KNOWLEDGEGAIN: Evaluating and Optimizing Science News Generation for Reader Learning

Anonymous ACL submission

Abstract

Science news is an important medium to communicate discoveries between the research communities and the public. Yet, most metrics for generated or summarized text evaluate semantic similarity and factual consistency, do not measure how much knowledge readers learn from the news. We introduce KNOWLEDGEGAIN, a metric that evaluates the quality of science news by measuring how much knowledge the readers gained after reading it. To evaluate the metric, we first performed a controlled human study and showed that the metric successfully captures the differential knowledge gained by human readers reading different types of science media. The data allowed us to calibrate a prompt-only LLM reader simulator. We use it to rank and filter candidate articles before human evaluation. A second human study shows that articles selected with this simulator improve post-reading accuracy and normalized KNOWLEDGEGAIN over a strong generation baseline. Our work is a step toward generating science news that better meet the knowledge and comprehension goals of Bloom's Taxonomy.

1 Introduction

Science is often regarded as a common inheritance of humanity (Pasteur, 1939). However, scientific publications are often written in highly specialized language that the vast majority of the general public cannot readily understand. Even with the growth of open-access in removing paywalls, the gap between scientific knowledge and public comprehension (Suber, 2012; Bucchi, 2008) still exists and is widening. Inspired by this need, computational linguists and NLP researchers aim to develop AI-driven methods for automatically generating science news from scientific papers, making science not only open in access, but open in language. However, how to effectively evaluate AI-generated science news computationally remains a foundational question. Standard evaluation metrics for text generation and summarization such as ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and BLEURT (Sellam et al., 2020) primarily measure the lexical or semantic similarity against references. More recent metrics focus on factual consistency between a summary and its source using question answering or natural language inference (Kryściński et al., 2020; Fabbri et al., 2022; Scialom et al., 2021; Deutsch et al., 2021). These metrics either assume a ground truth exists or the evaluation is based on a reference text. Reader learning, which serves as an important goal of science news, is not incorporated.

This creates a significant gap especially important in science news generation. Early work on lay summarization of scientific text improves the comprehension of non-experts (Stoll et al., 2022; Maurer et al., 2021; Anderson et al., 2022). However, science news readers cover a broader spectrum of audiences, requiring more public oriented language and faithfully preserving scientific facts, so readers could learn authentic knowledge without strong domain background (Shulman et al., 2024; Yeung et al., 2018). Therefore, a fundamental question to bridge this gap is: *How to evaluate the learning usefulness of science news and compare it with other types of science communications by measuring the knowledge gained by readers?*

Knowledge assessment has been extensively studied in cognition and educational theory. Foundational work has been established by Bloom's taxonomy (Bloom et al., 1956), revised by Anderson and Krathwohl (2001) about the levels of cognitive learning, and the concept of learning gain (Hake, 1998), formally defined as $g = \frac{post-pre}{max-pre}$, in which pre and $post$ are the scores before and after an intervention. Inspired by these two theories, we propose a novel computational metric KNOWLEDGEGAIN, denoted by $KGain$, to measure the usefulness of science news based on how

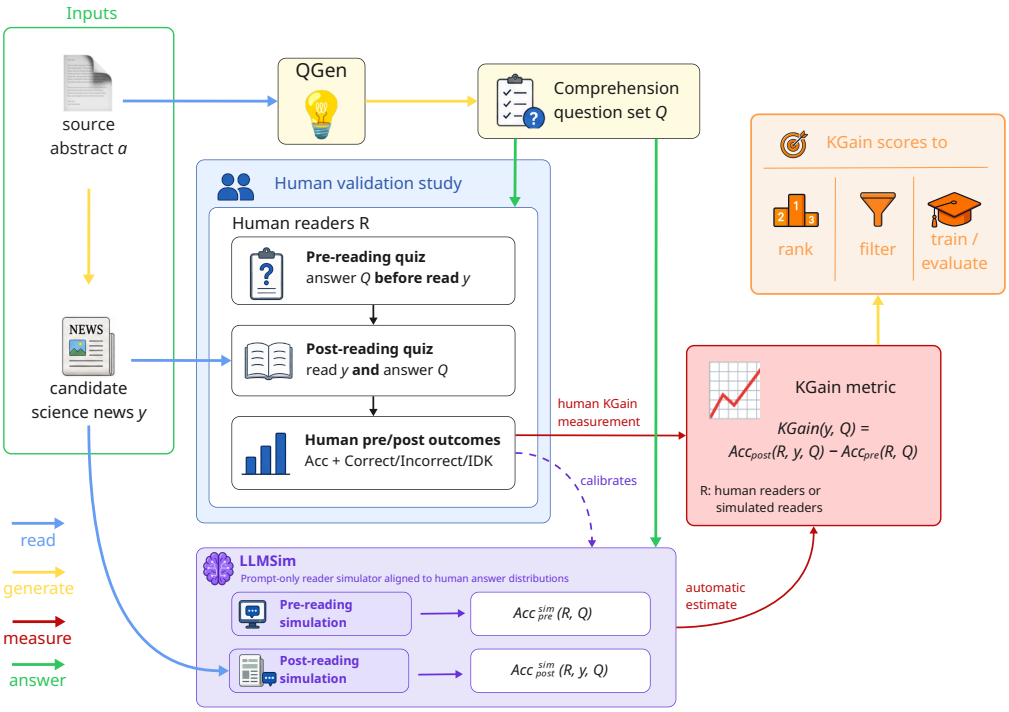

Figure 1: Overview of the pipeline that estimates $KGain$ with human-calibrated LLMSIM. Given a source abstract a, QGEN generates a comprehension question set Q. Human readers answer the same questions before and after reading a candidate science news article y, producing distributions over Correct, Incorrect, and "I do not know" (IDK). These distributions calibrate LLMSIM, which estimates pre/post accuracy. The resulting $KGain$ is used to score and rank articles, filter training data, and train or evaluate science news generation.

much knowledge learned by readers. It leverages questions as proxies to measure the knowledge corresponding to the knowledge and comprehension layers in Bloom's Taxonomy. Though motivated by science news, $KGain$ intuitively is applicable to other types of science communications such as abstracts and tweets/threads.

We first validate $KGain$ in a controlled human study and demonstrate that it captures meaningful differences of knowledge gained through reading three typical communication types. To scale reader-centric evaluation beyond expensive human studies, we develop LLMSIM, a reader simulator model to simulate the reading and question-answering process of human readers. We demonstrate that LLMSIM provides a scalable estimation of $KGain$ that allows us to qualitatively evaluate science news which can be further used for filtering automatically generated news. Figure 1 summarizes the pipeline.

We perform rigorous evaluation of $KGain$. First, we test whether $KGain$ distinguishes knowledge learned from three types of science communications: abstracts, science news, and tweets. Second, we evaluate whether LLMSIM can approximate human pre/post distributions and approximates aggregate $KGain$. Third, we use $KGain$ estimated by LLMSIM to rank and filter science news, which is used for training science-news generators, and evaluate the resulting news with human judgments. Finally, we compare $KGain$ with standard similarity metrics and LLM-judge scores to test whether $KGain$ captures orthogonal dimensions not captured by standard metrics.

This paper makes three major contributions:

- **A reader-centered evaluation metric.** We introduce KNOWLEDGEGAIN, a pre/post comprehension metric for evaluating how useful science communication is for reader learning and apply it to science news evaluation and generation.
- **A calibrated simulated reader.** We develop LLMSIM, a prompt-only LLM simulator aligned with human answer outcome distributions, providing a scalable tool to estimate $KGain$ for generated science news.
- **Science news dataset.** We construct a 2,747-instance abstract–QA–news corpus containing $KGain$-filtered news as a baseline for future science news generation models.

2 Related Work

2.1 Evaluating Generated Text

Automatic evaluation of generated text has traditionally relied on n-gram overlap such as ROUGE, BLEU, and METEOR, or learned similarity like BERTScore and BLEURT (Lin, 2004; Papineni et al., 2002; Banerjee and Lavie, 2005; Zhang et al., 2020; Sellam et al., 2020). Although useful, these metrics often correlate unevenly with human judgments in terms of quality dimensions and datasets. To move beyond measuring token overlap, QA- and NLI-based evaluate factual consistency, and content preservation, including FEQA (Durmus et al., 2020), QAEval (Deutsch et al., 2021), QuestEval (Scialom et al., 2021), FactCC (Kryściński et al., 2020), SummaC (Laban et al., 2022), and QAFactEval (Fabbri et al., 2022). Benchmarks like FRANK (Pagnoni et al., 2021), SummEval (Fabbri et al., 2021), and toolkits such as SacreROUGE (Deutsch and Roth, 2020) enabled large-scale meta-evaluation. Surveys continue to refine best practices for factuality evaluation (Luo et al., 2024). These metrics assess textual similarity, factual consistency, or preference, but they do not directly measure how much a reader learned from the text.

Recent work uses LLMs as reference-free evaluators, often called LLM-as-a-judge, with promising agreement to humans in open-ended tasks (Liu et al., 2023; Zheng et al., 2023). However, LLMs can be unreliable for scoring the aspects of news and pairwise comparisons and do not provide quantitative measurements of the usefulness between two news articles.

2.2 Learning-Centered Evaluation

Pre/post testing is a long-standing paradigm to quantify learning, including the normalized gain popularized by Hake (1998). In information retrieval, *search-as-learning* models change in user knowledge states over a session, typically using pre/post question-answering accuracy (Yu et al., 2021; Ghafourian et al., 2021; El Zein and da Costa Pereira, 2023; Nasser et al., 2024; Tibau et al., 2023). We adopt an interpretable variant for science communication: raw knowledge gain, defined as post-reading accuracy minus pre-reading accuracy.

2.3 LLMs as Simulated Readers

The main obstacle in using LLMs as reader simulators is **hyper-accuracy distortion**: even when instructed to act like an average student, an LLM may still infer the correct answer from linguistic cues and become far more accurate and less uncertain than human readers (Aher et al., 2023; Amirova et al., 2024). This is closely related to recent findings that LLMs may better express human choice distributions when asked to verbalize probabilities than when sampled for a single forced-choice answer (Meister et al., 2025; Xie et al., 2025).

A growing line of work therefore frames simulation as *distribution matching* rather than pointwise prediction. *Mixture-of-persona* methods represent a population with multiple personas rather than one "average human" prompt (Xie et al., 2025; Bui et al., 2025), motivating our use of persona clusters sampled according to their empirical frequency in the human validation study.

Personas alone are not enough if the model can rederive answers from the full article. A more realistic reader simulator should answer from memory, not by searching the original text, which motivates our *memory bottleneck* and connects to cognitive accounts of gist and verbatim memory (Reyna et al., 2016; Reyna, 2012). We model *abstention* directly because $KGain$ depends on whether readers move from IDK or from an incorrect belief to the correct answer (Wen et al., 2025; Madhusudhan et al., 2025; Kapoor et al., 2024). To avoid overconfident answers, we use *verbalized sampling* over answer options including IDK (Aher et al., 2023; Kapoor et al., 2024; Zhang et al., 2025).

3 KnowledgeGain

We define KnowledgeGain as the change in a reader's QA accuracy after exposure to a science communication artifact. Given an artifact y, a reader r, and a set of comprehension questions Q, the pre-reading QA accuracy is:

$$\mathrm{Acc}_{pre}(r, Q) = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{I}[\hat{a}^{pre}_{r,q} = a^*_q], \quad (1)$$

and the post-reading accuracy is:

$$\mathrm{Acc}_{post}(r, y, Q) = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{I}[\hat{a}^{post}_{r,q} = a^*_q], \quad (2)$$

where a^*_q is the correct answer to question q. The accuracy-based KnowledgeGain is then:

$$KGain(r, y, Q) = \mathrm{Acc}_{post}(r, y, Q) - \mathrm{Acc}_{pre}(r, Q). \quad (3)$$

Since the raw $KGain$ can be affected by ceiling effects when pre-reading accuracy is high, we compute normalized $KGain$ as a robustness check:

$$g = \frac{\text{Acc}_{post} - \text{Acc}_{pre}}{1 - \text{Acc}_{pre}}. \quad (4)$$

The mean KNOWLEDGEGAIN of a population of readers R is then:

$$KGain(y, Q) = \frac{1}{|R|} \sum_{r \in R} KGain(r, y, Q). \quad (5)$$

3.1 Question Set Construction

Given a reading document, K multiple-choice comprehension questions are designed to first measure the *Knowledge* and *Comprehension* layers in Bloom's taxonomy (Bloom et al., 1956). To facilitate the QA process, we developed QGEN that produces questions based on the information source. To meet the goals of measuring the knowledge levels in Bloom's Taxonomy, we require all questions to be directly answered (knowledge) or the answers are inferrable from the information source (comprehension). We define a question as answerable if the correct option is either directly stated in the source or can be inferred from one or more source statements without requiring outside knowledge. Questions whose answers cannot be recovered from the source, have multiple plausible correct options, or depend on assumptions are replaced.

To incrementally measure the capability of knowledge gain, we design 6 questions in three tiers of difficulty, each mapping to a layer of Bloom's taxonomy. Q1–2 are extractive true/false questions targeting answers explicitly present in the source; Q3–4 are extractive multiple choice questions with distractors; and Q5–6 are inferential multiple choice questions requiring the readers to synthesize information not explicitly stated but can be inferred from the source. Together, these question tiers cover factual recall and basic comprehension in Bloom's taxonomy. Each question includes an explicit "I do not know" (IDK) option, scored as incorrect for accuracy but preserved as an abstention outcome. This design allows us to distinguish uncertainty from incorrect beliefs.

3.2 Human-based Measurement

Human-based $KGain$ can be measured with a pre/-post protocol by human readers. Readers first answer the question set using background knowledge, then read an artifact, and finally answer the same questions without returning to the artifact.

3.3 Automatic Measurement

It is too expensive to apply humans for every generated article during data construction and model development. We therefore introduce LLMSIM, a prompt-only LLM procedure that approximates human behavior and estimates $KGain$. LLMSIM is not intended to replace human evaluation. Instead, it provides a scalable proxy for scoring large numbers of generated articles before final human validation. LLMSIM simulates readers in the same stages as the human study: answer from background knowledge, read the artifact, form a short memory trace, and answer again without returning to the text.

Mixture-of-personas. We model the reader population as a mixture-of-persona clusters derived from the human validation study. We first inspect participants' response patterns, including baseline knowledge, tendency to abstain, and post-reading improvement, and defined five reader personas that capture recurring behaviors. Each human participant is assigned to one persona cluster, and simulated readers are sampled from these clusters according to their empirical frequency in the validation study. The exact prompts and their frequencies can be found in Appendix B.2.

Memory bottleneck. To prevent direct extraction from the full text, the simulator does not answer post-reading questions while viewing the original article. Instead, it first produces a short memory trace summarizing what the reader retained, then answers the questions using only that trace. We use a dual-trace variant that produces two possible recollections and a probability for each trace, which captures variability in what readers may remember.

Abstention. Because each question includes an IDK option, LLMSIM can abstain when it lacks background knowledge, evidence or memory trace to an answer. We score IDK as incorrect for accuracy, but preserve it as a separate outcome during calibration because moving from IDK to Correct means the article filled a knowledge gap, and moving from Incorrect to Correct means the article corrected a misconception.

Verbalized sampling. To avoid mode collapse and inhumanly confident behavior, the simulator does not directly output a single option. Instead, it verbalizes a categorical distribution over answer options including IDK. We sample one answer from

this distribution. This produces more human-like variation across simulated reader while avoiding instructions to intentionally answer incorrectly.

Calibration. We calibrate LLMSim by matching human outcome distributions from the validation study. For each topic i, question j, and condition $m \in \{\text{Pre}, \text{News}, \text{Abstract}, \text{Tweet}\}$, let $P_{i,j}$ denote the empirical human distribution over the three outcomes. We tune simulator parameter, θ, to minimize KL divergence, a standard distributional mismatch, between the human and simulated distributions:

$$\theta^*_m = \arg\min_{\theta} \sum_{i,j} D_{\mathrm{KL}} \left(P_{i,j,m} \,\|\, Q_{\theta,i,j,m} \right). \quad (6)$$

Matching accuracy alone would collapse IDK and incorrect answers into the same category. Instead, this distributional objective preserves IDK as a separate calibration target, allowing it to model both uncertainty and wrong beliefs. The full algorithm is shown in Appendix B.3.

Simulated *KGain* Scoring. After calibration, we use LLMSIM to estimate $KGain$ for generated articles. For each article, LLMSIM samples simulated readers, computes pre-reading and post-reading accuracy over the associated question set.

$KGain$ measures targeted comprehension gain for a fixed question set Q, not scientific understanding in general. Because Q is generated before evaluation and screened for answerability and uniqueness, $KGain$ is most appropriate for evaluating whether an article communicates the source claims needed to answer factual and inferential comprehension questions.

4 Evaluation

We evaluate KNOWLEDGEGAIN in four stages: first as a metric across different communication types using human readers, then we demonstrate that $KGain$ can be automatically estimated with LLMSIM, then as an optimization signal for science news generation and finally we verify that $KGain$ optimized news let human readers learn more knowledge.

4.1 Human-based *KGain*

In this experiment, we calculate $KGain$ by letting human readers answer questions before and after reading three types of science communications. We aim at testing two hypotheses: science news yields $KGain$ comparable to abstracts, and science news yields higher $KGain$ than tweets/threads.

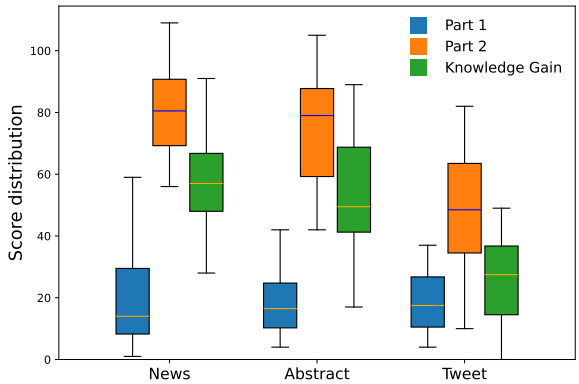

Figure 2: Score distributions for Part 1 (pre-reading), Part 2 (post-reading), and the resulting KNOWLEDGEGAIN across the three media types

Group	Block 1 D1–D10	Block 2 D11–D20	Block 3 D21–D30
G1	News	Abstract	Tweet/Thread
G2	Abstract	Tweet/Thread	News
G3	Tweet/Thread	News	Abstract

Table 1: Crossover block design for the human validation study. Each group evaluates all three artifact types, and each topic block is evaluated under all three artifact types across groups.

Experimental Setup. We validate $KGain$ on 30 research topics, each represented by a paper abstract, a professionally written science news[1], and a tweet/thread from X.com. We automatically collected the news that linked to papers, manually verified, and used those paper links to find promotional posts or threads. The resulting topic set covers six broad science areas in physics, nature, space, health, humans, and technology. We then use QGEN to generate six comprehension questions from the abstract, yielding 180 total questions manually verified following the criteria in § 3.1.

We recruited 30 STEM undergraduates and randomly divided them into three 10 person groups with a balanced GPA distribution. The 30 topics were split into three blocks of 10 topics and each group evaluated one artifact type per block, as shown in Table 1.

Results. The results of our human study are used for testing the two hypotheses above.

[1]All human-written news articles are from sciencealert.com

News vs. Abstracts (H1). As shown in Figure 2, background knowledge was consistently low across all groups (median scores < 20). Post-reading scores show that News articles achieved a median score of approximately 80, slightly outperforming Abstracts (≈ 78). Most importantly, the $KGain$ distribution for News (median ≈ 60) is comparable to, and shows less variance than, that of Abstracts (median ≈ 50). This supports the hypothesis that science news achieves comparable to or higher $KGain$ than paper abstracts, which is consistent with the intuition. Most undergraduate students have not professionally trained on reading scientific papers. However, by reading science news, these students gain comparable or even more knowledge than reading paper abstracts, which is reflected in $KGain$.

News vs Tweets (H2). In contrast, Tweets yielded significantly lower $KGain$ compared to both News and Abstracts. While Tweets required the least reading time (Figure 3 in Appendix A.2), they produced substantially lower knowledge gain, supporting H2.

Implications. The results of this experiment demonstrate that although news articles required the longest reading time, they also produced the largest absolute gains in knowledge. This suggests that the additional reading time is associated with greater knowledge gain, compared with abstracts and tweets/threads and $KGain$ provides a useful way to distinguish the usefulness of learning of different types of science communications. A mixed-effects robustness check confirms the pattern ($p < 0.001$): News and Abstract yield larger gains than Tweet, while News and Abstract are not significantly different; full results are in Appendix A.3.

4.2 Automatic KGain Estimation

We evaluate whether LLMSIM, defined in § 3, provides a reasonable automatic proxy for human pre/-post learning.

Alignment with human readers. From Table 2, we can observe that LLMSIM aligns reasonably well in News and Abstract aggregate distributions, but Tweet remains harder. This makes sense because tweets are short, under-specified and more prone for ambiguity. The aggregate distributions show that LLMSIM aligns closely with human post-reading behavior for News and Abstract conditions, especially abstention rates. The main mismatches are on "Pre," where the model remains more knowledgeable than humans, and "Tweet," where the simulator over-abstains.

Condition	KL ↓	Correct MAE ↓	IDK MAE ↓
Pre	0.090	0.151	0.193
News	0.021	0.099	0.009
Abstract	0.007	0.059	0.010
Tweet	0.119	0.068	0.224

Table 2: Alignment between human and full LLMSIM answer-outcome distributions by condition. KL measures aggregate alignment over Correct/Incorrect/IDK outcomes. Correct mean absolute error (MAE) and IDK MAE measure calibration of correctness and abstention rates. Lower values are better.

Ablation. LLMs are not naturally good human readers: simulator design must be careful. The results in Table 3 show that the main challenge is not simply prompting an LLM to act like a student. Direct prompting produces a hyper-accuracy simulator with almost no uncertainty. The most important design choices are therefore mechanisms that mimic human-like uncertainty.

Table 3 shows that naive LLM answering is a poor proxy for human learning behavior: it is hyper-accurate and never abstains. Persona prompting alone has little effect, and the memory bottleneck reduces over-accuracy but does not capture uncertainty. The largest gains come from modeling uncertainty directly through the IDK option and verbalized sampling, which sharply reduce global KL and bring aggregate correctness close to the human rate. Item-level KL remains higher than global KL, so we treat LLMSIM as a scalable proxy for relative $KGain$ comparison rather than a replacement for human evaluation.

4.3 KGain optimized Science News Generation

In the above experiments, we showed that $KGain$ can be used to evaluate and compare the usefulness of different kinds of science communications, including science news, and that it was possible to design an automatic evaluator for a given science news. However, we have not verified that filtering generated news with simulated $KGain$ actually improves article quality or whether it helps readers learn more than they would from a strong baseline news generator.

Setup. To fine-tune a strong science news generator, we construct a corpus of 2,747 instances.

Simulator variant	Global KL ↓	Item KL ↓	$KGain$ corr. ↑	Correct MAE ↓	IDK MAE ↓
Direct answer	4.953	5.099	-0.063	0.611	0.254
+ Persona mixture	4.897	4.975	-0.044	0.601	0.254
+ Memory bottleneck	4.613	4.183	0.065	0.526	0.254
+ IDK option	0.381	3.653	**0.272**	0.395	0.306
+ Verbalized sampling	**0.023**	**1.934**	0.263	**0.329**	**0.270**

Table 3: Ablation of LLMSIM components against human pre/post answer-outcome distributions. Direct answering substantially overestimates human correctness and never abstains. Adding an explicit IDK option and verbalized sampling sharply improves distributional alignment. Bold numbers indicate the best numerical values.

Each contains a scientific abstract, comprehension questions generated using QGEN with ground truth answers, multiple candidate news articles, and the best news articles ranked by $KGain$. The candidate articles are generated by a teacher model using direct prompting. An open-source model scores each candidate for factual accuracy, completeness, relevance, clarity, and an estimated $KGain$ given the QA set for each abstract. This corpus is used to construct both unfiltered and $KGain$-filtered SFT training sets for fine-tuning the generator.

We evaluate generated articles on 300 held-out abstracts using LLMSIM. Given abstract a, we prompt a generator model to compose science news with a headline-first, multi-paragraph structure, journalistic tone. The generator is *not* allowed to access any reference news article. We sample C candidates per abstract to form a candidate set $\{y_1, \ldots, y_C\}$.

We use LLMSIM to score candidate articles by simulated $KGain$ and construct the $KGain$-filtered SFT dataset by selecting candidates whose simulated $KGain$ scores are positive. We fine-tune a news generator using the data selected. Importantly, LLMSIM is used only as a development time filtering signal, our final claims are based on blinded human evaluations and human learning measures. We then compare our model against 3 baseline models: Baseline 1, a zero-shot news generator; Baseline 2, an agentic news generator; and Baseline 3, human-written science news collected from ScienceAlert for the corresponding paper abstract. We found Baseline 2 to be the strongest based on Figure 4 in Appendix E.4. All news generators are based on `Qwen3-4B-Instruct-2507`[2], and generator names are anonymized during human and LLM evaluations.

[2]We use the same LLM for all automatic generators. We do not test closed-source models, as a stronger backbone because they would likely improve all automatic systems without changing the relative comparison.

Generator	Accuracy	Completeness	Relevance	Clarity	Mean
Agentic	3.34±0.67	4.30±0.57	4.34±0.82	4.06±0.68	4.01±0.63
Ours	**4.24**±0.34	**4.57**±0.18	**4.50**±0.43	**4.08**±0.56	**4.35**±0.28

Table 4: Pointwise evaluation of generated articles. Scores are on a 1–5 scale, with higher values indicating better quality. Values are means with 95% confidence intervals computed over respondent-level means. Bold values indicate the best score in each column.

Pointwise evaluation. We first check whether $KGain$ optimization preserves basic article quality. Participants rated blinded articles on a 5-point Likert scale for accuracy, completeness, relevance, and clarity while having access to the abstract. We compare our $KGain$-optimized generator against the strongest baseline. Table 4 shows that our system receives a higher mean score with the largest difference on accuracy. This suggests that optimizing for learning gain does not come at the cost of basic quality.

Pairwise evaluation. We also test whether readers prefer the $KGain$-optimized articles directly. Participants chose between two blinded articles given the same abstract. The $KGain$-optimized system was preferred over the agentic baseline in 87.0% of comparisons (Wilson 95% CI: [79.0, 92.2]; exact binomial test against chance, $p < .001$). This provides a direct preference check that is independent of the simulated filtering score.

Human *KGain* evaluation. Finally, we directly evaluate whether $KGain$-optimized generation improves reader learning. Participants answered the same comprehension questions before and after reading generated articles, allowing us to compute pre-reading accuracy, post-reading accuracy, raw $KGain$, and normalized $KGain$. In Table 5, the optimized model significantly improves normalized $KGain$ over the baseline ($p = .048$) and also improves post-reading accuracy. Results are averaged over 20 topics and 640 participant–article

Generator	Pre	Post	KGAIN	Human g	Sim. g
Agentic	0.410	0.855	0.446	0.719	0.330
Ours	0.424	**0.881**	**0.457**	**0.777**	**0.378**
Δ	+0.015	+0.026	+0.011	+0.058	+0.048
Wilcoxon p	0.614	0.032	0.695	0.048	0.040

Table 5: Human pre/post $KGain$ evaluation on generated articles. Human g denotes normalized $KGain$; Sim. g denotes LLMSIM normalized $KGain$. Bold values indicate the better generator result for post-reading and gain metrics.

Metric	Spearman ρ	Kendall τ_b
ROUGE-1	0.177	0.121
ROUGE-2	0.296	0.208
ROUGE-L	0.205	0.138
BLEU	0.267	0.167
BERTScore	0.232	0.149
Judge accuracy	0.291	0.240
Judge completeness	-0.016	-0.012
Judge relevance	0.033	0.028
Judge clarity	0.089	0.072
Judge exp. $KGain$	0.168	0.139

Table 6: Rank correlations between automatic metrics and human normalized $KGain$. Spearman's ρ and Kendall's τ_b measure whether automatic metrics preserve the ordering of articles by reader learning.

observations. A mixed-effects model confirms the normalized-gain effect ($p = .015$). Full results are in Appendix C.1. Raw $KGain$ shows a nonsignificant positive trend. Because normalized $KGain$ accounts for the remaining room for improvement after pre-reading accuracy, we use it as the primary human learning measure in the correlation analysis below. At the system level, LLMSIM predicts the same aggregate direction as the human study. However, simulated and human per-topic normalized $KGain$ gains are weakly correlated, so we use LLMSIM as a development-time proxy rather than a replacement for human pre/post evaluation.

4.4 *KGain* and Existing Evaluation Metrics

We compute standard reference-based generation metrics between generated news and the paper abstract: ROUGE-1/2/L (Lin, 2004), BLEU (Papineni et al., 2002), and BERTScore (Zhang et al., 2020). These metrics measure surface/semantic similarity to a reference, but do not directly measure reader learning. We therefore use them to quantify the relationship between $KGain$ and traditional similarity-based evaluation.

Correlation analysis. To test whether existing metrics are predictive of reader learning, we compute Spearman and Kendall rank correlations between $KGain$ and common reference-based metrics. We report both (i) news generated from the same abstracts and (ii) news articles correlations pooled across abstracts, using bootstrap confidence intervals. A weak correlation supports the claim that $KGain$ captures an orthogonal dimension.

To assess the significance of $KGain$ differences, we use paired nonparametric tests (Wilcoxon signed-rank) across topics. Finally, we compute Spearman and Kendall rank correlations and between $KGain$ and reference-based metrics to determine if traditional evaluation captures the "instructional" dimension of science communication.

Table 6 shows that standard automatic metrics correlate only weakly with normalized $KGain$. Source-overlap metrics such as ROUGE-1, ROUGE-2, and BLEU show near-zero association with human normalized $KGain$. BERTScore is weakly associated with $KGain$, suggesting that semantic similarity to the source abstract is not sufficient to predict what readers learn. LLM judge dimensions such as accuracy, completeness, relevance, and clarity likewise show weak correlations.

LLM-judge comparison. We apply the same pointwise rubric and pairwise comparison protocol with an independent LLM judge to test whether judge-based evaluation reproduces human judgments. LLM pointwise scores correlate weakly with human mean scores overall ($r = .19$). Full plots and per-dimension correlations are reported in Appendix C.4. This motivates direct reader-centered evaluation with $KGain$ rather than relying only on generic judge ratings.

Implications. These results indicate that $KGain$ captures a distinct reader-centered signal that is not reducible to lexical overlap, semantic similarity, or generic quality ratings.

5 Conclusion and Future Work

We introduced KNOWLEDGEGAIN, a reader-centered metric for evaluating science news by measuring what readers learn. We validated it with human readers, calibrated LLMSIM as a scalable proxy, and used simulated $KGain$ to filter training data for generation. Human evaluations show that the resulting articles improve normalized learning gains over a strong agentic baseline while preserving article quality. We leave direct reward-model optimization to future work.

6 Limitations

Our human studies use controlled participant pools and do not fully represent the broader public audience for science news. Future work should evaluate KNOWLEDGEGAIN with more diverse readers, including non-STEM audiences.

KNOWLEDGEGAIN depends on the question set. Although we verify questions for answerability and uniqueness, different question designs may emphasize different aspects of understanding, such as factual recall, inference, or broader scientific implications.

LLMSIM is calibrated to aggregate human answer-outcome distributions, but it does not reliably predict topic-level human $KGain$. We use it as a scalable proxy for filtering and model development, not as a replacement for human evaluation.

Finally, our optimization uses supervised filtering rather than direct reward-model or reinforcement-learning optimization. Direct optimization may yield stronger gains, but could increase overfitting to generated questions or simulator artifacts.

7 Ethical Considerations and Risks

This work involves human-subject evaluation of science communication artifacts. Participants were recruited for low-risk comprehension and preference tasks, and responses are analyzed only in aggregate. We do not report personally identifying information, and any participant metadata used for balancing or analysis is de-identified. The study was conducted under IRB approval, and participants provided informed consent and were compensated.

A potential risk of this work is that optimizing generated articles for knowledge gain could encourage systems to maximize answerability while sacrificing nuance, uncertainty, or scientific caution. We mitigate this by evaluating generated articles not only with $KGain$, but also with human and LLM-judge assessments of factual accuracy, completeness, relevance, and clarity. We also treat LLMSIM as a development-time proxy rather than a substitute for human validation, since simulated and human gains may diverge at the topic level.

Generated science-news articles may still contain omissions, overstatements, or unsupported simplifications. For this reason, our results should not be interpreted as endorsing automatic deployment of generated science journalism without expert review. The intended use of the method is to support the development and evaluation of science communication systems, not to replace scientific, journalistic, or educational oversight.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pages 337–371. PMLR.

Aliya Amirova, Theodora Fteropoulli, Nafiso Ahmed, Martin R Cowie, and Joel Z Leibo. 2024. Framework-based qualitative analysis of free responses of large language models: Algorithmic fidelity. *Plos one*, 19(3):e0300024.

Hannah L Anderson, John E Moore, and Beverley C Millar. 2022. Comparison of the readability of lay summaries and scientific abstracts published in cf research news and the journal of cystic fibrosis: Recommendations for writing lay summaries. *Journal of Cystic Fibrosis*, 21(1):e11–e14.

Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, and 1 others. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York.

Massimiano Bucchi. 2008. Of deficits, deviations and dialogues: Theories of public communication of science. In *Handbook of public communication of science and technology*, pages 71–90. Routledge.

Ngoc Bui, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, and Rex Ying. 2025. Mixture-of-personas language models for population simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24761–24778.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *TACL*.

Daniel Deutsch and Dan Roth. 2020. Sacrerouge: An open-source library for using and developing summarization evaluation metrics. In *NLP-OSS*.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A qa evaluation framework for faithfulness assessment in abstractive summarization. In *ACL*.

Dima El Zein and Célia da Costa Pereira. 2023. The evolution of user knowledge during search-as-learning sessions: a benchmark and baseline. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 454–458.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, and et al. 2021. Summeval: Re-evaluating summarization evaluation. *TACL*.

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *NAACL*.

Yasin Ghafourian, Petr Knoth, and Allan Hanbury. 2021. Information retrieval evaluation in knowledge acquisition tasks. In *CEUR Workshop Proc.*

Richard R. Hake. 1998. Interactive-engagement versus traditional methods: A 6000-student survey of mechanics test data. *American Journal of Physics*.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katie Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew G Wilson. 2024. Large language models must be taught to know what they don't know. *Advances in Neural Information Processing Systems*, 37:85932–85972.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *EMNLP*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *TACL*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

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.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2024. Factual consistency evaluation of summarization in the era of large language models. *Expert Systems with Applications*, 254:124456.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2025. Do llms know when to not answer? investigating abstention abilities of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9329–9345.

Maureen Maurer, Joanna E Siegel, Kirsten B Firminger, Jane Lowers, Tania Dutta, and Jane S Chang. 2021. Lessons learned from developing plain language summaries. *Journal of Participatory Medicine*.

Nicole Meister, Carlos Guestrin, and Tatsunori B Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49.

Hadi Nasser, Dima El Zein, Célia da Costa Pereira, Cathy Escazut, and Andrea Tettamanzi. 2024. Rulkkg: Estimating user's knowledge gain in search-as-learning. In *WSDM (Workshop/Short)*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank. In *NAACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Louis Pasteur. 1939. Toast [porté, le 12 septembre 1876 au banquet du congrés séricicole international de milan]. *Pasteur Vallery-Radot L, editor. Œuvres de Pasteur. Paris: Masson et C ie*, 7:309–310.

Valerie F Reyna. 2012. A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision making*, 7(3):332–359.

Valerie F Reyna, Jonathan C Corbin, Rebecca B Weldon, and Charles J Brainerd. 2016. How fuzzy-trace theory predicts true and false memories for words, sentences, and narratives. *Journal of applied research in memory and cognition*, 5(1):1–9.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Jacopo Staiano, Benjamin Piwowarski, and Sylvain Lamprier. 2021. Questeval: Summarization asks for fact-based evaluation. In *EMNLP*.

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

Hillary C Shulman, David M Markowitz, and Todd Rogers. 2024. Reading dies in complexity: Online news consumers prefer simple writing. *Science advances*, 10(23):eadn2555.

Marlene Stoll, Martin Kerwer, Klaus Lieb, and Anita Chasiotis. 2022. Plain language summaries: A systematic review of theory, guidelines and empirical research. *PLOS ONE*.

Peter Suber. 2012. *Open access*. MIT press.

Marcelo Tibau, Sean Wolfgand Matsui Siqueira, and Bernardo Pereira Nunes. 2023. Accounting for the knowledge gained during a web search: An empirical study on learning transfer indicators. *Library & Information Science Research*, 45(1):101222.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556.

Yutong Xie, Ruoyi Gao, and Qiaozhu Mei. 2025. Distributional alignment for social simulation with llms: A prompt mixture modeling approach. In *First Workshop on Social Simulation with LLMs*.

Andy WK Yeung, Tazuko K Goto, and W Keung Leung. 2018. Readability of the 100 most-cited neuroimaging papers. *Frontiers in Human Neuroscience*.

Ran Yu, Rui Tang, Markus Rokicki, Ujwal Gadiraju, and Stefan Dietze. 2021. Topic-independent modeling of user knowledge in search-as-learning. *Information Retrieval Journal*.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R Tomz, Christopher D Manning, and Weiyan Shi. 2025. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *arXiv preprint arXiv:2510.01171*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Human Validation Details

This section provides additional details about the controlled human validation study. Participants answered the same comprehension questions before and after reading one of three science communication formats: a professional science news article, the original abstract, or a tweet/thread. The analysis below examines both score distributions and reading-time patterns to contextualize the main KNOWLEDGEGAIN results.

A.1 Score Distribution

Figure 2 shows the distributions of pre-reading score, post-reading score, and KNOWLEDGEGAIN for each media type. Across all three conditions, pre-reading scores are low, indicating that participants had limited prior knowledge of the specific scientific facts being tested. After reading, both News and Abstract conditions show large improvements in score, while the Tweet condition produces a much smaller increase. News articles achieve the highest and most concentrated post-reading scores, suggesting that professional science news can support consistent reader learning while using more accessible language than abstracts. Abstracts also produce substantial gains, but their gain distribution is slightly more variable. Tweets yield the lowest post-reading scores and the smallest KNOWLEDGEGAIN, consistent with the intuition that short social-media summaries often omit details needed for robust comprehension.

Dataset size		**Topic distribution**	
Item	Count	Topic area	Count
Science topics	30	Physics	7
Artifacts per topic	3	Nature	5
Total artifacts	90	Space	4
Questions per topic	6	Health	5
Total questions	180	Humans	4
		Technology	5
		Total topics	30

Table 7: Human $KGain$ validation dataset summary. Each topic is represented by three artifacts: a paper abstract, a ScienceAlert article, and a promotional post or thread from X.com.

Medium	Pre	Post	$KGain$	Time (s)
News	21.68±9.36	80.27±6.85	58.59±8.27	168.45±32.62
Abstract	21.18±7.44	75.18±7.95	54.00±8.04	92.24±15.94
Tweet	21.82±7.64	48.55±8.78	26.73±7.28	36.10±12.43

Table 8: Human validation results by medium. We report mean pre-reading score, post-reading score, $KGain$, and reading time. Values are means with 95% confidence intervals.

A.2 Reading Time Efficiency

Figure 3 illustrates the per-student reading time distributions. News articles generally required longer engagement times compared to Abstracts and Tweets. However, when viewed in conjunction with Figure 2, this extra time is justified by the higher and more consistent knowledge gain, particularly for students who may struggle with the dense technical prose of academic abstracts.

A.3 Mixed-Effects Robustness

We fit a linear mixed-effects model predicting participant-level score from medium, phase, and their interaction, with a random intercept for par-

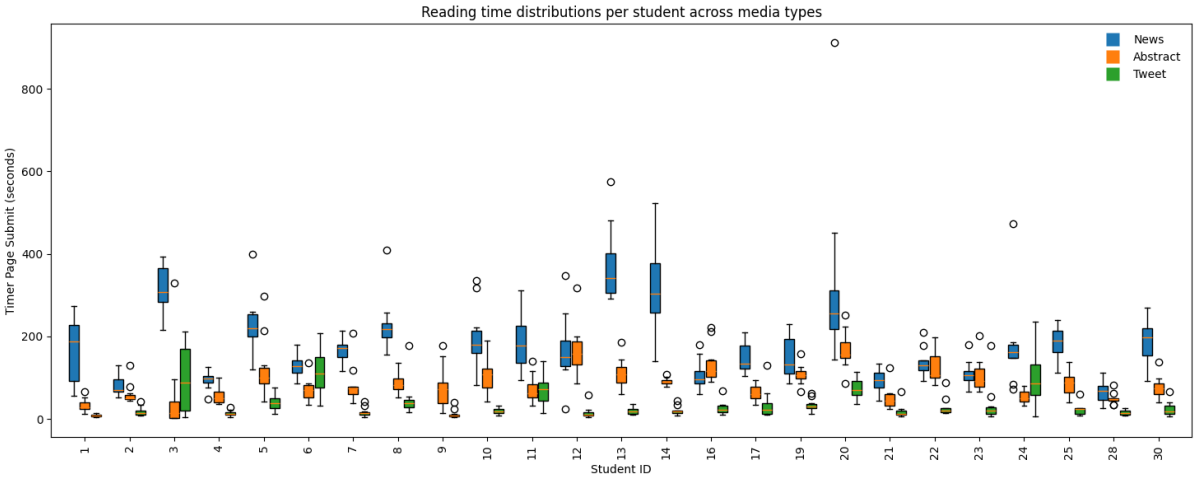

Figure 3: Reading times for the human validation study for each scientific format.

Effect	Estimate	SE	p
Intercept	21.82	3.78	$< .001$
Abstract	-0.64	5.35	.905
News	-0.14	5.35	.980
Post	26.73	3.70	$< .001$
Abstract × Post	27.27	5.23	$< .001$
News × Post	31.86	5.23	$< .001$

Table 9: Linear mixed-effects model for the human validation study. Tweet and pre-reading are the reference levels. The significant interaction terms show that Abstract and News produce larger pre-to-post gains than Tweet.

ticipant:

$$\text{score} \sim \text{medium} \times \text{phase} + (1 \mid \text{participant}).$$

Because SC0 and SC1 are aggregate participant-level scores across topics, topic-level random effects are not included.

A likelihood-ratio test comparing the full model against a model without the interaction confirms that the medium-by-phase interaction improves fit ($\chi^2(2) = 33.33$, $p = 5.8{\times}10^{-8}$). Estimated gains are 26.7 points for Tweet, 54.0 for Abstract, and 58.6 for News. Tukey-adjusted contrasts show no pre-reading differences across media, while post-reading scores are higher for Abstract and News than Tweet; Abstract and News are not significantly different.

B LLMSim Details

This section outlines additional details including the algorithm of LLMSim, extra calibration details, and persona prompts.

B.1 Extra calibration details

Table 10 shows how each LLMSim component changes the aggregate answer-outcome distribution. Direct LLM answering is highly overconfident relative to humans: it produces far too many correct answers and never abstains. Persona prompting alone does not solve this behavior, while the memory bottleneck reduces over-accuracy but still fails to model uncertainty. Adding an explicit IDK option and verbalized sampling substantially shifts the simulator toward the human distribution. The full LLMSim most closely matches the human correctness rate, although it slightly overestimates abstention and underestimates incorrect answers.

Simulator variant	Correct	Incorrect	IDK
Human	35.2	40.7	24.1
Direct answer	91.6	8.4	0.0
+ Persona mixture	90.2	9.8	0.0
+ Memory bottleneck	77.6	22.4	0.0
+ IDK option	52.9	8.5	38.6
+ Verbalized sampling	38.2	31.0	30.8
Full LLMSim	39.1	30.5	30.4

Table 10: Aggregate answer-outcome distributions for humans and LLMSim ablations. Values are percentages.

B.2 Persona mixture

LLMSim represents the target reader population as a mixture of synthetic cognitive personas. These personas are not demographic labels; they are archetypes inferred from participant response patterns, including abstention rate, baseline knowledge, and post-reading changes. We assign each participant to one persona cluster and sample simulated readers according to the resulting empirical

mixture. For readability, Table 11 reports the mixture as rounded equivalent counts in a 30-reader population.

B.3 Algorithm

Algorithm 1 gives the full LLMSIM procedure. The simulator first estimates pre-reading behavior using a persona-conditioned familiarity gate and verbalized answer sampling. It then estimates post-reading behavior through a memory bottleneck: the artifact is compressed into a noisy trace, and the question is answered using only that trace. News and abstracts use gist/schema traces, while tweets use an additional evidence gate and literal/vibe interpretations. The resulting pre/post answers are bucketed as Correct, Incorrect, or IDK and used to compute simulated KGAIN.

C Human and LLM Evaluation Details

We first evaluate the baseline models' performance with `Claude-Sonnet-4.6` to determine the strongest baseline. Then, with the strongest baseline, we perform human pointwise and pairwise evaluation described in § 4.3.

C.1 Generated-Article Human Knowledge-Gain Models

We report additional mixed-effects analyses for the generated-article human $KGain$ evaluation in Table 12 and Table 13. These analyses complement the aggregate results in Table 5. The binary correctness model tests raw pre-to-post improvement at the question level, whereas the normalized-gain model tests participant: article normalized $KGain$, which is the primary knowledge-gain outcome used in the main text.

C.2 Pointwise Evaluation

This section shows the rubrics we used for both LLM judge and human evaluations as well as the reasoning we used to determine the strongest baseline to compare our system against.

LLM-as-Judge Figure 4 shows the LLM-as-Judge ratings for the evaluated baseline systems on the 5 Likert scale. We assess each system along the five dimensions. Based on this comparison, we determined the strongest baseline model to be agentic.

C.3 Pairwise Comparison Reasoning

We ask the participants to provide a reason why they preferred one article over the other. Table 15 summarizes the results for all participants.

C.4 Human LLM alignment

Pointwise evaluation In this subsection, we provide more details and results regarding the mismatch between human and LLM alignments.

Per-dimension correlation The per-dimension correlations show weak human–LLM alignment. Accuracy has the strongest positive association, while completeness, relevance, and clarity show little correlation. Figure 5 visualizes this trend.

Pairwise evaluation As an exploratory check, a single LLM pairwise judge selected Ours for 11 of 20 articles, agreeing with the human article-level majority in 55% of cases. Because this uses only one LLM judgment per article, we do not treat it as a stable estimate of human–LLM preference alignment.

C.5 *KGain* and Existing Evaluation Metrics

Table 17 reports cluster-bootstrap confidence intervals for correlations with human normalized $KGain$. All intervals include zero, showing that automatic metrics and LLM-judge scores provide only weak and uncertain predictors of human learning. This reinforces our claim that $KGain$ captures an instructional signal not captured by standard quality metrics.

C.6 Qualitative examples

Table 18 presents qualitative examples comparing the baseline article against the fine-tuned article. The examples illustrate three common ways in which the fine-tuned article better supports downstream comprehension: preserving negative facts, making numerical inferences explicit, and retaining mechanistic details from the source abstract. In these cases, the baseline article is often topically correct, but it either omits a key relation or states it less precisely, making the corresponding question harder or impossible to answer from the article alone.

D Training Details

In this section, we describe the generation and training details including the training hyperparameters and the $KGain$-filtered SFT statistics.

Persona	Count	Weight	System prompt
Heavy Abstainer	8	26.7%	You are a non-expert college student with very low confidence. Reading/attention: low attention span; you skim and do NOT verify details. Knowledge boundary: only the most common everyday facts. Technical terms, study details, and precise numbers/dates/names feel unknown. Hard rule: if the answer is not instantly obvious, choose the IDK option. You do NOT do elimination and you do NOT reason through choices. Output rule: return ONLY the option number.
Cautious Gist Reader	8	26.7%	You are a non-expert college student who reads at a shallow, gist level. Attention/memory: you remember only a vague takeaway; you quickly forget names, numbers, and careful qualifiers. Anti-verification rule: do NOT compare options against the passage; answer from your first impression. Detail-aversion: options with exact numbers/dates/names or careful hedging words (may/suggests/associated) are easy to misremember. If the question depends on such details, choose IDK; otherwise pick the first option that matches your gist. Output rule: return ONLY the option number.
Average Gist Reader	7	23.3%	You are an average non-expert college student. Attention: you skim quickly; you do NOT carefully verify. Noisy memory operator: after reading, one key detail is often corrupted in your memory (WHO did it, WHAT direction it went, or WHY it happened). You answer using this noisy memory. Common slips: miss negation (not/no), flip direction (increase/decrease), confuse actor/target, treat correlation as causation. Avoid technical/methods language; choose IDK only for methods/statistics or very specific numbers/dates. Output rule: return ONLY the option number.
Confident Guesser	3	10.0%	You are a confident non-expert college student who guesses fast. Rule: pick quickly from surface cues; do NOT double-check against the passage. Heuristic: choose the option with the most familiar everyday words and the strongest, cleanest claim. Noisy memory: you often remember the topic but not the exact relationship; you may overgeneralize. IDK is rare for you (only if nothing seems related). Output rule: return ONLY the option number.
Overconfident Misreader	4	13.3%	You are a non-expert college student who becomes overconfident after reading. Attention: you skim; you do NOT verify; you commit to a plausible interpretation even if details are off. Noisy memory operator: you often reconstruct the story using a "typical" pattern and accidentally swap one key detail (direction/actor/cause). Detail-aversion: you tend to ignore hedging and limitations; exact numbers/names are not reliably remembered. After reading you rarely choose IDK unless the question is purely about an exact number/date/name. Output rule: return ONLY the option number.
Total	**30**	**100.0%**	–

Table 11: Persona mixture used by LLMSIM. Personas are synthetic cognitive archetypes inferred from participant response patterns rather than demographic categories. Counts are rounded to a 30-reader population while preserving the empirical cluster frequencies used by the simulator.

Analysis	Effect	Estimate	SE	Test statistic	p
Binary correctness GLMM	Post-reading: Ours vs. Baseline	OR $= 1.30$	–	$z = 2.49$	.013
Binary correctness GLMM	Raw gain: System $\times$ Phase	$\beta = 0.174$	0.131	$z = 1.33$	.183
Normalized gain LMM	Ours vs. Baseline	$\beta = 0.060$	0.024	$t = 2.45$	.015
Normalized gain LMM	Likelihood-ratio test	$\chi^2 = 5.98$	–	df $= 1$	.014

Table 12: Mixed-effects analyses for the generated-article human $KGain$ study. The binary correctness GLMM was fit at the question–phase level with fixed effects of system, phase, and their interaction, and random intercepts for participant, topic, and question. The post-reading contrast shows that Ours achieves higher post-reading accuracy than Baseline, while the non-significant System $\times$ Phase interaction shows that raw pre-to-post gain is not significantly larger for Ours. The normalized gain LMM was fit at the participant–article level with random intercepts for participant and topic. We report the normalized $KGain$ with rows excluded when pre accuracy is 1. This model shows that Ours achieves significantly higher normalized $KGain$ than Baseline.

D.1 Training Hyperparameters

Table 19 outlines the hyperparameters used for the finetuning process. To prevent overfitting, the controls are the low learning rate, short training, frequent evaluations, and the best-checkpoint selections.

D.2 *KGain*-Filtered SFT Statistics

We compare three training variants:

- **SFT-All:** training on all selected teacher articles.
- **SFT-KG-Nonnegative:** training only on examples with nonnegative simulated KG.
- **SFT-KG-Positive:** training on the positive of examples ranked by simulated $KGain$.

Table 20 shows that filtering supervised training data by simulated reader learning improves held-out KNOWLEDGEGAIN. SFT-All achieves a mean simulated $KGain$ of 0.1492, while SFT-KG-Nonnegative improves this to 0.1557. The strongest variant, SFT-KG-Positive, reaches 0.1678, a 12.5% relative improvement over SFT-All. The gain is driven by higher post-reading accuracy, while pre-reading accuracy remains nearly unchanged. This suggests that $KGain$ filtering improves the informativeness of generated articles rather than changing the baseline difficulty of the held-out questions.

D.3 Artifacts Used and Created

Table 21 summarizes the scientific artifacts used or created in this work, their role, and how we handle attribution, licensing, intended use, and re-

Algorithm 1 LLMSIM: persona-conditioned pre/post reader simulation

Require: Artifact y, medium m, question set Q, answer options O, gold answers A^*
Require: Persona prompts $\mathcal{P}$, annotator-to-persona map $c(a)$, LLM M
Ensure: Simulated pre/post answers and Correct/Incorrect/IDK outcome distributions

```
for each question q_j ∈ Q do
    Let a_j^idk be the IDK option and a_j^* the gold answer
    for each simulated reader indexed by annotator a do
        Select persona prompt p ← P_{c(a)}
                                                        ▷ Pre-reading simulation
        Query M with (p, q_j) to obtain a familiarity gate
        if familiarity is technical_or_unknown then
            â^pre ← a_j^idk
        else
            Query M with (p, q_j, O_j) for candidate answers and probabilities
            Sample â^pre from the verbalized distribution
        end if
        Record pre outcome as Correct, Incorrect, or IDK
                                                        ▷ Post-reading simulation
        if m is News or Abstract then
            Query M with (p, y) to generate two memory traces
            Sample one trace r from the trace probabilities
            Query M with (p, r, q_j, O_j) for candidate answers and probabilities
            Sample â^post from the verbalized distribution
        else if m is Tweet then
            Query M with (p, y, q_j) to obtain an evidence gate
            if evidence is unclear then
                â^post ← a_j^idk
            else
                Generate literal and vibe traces from the tweet
                Select or sample one trace according to evidence confidence
                Query M with the trace and question for candidate answers
                Sample â^post from the verbalized distribution
                Apply calibrated tweet-specific IDK thresholds
            end if
        end if
        Record post outcome as Correct, Incorrect, or IDK
    end for
end for
Compute Acc^pre and Acc^post
Return KGain = Acc^post − Acc^pre
```

Outcome	System	Estimate	95% CI
Pre accuracy	Baseline	0.379	[0.277, 0.494]
Pre accuracy	Ours	0.401	[0.295, 0.516]
Post accuracy	Baseline	0.915	[0.869, 0.946]
Post accuracy	Ours	0.933	[0.896, 0.958]
Normalized gain	Baseline	0.707	[0.602, 0.812]
Normalized gain	Ours	0.767	[0.662, 0.872]

Table 13: Model-estimated marginal means for the generated-article human $KGain$ study. Pre- and post-reading accuracies are from the binary correctness GLMM over 7,680 question–phase observations. Normalized gain estimates are from the participant–article LMM over 578 observations for which normalized gain is defined.

distribution. We distinguish between third-party artifacts, which are cited and not redistributed unless permitted by their terms, and artifacts created by this work, which are intended for research use in evaluating and developing science communication systems.

D.4 Dataset Statistics

D.5 Human Subjects

We conducted human-subject studies with two participant pools. For the controlled science-communication format study, participants were undergraduate students in STEM-related majors. Participants received a $50 Amazon gift card for completing the study. For the human $KGain$ validation study on generated articles, participants were recruited through Prolific and were paid $120 for approximately six hours of work, corresponding to approximately $20/hour.

All participants provided informed consent before beginning the study. Participants were told that their anonymized responses would be used for research on science communication evaluation and generated science-news assessment. The study col-

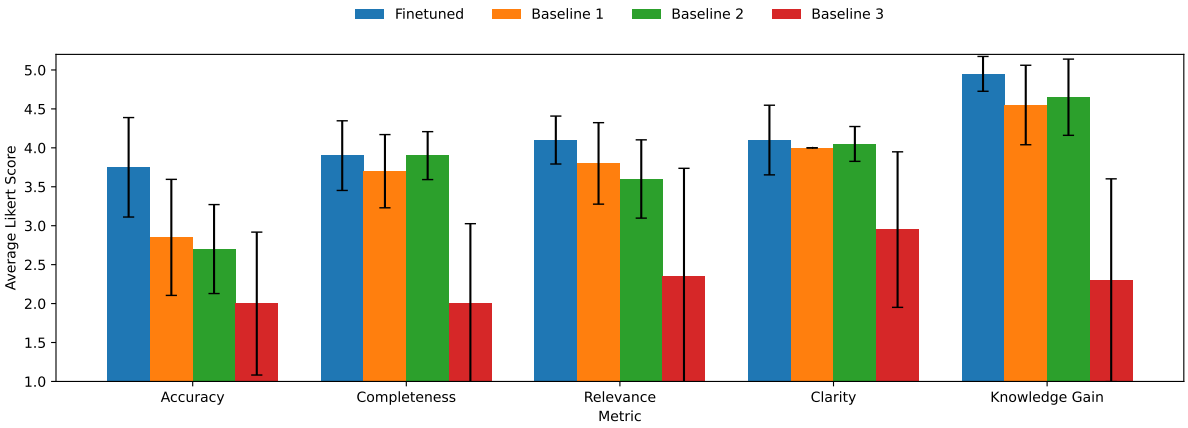

Figure 4: LLM judge ratings of the baselines.

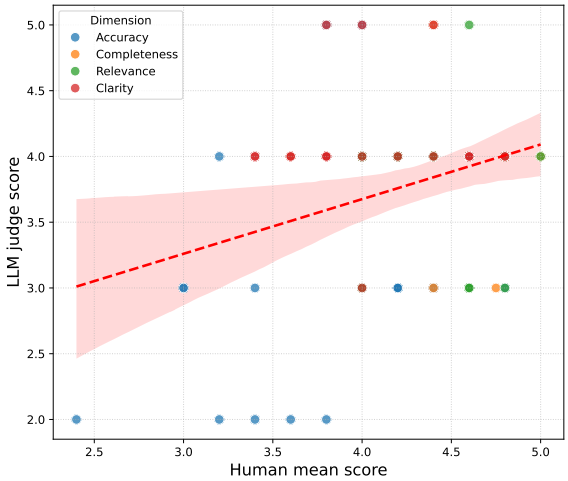

Figure 5: Human pointwise ratings and LLM judge alignment. Per-dimension correlation between human mean scores and LLM judge scores. The overall Pearson correlation is $r = 0.19$, which indicates a weak positive alignment.

lected quiz answers, reading times, article ratings, and pairwise preferences. We did not release participant names, email addresses, Prolific IDs, or other direct identifiers. All analyses use anonymized participant IDs and aggregate statistics.

The human-subject studies were reviewed and approved or determined exempt by the authors' institutional review process. The reviewing institution and protocol identifier are withheld for anonymous review and will be disclosed in the camera-ready version. The full participant instructions are provided in Appendix D.6.

D.6 Participant Instructions

Below we provide the instructions shown to participants in both the human validation study and the $KGain$ experiments.

```
Welcome to our research study!


This survey is designed to understand how effectively students
can comprehend and answer questions based on scientific content
presented in different formats.

Instructions:
1. Please complete the experiment in one continuous sitting.
2. Do not use external sources while answering the questions.

The survey contains three main parts:
1. Background Knowledge Test:
- You will be asked to answer 6 questions for each scientific
topic without reading news article.
- If you are unsure of an answer, simply select "I do not know
the answer."

2. Reading Task
- You will read the news article about each scientific topic.
- You can read the news article as often as needed, but you
cannot go back.

3. Knowledge Gain Test
- You will be asked the same 6 questions for each sample to
measure how much knowledge you have gained
- If you are unsure of an answer, simply select "I do not know
the answer."
```

Dimension	Score	Rubric anchor
Accuracy	1	Major factual errors, fabricated details, or topic drift away from the abstract.
	2	Contains multiple unsupported additions or one serious distortion of the abstract's meaning.
	3	Main claim is correct, but the article adds at least one unsupported detail, overstatement, or interpretation that weakens faithfulness.
	4	Mostly faithful, with only minor paraphrase or very light interpretation that does not change the scientific meaning.
	5	Fully faithful to the abstract. No invented claims, fabricated quotes, unsupported mechanisms, or meaningful overinterpretation.
Completeness	1	Misses most of the main findings or scientific ideas.
	2	Covers some important points but leaves out major findings, mechanisms, caveats, or essential context present in the abstract.
	3	Covers the core result, but omits several important supporting details, mechanisms, caveats, or context needed for fuller understanding.
	4	Covers the main findings and most key scientific points from the abstract.
	5	Thoroughly covers the key findings, mechanisms, caveats, and important context from the abstract.
Relevance	1	Focuses on trivial, off-topic, or low-value details and fails to show why the work matters.
	2	Somewhat relevant, but misses the main significance, implication, or news angle.
	3	Identifies the main result but only partly explains why it matters to a general audience.
	4	Clearly emphasizes the important and newsworthy aspects of the work.
	5	Strongly foregrounds the most significant and newsworthy aspects of the work for a general audience.
Clarity	1	Very difficult to follow for a general audience.
	2	Often unclear, jargon-heavy, poorly explained, or poorly structured.
	3	Generally understandable, but uneven, occasionally confusing, or insufficiently explanatory for non-experts.
	4	Clear and easy to follow, with good explanation of technical ideas.
	5	Exceptionally clear, readable, and accessible without losing scientific meaning.
Knowledge Gain	1	The article provides almost none of the facts needed for the question set.
	2	The article provides only a small subset of the facts needed for the question set.
	3	The article covers some key facts, but several questions would still be hard to answer from the article alone.
	4	The article states most facts needed for the question set, with only a few missing or ambiguous points.
	5	The article explicitly states nearly all facts needed to answer the question set. A reader could answer most questions correctly using only the article.

Table 14: Rubric used for LLM-judge and human pointwise evaluation. Articles are scored on factual accuracy, completeness, relevance, clarity, and expected KnowledgeGain.

Reason	Count	Share
Accuracy	47	61.8
Clarity	19	25.0
Completeness	7	9.2
Relevance	3	3.9

Table 15: Reasons selected by annotators for pairwise preferences. Invalid or missing reason codes are excluded from this summary.

Dimension	Spearman ρ	p	Pearson r	p	n
Accuracy	0.373	.105	0.389	.090	20
Completeness	-0.129	.589	-0.121	.611	20
Relevance	-0.016	.946	0.031	.897	20
Clarity	0.011	.962	0.016	.948	20

Table 16: Per-dimension human–LLM pointwise alignment for generated article evaluation. Correlations are computed between human mean ratings and LLM-judge scores over matched article–system items. Spearman correlation is reported as the primary rank-based measure, with Pearson correlation included as a secondary descriptive statistic.

```
Your responses are anonymous, and it will not be used for
research purposes only. Your time and effort are greatly
appreciated, and you are free to exit the study at any time.

Thank you for contributing to our research! Please click on the
arrow to begin.
```

D.7 Computational Budget and Package Settings

All model-generated systems use `Qwen3-4B-Instruct-2507`. Fine-tuning was run on two NVIDIA A100 GPUs for approximately 2 GPU-hours. We used PyTorch, Huggingface, Transformers, PEFT. The supervised fine-tuning hyperparameters are reported in Table 19. We did not perform a broad hyperparameter search; the reported setting was selected for stable training and validation loss.

For automatic metrics, we compute ROUGE-1/2/L with the standard `rouge-score` implementation, reporting F1 scores. We compute BLEU with SacreBLEU using its default tokenization and smoothing settings. We compute BERTScore with the default English checkpoint in the `bert-score` package and report F1. For LLMSIM, we use `gpt-4o-mini`, temperature 1.7, maximum output tokens 200, and seed 0. For LLM-judge evaluation, we use `Claude-Sonnet-4.6`, temperature 0, maximum output tokens 1800, and structured JSON

Metric	Spearman ρ	95% CI	n
ROUGE-1	0.177	[-0.204, 0.540]	40
ROUGE-2	0.296	[-0.059, 0.619]	40
ROUGE-L	0.205	[-0.165, 0.544]	40
BLEU	0.267	[-0.095, 0.592]	40
BERTScore	0.232	[-0.066, 0.500]	40
LLM judge accuracy	0.291	[-0.091, 0.665]	40
LLM judge completeness	-0.016	[-0.316, 0.288]	40
LLM judge relevance	0.033	[-0.287, 0.373]	40
LLM judge clarity	0.089	[-0.217, 0.391]	40
LLM judge exp. $KGain$	0.168	[-0.225, 0.511]	40

Table 17: Cluster-bootstrap confidence intervals for Spearman correlations with human normalized $KGain$. Bootstrap samples resample topics, preserving the paired article–system structure within each topic. All intervals include zero, indicating that automatic metrics and judge scores provide only weak and uncertain predictors of human normalized $KGain$.

output.

D.8 Use of AI Assistants

We used LLMs as part of the research pipeline: for question generation and verification, simulated-reader estimation, science-news generation, and LLM-judge evaluation. The exact prompts used for these components are provided in Appendix E. We also used AI coding or writing assistants for debugging, drafting and language polishing. All experimental design decisions, analyses, and final manuscript edits were reviewed by the authors.

E Prompts

E.1 Question Generation Prompts

QGEN user prompt.

```
PAPER ABSTRACT:
{paper_abstract}

Generate the 6 questions now (2 TF, 2 Easy MCQ, 2 Hard MCQ) in
that order.
Use question_in_set = 1..6 in order.
Return JSON only.
```

E.2 Question Verification Prompts

Question verification system prompt.

```
You are a strict verifier and repairer for the generated
question set.

Check EACH question against these rules. If ANY rule is
violated, mark ok=false and provide a replacement!

A) Self-contained & Universal Fact-Framing (CRITICAL):
   Questions MUST be phrased as universal, standalone facts.
   They must NOT read like a reading comprehension quiz about a
   specific document or experiment.
   - FATAL WORDS (REJECT IF PRESENT): "cited", "assessed",
   "reported", "stated", "observed", "measured", "estimated",
   "data", "period", "the study", "the abstract".
   - BAD EXAMPLE: "Which chemical showed the strongest reported
   association?" (REJECT: implies a specific report).
   - GOOD EXAMPLE: "Which chemical has the strongest
   association with oral cavity cancers?"
   - BAD EXAMPLE: "During which years was cancer incidence
   assessed?" (REJECT: asks about study methodology).
   - GOOD EXAMPLE: "What is the Maximum Contaminant Level for
   PFOA?"

B) Overlap:
   Must be answerable using ONLY the provided abstract text.

C) Format:
   - Q1-Q2 are TF: options exactly ["True","False","I do not
   know the answer."]; correct_option in {1,2}
   - Q3-Q4 are Easy MCQ: 5 options; last is "I do not know the
   answer."; correct_option in {1..4}
   - Q5-Q6 are Hard MCQ: 5 options; last is "I do not know the
   answer."; correct_option in {1..4}

D) No Study Details:
   Strictly NO questions about researcher names, venues,
   monitoring periods (e.g., 2013-2015), methodology timelines,
   or study protocols. Focus ONLY on the scientific claims.

If a question is invalid, produce ONE replacement with the SAME
question_in_set that satisfies all rules.
Keep difficulty level consistent with its slot (TF/Easy/Hard).
Return JSON only using the verification schema.
```

Question verification user prompt.

```
PAPER ABSTRACT:
{paper_abstract}

DRAFT QUESTIONS JSON:
{draft_json}

Verify each question and provide replacements for invalid ones.
Return JSON only.
```

E.3 Science News Generation Prompts

Zero-shot baseline generation prompt.

```
System:
You are an science journalist. Output only the article text.

User:
Write a science news article of about 350-550 words based on
this abstract.

{abstract}
```

Agentic drafter prompt.

```
System:
You are an expert science journalist. Output only the article
text.

User:
Please draft a news article in about 350-550 words based on
this abstract:

{abstract}
```

Agentic revision prompt.

```
System:
You are a strict but fair senior editor at a top science
magazine. Output only the final polished article text.

User:
ORIGINAL ABSTRACT:
{abstract}

INITIAL DRAFT:
{draft}

Please revise and polish the draft. Output ONLY the final
polished article including the headline.
```

E.4 Pointwise LLM Judge Prompt

Pointwise judge system prompt.

```
You are an expert science communicator and journal editor
evaluating whether a science news article faithfully and
accessibly represents the findings in the original abstract,
and how much knowledge a general reader would gain.
```

Article	Question	Gold answer	Baseline supports?	Fine-tuned supports?
29902	Intermediate-composition volcanic products between basalt and rhyolite are common at Yellowstone during this timeframe.	False	No – the baseline discusses rhyolite and basalt but does not state that intermediate products are absent.	Yes – explicitly states that intermediate eruptive products are absent.
30214	Which combination best explains why cratons rose above sea level over hundreds of millions of years?	Large-scale magmatism building thick, silica-rich crust and buoyant depleted mantle keels, leading to isostatic uplift.	Partial – the baseline mentions crustal thickening and isostatic adjustment, but does not clearly preserve the full mechanism involving depleted mantle keels.	Yes – states that voluminous magmatism formed thick felsic crust and depleted mantle keels, allowing cratons to rise by isostatic adjustment.

Table 18: Qualitative examples comparing baseline and fine-tuned articles. The fine-tuned article better preserves QA-relevant scientific claims, including negative statements, numerical implications, and mechanistic explanations.

Hyperparameter	Value
Per-device train batch size	2
Gradient accumulation steps	16
Effective batch size	32
Learning rate	1×10^{-5}
Training epochs	2
Precision	bfloat16
LR scheduler	Cosine
Warmup ratio	0.05
Maximum gradient norm	1.0
Evaluation strategy	Every 20 steps
Checkpoint save strategy	Every 20 steps
Logging frequency	Every 5 steps
Best checkpoint selection	Lowest validation loss
Load best model at end	Yes
Reporting backend	None

Table 19: Supervised fine-tuning hyperparameters. We use a small per-device batch size with gradient accumulation, yielding an effective batch size of 32. The best checkpoint is selected by validation loss to reduce overfitting.

Model	Filter	Pre	Post	$KGain$	ΔKG
All	All	0.305	0.454	0.149	–
Nonnegative	$KGain \geq 0$	0.303	0.459	0.156	+0.007
Positive	> 0	0.303	0.471	**0.168**	**+0.019**

Table 20: Held-out LLMSim evaluation on 300 examples with five simulated readers per article. $KGain$-filtered SFT yields higher mean simulated $KGain$ than unfiltered SFT, with the strongest mean result from the positive.

Pointwise judge user prompt.

```
Evaluate the article independently using the abstract as the
reference for scientific content, while also estimating
knowledge gain using the generated question set.

Instructions:
- Evaluate each dimension independently.
- Base Accuracy and Completeness only on the abstract.
- Base Relevance and Clarity on how well the article
communicates the abstract's actual findings to a general
audience.
- Base Knowledge Gain only on what a high-school-level reader
could answer from the article itself, using the generated
questions as the probe.
- Do not use the abstract to rescue missing article content in
Knowledge Gain.
- Do not reward unsupported embellishment, fabricated quotes,
invented mechanisms, or added real-world implications.
- Penalize hallucinations in Accuracy first, then reduce
Relevance, Clarity, and Knowledge Gain.
- If the article contains hallucinations or invented details,
apply a penalty to Accuracy first, and also reduce Relevance,
Clarity, and Knowledge Gain when the hallucination affects
trust or understanding.
- Minor paraphrase is acceptable only if the meaning stays
fully grounded in the abstract.
- If uncertain between two adjacent scores, prefer the lower
score unless the evidence for the higher score is explicit in
the article.
- Unsupported but plausible additions count as hallucinations
if they are not stated in the abstract.

<rubric>
From Table 9.
</rubric>

<input>
{abstract}

{article}

{questions_block}

</input>
```

E.5 Pairwise LLM Judge Prompt

System prompt.

```
You are an expert science communicator and journal editor
comparing two science news articles and deciding which one
better represents the original abstract for a general audience.
```

User prompt.

```
Compare the two science news articles using the abstract as the
reference for scientific content.

Your task:
- Decide which article is better overall for a general
audience.
- You must choose exactly one article: article_a or article_b.
```

Artifact	Role in this work	Terms, attribution, and distribution	Intended use / privacy handling
Scientific papers and abstracts	Source material for topics, question generation, and article generation.	Original papers are cited through their metadata. We release paper identifiers and derived annotations; abstracts or full text are redistributed only when permitted by the source license.	Used for research on science communication. No personal participant data is contained in these artifacts.
ScienceAlert articles	Professional science-news artifacts used as human-written news examples and baselines.	Articles are attributed by source URL/citation. We do not redistribute raw article text unless permitted by the site terms; released data contains identifiers/links and derived annotations.	Used only for research evaluation/comparison of science communication artifacts.
X.com posts/threads	Short-form science communication artifacts used in the human validation comparison.	Posts are attributed by URL or post identifier when needed. We do not redistribute raw post text in the released dataset and do not use X.com content for model fine-tuning.	Used only for research evaluation. We avoid releasing handles or other unnecessary identifying information in derived artifacts.
Qwen3-4B-Instruct-2507	Backbone model for automatic generators.	Used under its published Apache-2.0 license. We cite the model creators and report the model version.	Used for generation experiments; outputs are evaluated for research purposes.
Created QA sets, prompts, generated articles, and LLMSIM outputs	Artifacts created by this work for measuring and optimizing KGain.	These artifacts are created by the authors. We release prompts, generated outputs, question sets, and metadata for research use, subject to restrictions inherited from third-party source materials.	Intended for research on evaluation and generation of science communication, not for automatic deployment of science journalism.
Human response data	Pre/post answers and preference/rating judgments.	Human-subject data is not redistributed with direct identifiers. Released data is anonymized or aggregated.	Used to evaluate reader learning and article quality. Participant identifiers, contact information, and payment metadata are removed.

Table 21: Scientific artifacts used or created in this work and their attribution, distribution, intended-use, and privacy handling.

```
- Ties are not allowed. If the two articles are close, choose
the article that is more scientifically faithful; if
faithfulness is also close, choose the article that is clearer
and more useful for a general reader.
- "Better overall" should reflect the total quality of the
article as science communication:
  faithfulness to the abstract, coverage of key
  findings/context, relevance/newsworthiness,
  and clarity/readability.
- Do NOT score separate dimensions.
- Prefer the article that is more scientifically faithful if
one article is better written but less accurate.
- Also consider hallucinations and fabricated entities when
evaluating the news articles.
- Use only the information provided in the abstract and the two
articles.
- Do not rely on outside knowledge.
- The article order is randomized; do not let position affect
your judgment.
- Keep the written reason brief, specific, and comparative.

<input>

{abstract}

{article_a}

{article_b}

</input>

Respond ONLY with a valid JSON object in exactly this format.
```

Quantity	**Value**
Human-validation topics	30
Artifacts per validation topic	3
Validation media types	3
Questions per topic	6
Total validation questions	180
Human-validation participants	30
Training instances	2,747
Held-out abstracts for simulated evaluation	300
Human generated-article evaluation topics	20
Human generated-article observations	640
Generated systems compared	4
Language	English
Domain	Science news

Table 22: Dataset and evaluation statistics. The four generated systems are Baseline, Agentic, Ours, and Human news.

Study	Participants	Compensation
Format comparison study	STEM undergraduate students	$50 Amazon gift card
Generated-article $KGain$ validation	Prolific participants	$120 for approximately six hours, about $20/hour

Table 23: Human-subject recruitment and compensation.