanation.

Examples:

- "Debugging this problem is like trying to

↳ untangle

headphones while wearing gloves." → metaphor

- "DNA is like a recipe book." → NOT metaphor

9. NOT METAPHORS (SUMMARY WITH EXAMPLES)

Always score 0 for:

- purely literal descriptions

- standard technical terminology

- metonymy (institutions standing for people)

- shape or visual resemblance only

- routine evaluative or descriptive verbs

Examples (Score 0):

- "The heart pumps blood."
- "The White House announced new measures."
- "a mushroom-shaped cloud"

10. CONSERVATIVE TIE-BREAKING RULE

When the case is borderline or ambiguous:

- Prefer Score 0 over Score 10.
- Humans apply a conservative standard; match that standard here.

Scaffolding.

STEP 1: DEFINITION - What is scaffolding?

Scaffolding means the explanation BUILDS

- ↳ UNDERSTANDING by introducing ideas in order
- ↳ of increasing complexity.

WHAT MAKES AN IDEA "SIMPLER" vs "MORE COMPLEX"?

1. STANDALONE vs DEPENDENT:

- Simple: "Atoms are tiny particles"
- ↳ (understood alone)
- Complex: "Electrons orbit the nucleus"
- ↳ (requires knowing what atoms are)

2. FEWER vs MORE PREREQUISITES:

- Simple: "Plants need sunlight" (common knowledge)
- Complex: "Chlorophyll absorbs photons to split water molecules" (requires chemistry background)

3. CONCRETE vs ABSTRACT:

- Simple: "Ice melts when heated" (observable)
- Complex: "Phase transitions occur when thermal energy exceeds intermolecular bonds" (abstract principle)

4. SINGLE vs COMBINED CONCEPTS:

- Simple: "DNA stores genetic information"
- ↳ (one idea)
- Complex: "DNA replication errors during cell division cause mutations that may lead to cancer" (combines 4 concepts)

MINIMUM REQUIREMENT: At least 2 ideas where the second DEPENDS ON or DEEPENS the first.

A SINGLE fact or definition is NOT scaffolding
↳ (nothing to build on).

STEP 2: Examples of SCAFFOLDING PRESENT (Score 10):

SHORT ANSWER examples:

- "Cells are the basic building blocks of life."
- ↳ DNA inside cells contains instructions for making proteins."
- ↳ First establishes cells, then zooms into DNA
- ↳ - builds from whole to part
- "When you heat water, molecules move faster."
- ↳ This is why boiling water evaporates."
- ↳ First explains mechanism (molecules), then applies to observable phenomenon

- "Gravity pulls things toward Earth. That's why we don't float away."
- ↳ First states rule, then connects to everyday experience

LONGER ANSWER example:

- "Two facts motivate my research—first, diverse systems are healthier... My research asks if it matters that species are being gained and lost... By manipulating plant number and type, I can begin to uncover the mechanisms..."
- ↳ Starts with motivation → poses question → describes method (clear progression)

STEP 3: Examples of SCAFFOLDING ABSENT (Score 0):

SINGLE FACT (nothing to build on):

- "Photosynthesis converts sunlight into energy."
- ↳ Just one idea, no building possible
- "The mitochondria is the powerhouse of the cell."
- ↳ Single definition, no progression

UNORDERED LIST of facts:

- "DNA contains genes. Proteins are made from amino acids. Cells divide through mitosis."
- ↳ Multiple facts but no logical order or building between them

ALL AT ONCE (no progression):

- "Quantum entanglement occurs when particles become correlated such that measuring one instantaneously affects the other regardless of distance due to non-local correlations established at particle creation."
- ↳ All concepts crammed together with no build-up

COMPLEX BEFORE SIMPLE:

- "The non-local correlations from entanglement happen because... basically, particles can be connected."
- ↳ Starts with the hard part, then gives the simple explanation

LOGICAL FLOW WITHOUT BUILDING (NOT scaffolding):

- "Scientists collected samples. They analyzed the samples. They published their findings."
- ↳ Ideas flow logically but stay at same complexity level - no building
- "The experiment tested temperature. Results showed higher temps increased reaction speed. This confirmed the hypothesis."
- ↳ Coherent narrative, but each sentence is equally complex - no building
- "Photosynthesis happens in leaves. It uses sunlight. Plants need it to survive."
- ↳ Related facts that flow, but none requires understanding the previous one

KEY DISTINCTION: Scaffolding requires INCREASING DEPTH, not just logical coherence.

STEP 4: Common edge cases:

SHORT != NO SCAFFOLDING:

- A 2-sentence answer CAN have scaffolding if idea #1 prepares for idea #2
- Length doesn't matter; logical ordering does

TRANSITION WORDS DON'T GUARANTEE SCAFFOLDING:

- "because", "therefore", "this means" can
- indicate building, but not always

WITH transitions, IS scaffolding:

- "Gravity pulls objects toward Earth. Because of
- this, dropped objects fall."
- Simple rule first, then applies it (builds)

WITH transitions, NOT scaffolding:

- "Entropy increases because systems tend toward
- disorder because energy spreads out."
- "because" used twice but all concepts at same
- complexity level (no building)
- "Quantum tunneling occurs because
- wavefunctions extend beyond barriers."
- Single complex statement, transition word
- doesn't help

WITHOUT transitions, IS scaffolding:

- "Atoms have electrons. Electrons carry charge.
- Charge creates electricity."
- Clear simple→complex progression despite no
- explicit connectors

CHRONOLOGICAL ORDER != SCAFFOLDING:

- "First I added water, then I heated it, then it
- boiled" is just a sequence
- Scaffolding requires conceptual building, not
- just time order

LOGICAL FLOW != SCAFFOLDING:

- A well-organized answer can flow smoothly
- without scaffolding
- Scaffolding requires each idea to be MORE
- COMPLEX than the previous
- If all ideas are at the same complexity level,
- it's NOT scaffolding (even if coherent)

FINAL DECISION:

- Score 10: Ideas are ordered from
 - simple→complex, with building/dependence
 - Score 0: Single idea, random order, or
 - complex-before-simple
- Do not use intermediate scores.

Connection to everyday life.

STEP 1: Check for ANY named reference from the

→ following categories.

SCORE 10 for ANY of these:

A) POPULAR CULTURE:

- Movies, TV shows, books, documentaries (e.g.,
- "Jurassic Park", "Game of Thrones", "The
- Matrix")
- Music, bands, songs (e.g., "like a Beatles
- song", "Beethoven's 5th")
- Video games, board games (e.g., "Monopoly",
- "Minecraft", "chess")
- Internet culture, memes, viral moments
- Fictional characters (e.g., "Batman", "Sherlock
- Holmes")

B) HISTORICAL FIGURES AND EVENTS:

- Famous scientists, leaders, artists (e.g.,
- "Galileo", "Einstein", "Napoleon", "Carl
- Sagan")
- Historical events (e.g., "the French
- Revolution", "World War II", "the Cold War")

- Famous inventions or discoveries (e.g., "the
- Tsar Bomba", "the Manhattan Project")

C) ORGANIZATIONS AND BRANDS:

- Companies, brands, products (e.g., "Google",
- "Netflix", "American Apparel", "Teflon")
- Government agencies (e.g., "NASA", "FBI",
- "FCC")
- Well-known institutions (e.g., "Harvard", "the
- Smithsonian")

D) PUBLICATIONS AND MEDIA:

- Magazines, newspapers (e.g., "The Economist",
- "Psychology Today", "Time Magazine")
- Websites, platforms (e.g., "Yahoo Answers",
- "Wikipedia", "Reddit")

E) FAMOUS PLACES:

- Landmarks (e.g., "Times Square", "the Eiffel
- Tower", "Chernobyl")
- Well-known locations used as reference points

STEP 2: ALSO check for concrete everyday

→ activities.

SCORE 10 for relatable human experiences that

- ground abstract concepts:
- Household tasks: "like doing laundry", "like
- vacuuming"
- Food & cooking: "like mixing cake batter",
- "like grocery shopping"
- Common experiences: "like getting stuck in
- traffic", "like waiting for coffee"
- Childhood activities: "like a child spinning in
- a circle with arms outstretched"
- Social activities: "like hosting a dinner
- party"

KEY: The activity must be something a general

→ audience experiences personally.

STEP 3: CHECK THE "EXTERNAL TO TOPIC" RULE.

IMPORTANT: The reference must be EXTERNAL to the

→ question topic.

SCORE 0 if the reference is ALREADY PART of the

→ question:

- Question asks "how has our knowledge of
- dinosaurs changed since Jurassic Park"
- Mentioning "Jurassic Park" in the answer does
- NOT count (it's part of the question)
- Question asks about NASA missions → Mentioning
- "NASA" does NOT count
- Question asks about Einstein's theories →
- Mentioning "Einstein" does NOT count

SCORE 10 if the reference is EXTERNAL:

- Question about physics → Mentioning "Galileo"
- DOES count (external historical figure)
- Question about nuclear reactors → Mentioning
- "the Cold War" or "Tsar Bomba" DOES count
- Question about space → Mentioning "NASA" DOES
- count (if NASA wasn't in the question)

STEP 4: What does NOT count (Score 0).

- Generic physics analogies: "like a ball rolling
- downhill", "like waves in water"
- Nature analogies: "like a river flowing", "like
- clouds forming"

- Abstract comparisons: "like a sponge"
 $\hookrightarrow$ absorbing", "like a balloon"
- Vague references: "as we all know", "in
 $\hookrightarrow$ everyday life"
- Technical jargon without cultural grounding
- References already embedded in the original
 $\hookrightarrow$ question (see Step 3)

STEP 5: FINAL DECISION.

Score 10 if:

- You found ANY named reference from Step 1
 $\hookrightarrow$ categories, AND
- The reference is EXTERNAL to the question topic
 $\hookrightarrow$ (passes Step 3)

OR:

- You found a concrete everyday activity from
 $\hookrightarrow$ Step 2, AND
- The activity is EXTERNAL to the question topic

Score 0 otherwise.

EXAMPLES:

- "Galileo discovered the moons of Jupiter" in a
 $\hookrightarrow$ question about CERN $\rightarrow$ Score 10 (Galileo is
 $\hookrightarrow$ external)
- "the Tsar Bomba" in a question about killing
 $\hookrightarrow$ stars $\rightarrow$ Score 10 (famous Cold War bomb,
 $\hookrightarrow$ external)
- "Psychology Today" in any answer $\rightarrow$ Score 10
 $\hookrightarrow$ (named publication)
- "NASA" in a question NOT about NASA $\rightarrow$ Score 10
 $\hookrightarrow$ (famous organization)
- "like a child spinning" in a physics question $\rightarrow$
 $\hookrightarrow$ Score 10 (relatable everyday activity)
- "Jurassic Park" in a question asking about
 $\hookrightarrow$ Jurassic Park $\rightarrow$ Score 0 (NOT external)

Do not use intermediate scores.

A.6 Synthetic Preference Label Noise: Llama-2-7B vs. GPT-3.5 Win Rate

To quantify the noise in the synthetic preference labels (GPT-3.5 as *chosen*, Llama-2-7B as *rejected*), we evaluated Llama-2-7B on the held-out test set (91 questions) using the same rubric pipeline as all other models and compared its per-question aggregate score against GPT-3.5 (CoT-prompted). Table 6 summarizes the results.

Model	Questions won	Mean agg. score
Llama-2-7B	23 / 91 (25.3%)	0.170
GPT-3.5 (CoT)	68 / 91 (74.7%)	0.330

Table 6: Per-question rubric win rate of Llama-2-7B vs. GPT-3.5 (CoT-prompted) on the test set. A "win" means the model's weighted aggregate rubric score was higher on that question. The 25.3% win rate for Llama-2-7B confirms that the GPT-3.5 > Llama-2-7B preference ordering is a noisy approximation rather than a clean ground truth.

Teacher	Unprompted	Prompted
GPT-3.5 (original)	24.5%	54.2%
GPT-5	24.3%	57.5%
Kimi K2.6	34.1%	59.3%

Table 7: Aggregate rubric scores of Llama-3.1-8B students distilled from different teachers, evaluated on the held-out test set in unprompted and prompted modes. Paired Wilcoxon p -values vs. the GPT-3.5 student: GPT-5 unprompted $p=0.25$, prompted $p<0.001$; Kimi unprompted $p<0.001$, prompted $p=0.0009$.

A.7 Teacher Robustness: Second-Teacher Distillation

To test whether our findings depend on the specific teacher (GPT-3.5), we regenerated the synthetic SFT dataset with two additional teachers, GPT-5 (a substantially newer OpenAI model) and Kimi K2.6 (Moonshot; non-OpenAI), and repeated the identical SFT procedure on the same Llama-3.1-8B base. Table 7 reports the resulting students' aggregate rubric scores against the original GPT-3.5-distilled students, in both unprompted and prompted inference modes. Unprompted, the GPT-5 student is statistically indistinguishable from the GPT-3.5 student (24.5% vs. 24.3%; paired Wilcoxon signed-rank over the 91 test questions, $p=0.25$); prompted, GPT-5 yields a small but significant edge (54.2% vs. 57.5%; $p<0.001$). The Kimi student is higher in both modes (34.1% unprompted, 59.3% prompted; $p<0.001$ and $p=0.0009$ vs. GPT-3.5). Per dimension (Table 8), GPT-5 matches GPT-3.5 on stylistic dimensions in both modes, with only prompted scaffolding differing (0.50 vs. 0.70, $p<0.001$), a structural rather than stylistic gap. Kimi has a different mix, with more analogy and metaphor, so teacher identity still shapes style; both additional teachers nonetheless produce strong students, including outside OpenAI.

A.8 Deployment Considerations

Distilling the communication behavior into the student weights changes the deployment trade-off relative to inference-time prompting. The instructional prompt adds approximately 200 input tokens to every teacher-model query. At the time of our experiments, GPT-3.5 Turbo was priced at \$0.50 per million input tokens and \$1.50 per million output tokens and generated at 54.2 tokens/s. By contrast, the unprompted Synthetic-SFT Llama-3.1-8B requires no specialized prompt and runs locally at approximately 10.3 tokens/s on a single RTX 4090

Dimension	Unprompted			Prompted		
	GPT-3.5	GPT-5	Kimi	GPT-3.5	GPT-5	Kimi
Jargon	.64	.63	.60	.67	.67	.66
Readability	.36	.37	.28	.57	.58	.42
Scaffolding	.22	.23	.36	.50	.70***	.61
Analogy	.18	.15	.52	.90	.88	.97
Metaphor	.07	.08	.29	.60	.61	.90
Humor	.00	.00	.00	.01	.00	.01

Table 8: Per-dimension normalized rubric scores for Llama-3.1-8B students distilled from GPT-3.5, GPT-5, and Kimi K2.6. Against GPT-3.5, GPT-5 differs significantly only on prompted scaffolding (** $p < 0.001$); stylistic dimensions remain n.s. Kimi shows a different rhetorical mix (higher analogy/metaphor), illustrating that teacher identity still shapes style.

using approximately 8 GB of VRAM in 4-bit quantization, with no per-query API charge. Despite this simpler inference setup, it achieves statistical parity with the prompted teacher in human preference (50.2%, $p=0.95$; Table 2). These throughput figures are illustrative rather than hardware-normalized: the teacher measurement reflects a hosted API, whereas the student measurement reflects local consumer hardware.