Information Research - Vol. 30 No. iConf (2025)

Improving scholarship accessibility with reinforcement learning

Haining Wang, Jason Clark, Hannah McKelvey, Leila Sterman, Gao Zheng, Zuoyu Tian, and Xiaozhong Liu

DOI: https://doi.org/10.47989/ir30iConf47530

Abstract

Introduction. A vast amount of scholarly work is published daily, yet much of it remains inaccessible to the general public due to dense jargon and complex language. We introduce a reinforcement learning approach that fine-tunes a language model to rewrite scholarly abstracts into more comprehensible versions.

Method. Our approach utilises a carefully balanced combination of word- and sentence-level accessibility rewards to guide the language model in substituting technical terms with more accessible alternatives, a task which models supervised fine-tuned or guided by conventional readability measures struggle to accomplish.

Analysis. We evaluate our model's performance through readability metrics, factual accuracy assessments and language quality measurements, comparing results against supervised fine-tuning baselines.

Results. Our best model adjusts the readability level of scholarly abstracts by approximately six US grade levels—in other words, from a postgraduate to a high school level. This translates to roughly a 90% relative improvement over the supervised fine-tuning baseline, while maintaining factual accuracy and high-quality language.

Conclusion. We envision our work as a step toward bridging the gap between scholarly research and the general public, particularly younger readers, and those without a college degree.

Introduction

Accessible language in science communication

At first glance, the daily publication of tens of thousands of scientific papers—many freely accessible through open science and open access initiatives—suggests few barriers to knowledge dissemination. However, two key facts challenge this perception and show that significant barriers remain. A recent survey by the US Department of Education found that more than half of US adults aged 16 to 74 (54%, or 130 million people) read at or below a sixth-grade level (Rothwell, 2020). Meanwhile, an analysis of the readability of biomedical research abstracts published from 1881 to 2015 found that scientific writing has become increasingly difficult to read over time (Plavén-Sigray et al., 2017). Even when intended to be accessible, scientific abstracts typically require a postgraduate level of reading comprehension due to jargon use and sentence structure (Wang and Clark, 2024). This discrepancy leaves a significant portion of the population—including young readers and adults without advanced degrees unable to fully engage with scientific works, even if these are made freely available online. The '*infodemic*' surrounding COVID-19 highlighted this issue: the urgent need for understandable information about the virus clashed with the complex presentation of scientific findings—leading many to turn to more digestible but less reliable narratives on social media (Wang et al., 2019; Islam et al., 2020; Calleja et al., 2021).

While the legal and medical fields have long been encouraged to use accessible language as a clear conduit for public engagement (Mazur, 2000; Petelin, 2010), momentum for the adoption of accessible language within scientific communities has been building, roughly since the start of the open science movement (Schriver, 2017). For instance, the National Institutes of Health (NIH) advocates for 'clear and simple' principles when communicating with audiences with limited health literacy, and the *Proceedings of the National Academy of Sciences of the United States of America* (PNAS) requires authors to submit a significance statement accessible to non-experts (Berenbaum, 2021; Pool et al., 2021). However, there are inherent conflicts between the specialised nature of communication among disciplinary peer scholars and the public-oriented dissemination of scientific findings. Even assuming that communicating scholarly works in plain language is possible, it will inevitably increase the communication cost among domain experts and create confusion at the more advanced levels, compared to the use of jargon and technical terms. In an era of increasingly specialized scientific research, this conundrum is not easily addressed by scientists or disseminators. Hence, given the current landscape, the widespread adoption of accessible language in scholarly works is unlikely in the near future.

In response, we propose addressing the need for communicating scientific findings to a broader audience by *rewriting scholarly abstracts with simpler words and grammar using language models.* Since readability is key to comprehending scholarship (Flesch, 1946; DuBay, 2004; Kerwer et al., 2021), we envision the resulting accessible narratives as paving the way for the 'last mile' of science, broadening access to scientific understanding and engagement, especially for younger readers and those without a university degree.

Challenges to effective simplification

Fine-tuning a language model using pairs of abstracts and their accessible versions is the *de facto* method for automating the rewriting of scholarly abstracts into more accessible versions (Xu et al., 2015; Goldsack et al., 2022; Joseph et al., 2023). Accordingly, we introduced the Scientific Abstract-Significance Statement (SASS) corpus (Wang & Clark, 2024), a dataset composed of paired abstracts and significance statements from diverse disciplines, with the latter targeting 'an undergraduate-educated scientist outside their field of specialty' (Berenbaum, 2021; Pool et al., 2021). Although the simplified abstracts generated from language models fine-tuned on the SASS corpus are approximately three grade levels more readable than the original abstracts, as measured by US grade-based readability scores (Wang & Clark, 2024, Sec. 6), the documents are still not sufficiently accessible; even the best models produce college-level texts. Additionally,

because the vocabulary used in significance statements is often just as complex as that found in the abstracts themselves (Wang & Clark, 2024, Sec. 3), the readability improvements are primarily due to shorter sentences, and technical terms remain inadequately addressed.

Alternatively, the optimisation of a language model can be guided by a chosen objective in an actor-critic manner (Ramamurthy et al., 2023). It is intuitive to choose an established document readability measure, such as the Automated Readability Index (ARI; see Section 4.2.2), to assess the overall readability of the outputs generated by the language model. However, we found that the optimisation of language models guided by ARI is highly unstable, often resulting in the production of seemingly more accessible versions that still contain many technical terms. Inspired by Riddell and Igarashi (2021), we decomposed the measurement of document readability into two distinct measures: one at the sentence level and one at the word level. We then prioritized word-level accessibility in the optimisation to encourage the model to use more accessible words instead of simply shortening sentences.

Contribution

Our work aims to serve as a bridge between scholarly works and the general public, particularly benefiting younger readers and those without a college degree.

1. We address the common challenges in science communication by rewriting scholarly abstracts at a high school reading level using a language model.

2. We identify the challenges language models face in properly addressing jargon and propose Reinforcement Learning from Accessibility Measures (RLAM) as a means to improve the models' use of accessible terms in their rewrites. RLAM-trained language models can significantly reduce the reading level of a scholarly abstract from a postgraduate level to a high school level, achieving a 3 grade-level reduction or about a 90% performance boost compared to models fine-tuned using the same corpus.

3. We observe systematic differences between reinforcement learning models guided by different rewards and conclude that disproportionate weights for sentence-level rewards contribute to unstable training and lower simplification quality.

Our code, model generations and training logs are available at https://github.com/ Wang-Haining/RLAM under a permissive licence.

Scientific abstract-significance statement (SASS) corpus

We used the scientific abstract significance statement (SASS) corpus in our experiments. This corpus is composed of 3,430 abstract-significance statement pairs derived from PNAS and divided into training (3,030 samples), validation (200 samples), and test sets (200 samples) (Wang & Clark, 2024). It covers a wide range of disciplines, ensuring diverse representation across various fields, as shown in Figure 1. Corpus statistics are shown in Table 1; refer to Section 4.2 for a detailed description of the measures.

Section	ARI ↓	F-K ↓	VOA	SL ↓	WA	WL ↓
Abstract	18.9 (2.8)	19.2 (2.4)	-0.43 (0.25)	25.4 (4.9)	12.0 (0.4)	5.3 (0.4)
Significance	18.1* (3.1)	18.6* (2.7)	-0.31* (0.26)	23.9* (5.3)	11.9* (0.4)	5.4* (0.4)

Table 1. Corpus statistics for the scientific abstract-significance statement (SASS) corpus. Metrics include ARI (Automated Readability Index), F-K (Flesch-Kincaid readability test), VOA (log ratio of proportion of words found in the VOA1500 vocabulary), SL (average sentence length and number of sentences), WA (word accessibility; log frequency per 1 billion tokens in English Wikipedia), and WL (average word length). Measures whose names are followed by a down arrow symbol (↓) indicate that lower values correspond to a more readable document. Numeric values in parentheses are the corresponding standard deviations. Paired t-tests were conducted for each metric comparing the abstracts and significance statements, with p-values adjusted using the Bonferroni correction for multiple comparisons. The observed differences in each of the measurements are statistically significant after adjusting for the grouped p-values at a significance level of 0.05.

We observed that significance statements are semantically coherent with their corresponding abstracts. The corpus statistics indicate that significance statements are more readable than abstracts, as shown by lower mean values in the Automated Readability Index (ARI) and Flesch-Kincaid readability test (F-K). This suggests that the SASS corpus can be useful in simplifying scholarly abstracts across diverse disciplines.

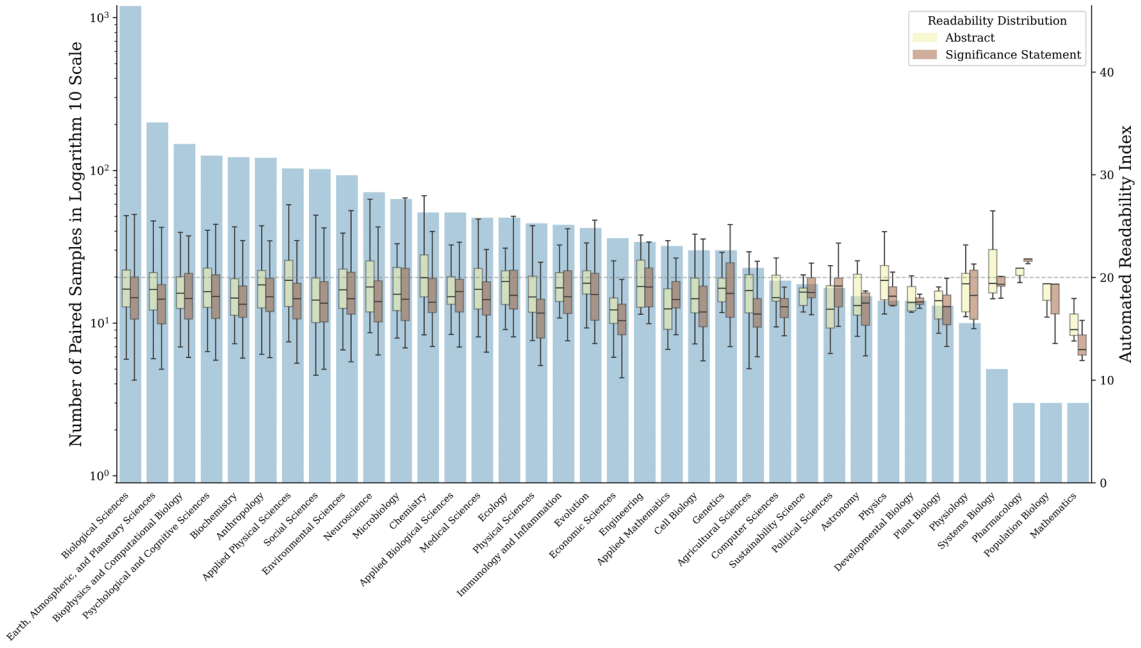

Figure 1. Discipline and readability distributions of abstracts and significance statements found in the training set of the Scientific Abstract-Significance Statement corpus. The count of paired samples in different disciplines is shown in blue bars on a log10 scale (disciplines with fewer than three samples are not shown). Readability is measured using the Automated Readability Index (ARI), which estimates the number of years of schooling required to understand a text. On average, abstracts have a readability slightly below 20 ARI, indicating a post-graduate level. Significance statements are generally more readable than their corresponding abstracts.

We also observed that word accessibility (i.e., log frequency per 1 billion tokens found in English Wikipedia) and average word length suggest that significance statements can be less accessible at

the word level than are their corresponding abstracts. Although the log ratio of words found in the VOA1500 vocabulary is slightly lower than in the corresponding abstracts, these 1,500 words are very basic and include a high proportion of function words. Considering that significance statements use approximately 1.5 fewer words on average, the increased use of VOA words may be a consequence of the higher use of function words to maintain grammaticality.

Reinforcement learning from accessibility measures

Language modelling via proximal policy optimisation

At the core of our approach is language modelling with Proximal Policy Optimisation (PPO) (Schulman et al., 2017) guided by two accessibility measures. A causal language model trained on large corpora can generate the next token based on the current sequence, which is useful in the context of reinforcement learning for developing a policy model that determines the most appropriate next token to maximize the expected return in terms of document readability.

The process begins with an input sequence s0 = (a_0, a_1,..., ai), where each a_i is from

a set of tokens W, and s_0 represents an abstract formatted in a simple template. The language model π_θ then generates a_0, a_1,..., aT_{-1} $\pi_\theta(\cdot \mid s_t)$, creating its accessible version until the maximum number of tokens T is reached, either due to the context length or an end-of-sentence token:

$$\pi_\theta(a_0, a_1, \ldots, a_{T-1}) = \prod_{t=0}^{T-1} \pi_\theta(a_t|s_t) \tag{1}$$

Our objective is to learn a policy model that, given an abstract, models the joint probability of tokens leading to a high reward in terms of accessibility while maintaining semantic coherence. Formally, this is expressed as:

$$J(\pi\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}(r(s_t, a_t) - \beta_{\mathrm{KL}}\mathrm{KL}(\pi\theta(a_t|\, s_t) \parallel \pi_{\theta_{\mathrm{SFT}}}(a_t|\, s_t)))\right] \tag{2}$$

Here, $J(\pi_\theta)$ represents the expected return when following policy π_θ. The reward $r(s_t, at)$ is estimated for each time step t in the trajectory τ = (s_0,a_0,s_1,a_1,..., sT_{-1},a_{T-1}), where s_t is the sequence of tokens at time t, formed as the concatenation of s_0 and the tokens a_0,a_1,...,a_{t-1}. This formula iteratively computes the rewards given the current sequence s_t and the token a_t chosen by the policy. The β_{KL}-weighted KL divergence term KL($\pi_\theta(a_t \mid s_t)$ // $\pi_{\theta\mathrm{SFT}}(a_t \mid s_t)$) is applied at every step of sequence generation to ensure the policy does not deviate significantly from the supervised fine-tuned model. This is crucial because, without such a constraint, the policy model might quickly learn to output whatever the reward model favours to maximize its return, which can lead to undesirable behaviours. For instance, the model might repeatedly output a frequent word ('is is is ...'), achieving a high reward based solely on accessibility measures but lacking meaningful content. Following Stiennon et al. (2020), β_{KL} is dynamically adjusted by targeting a specific KL divergence between π_θ and $\pi_{\theta\mathrm{SFT}}$ using a capped proportional controller in logarithmic space. The benefit of using a dynamic KL control, as opposed to a fixed one, is that it allows the model to adapt more flexibly to different stages of training, accommodating varying levels of KL divergence between the policy and SFT model.

In practice, we fine-tune the policy model in an actor-critic manner: while the policy model (the actor) generates sequences of tokens based on the current sequence s_t, the critic is an additional linear layer that takes the output of the language model's last layer and produces a scalar for time step t, estimating the expected cumulative reward of producing the token a_t, noted as $V_\rho(a_t)$. The problem is reduced to optimising at every step the expected cumulative reward of taking action a_t

in state s_t and following the policy π_θ thereafter ($V_\rho(s_{t+1}) = V_\rho(s_t,a_t)$), compared to the expected cumulative reward of being in state s ($V_\rho(s_t)$), termed as the advantage $d_t = V_\rho(s_{t+1}) - V_\rho(s_t)$.

We used the final reward of the entire generation (see Section 3.3) and back-propagate it through the sequence using Temporal Difference (TD) and Generalised Advantage Estimation (GAE) to estimate the advantage of each token:

$$d_t = \sum_{t'=t}^{T} (\gamma\lambda)^{t'-t} \left(r_T + \gamma V_\rho(s_{t'+1}) - V_\rho(s_{t'})\right) \quad (3)$$

where γ is the discount factor for future rewards, λ controls the bias-variance trade-off, and r_T is the final reward, which, in our case, is a linear combination of two accessibility measures. V_ρ is trained by minimizing the square error loss; see Equation 5.

We used the Proximal Policy Optimisation (PPO) (Schulman et al., 2017) clipped surrogate objective with importance sampling to more efficiently use offline samples to update the online policy. Importance sampling corrects for the discrepancy between the behaviour policy that generated the samples and the current policy by weighting the samples using the ratio of their probabilities under both policies. The PPO algorithm introduces a clipping mechanism to balance exploration and exploitation while preventing large, potentially harmful updates to the policy, see Equation 4. The whole RLAM algorithm is illustrated in Algorithm 1.

Algorithm 1 Training with Reinforcement Learning from Uncombined Accessibility Measures. The policy model is updated using the PPO clipped surrogate objective (Eq. 4), and the value model is updated by minimising a square-error objective (Eq. 5).

Input: initial policy model $\pi_{\theta_{SFT}}$, randomly initiated value head $V_{\rho_{init}}$, final reward function r_T for the last token, weighted by β_{KL} for KL divergence term; task prompts X; hyperparameters $\gamma,\lambda,$

$\pi_\theta \leftarrow \theta_{SFT}$, $V_\rho \leftarrow \rho_{init}$

for step = 1,..., M **do**

Sample a batch $\{s_0\}^n$ from X

Sample output sequences $\{a_0, a_1,..., a_{T-1}\}^n \sim \pi_\theta(\cdot \mid s_0)$ for each prompt s_0 in the batch ▷ Eq. 1

Compute final reward r_T for each sampled output sequence $\{a_0, a_1,...,a_{T-1}\}^n$. ▷Sec. 3.3

Distribute the final reward r_T to each token in the sequence through GAE ▷ Eq. 3

Compute advantages $\{d_t|s_t\}_{t=0}^{T-1}$, value targets $\{V_{targ}(s_t)\}_{t=0}^{T-1}$ for each sequence with V_ρ and compute KL divergence penalty $KL_t = KL(\pi_\theta(a_t \mid s_t) \,//\, \pi_{\theta_{SFT}}(a_t \mid s_t))$

for PPO iteration = 1,..., μ **do**

Update the policy model by maximizing the PPO clipped surrogate objective with KL penalty:

$$\theta \leftarrow \arg\max_{\theta} \frac{1}{n}\sum_{n=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\min\left(\frac{\pi_{\theta_{\text{online}}}(a_t|s_t)}{\pi_{\theta_{\text{offline}}}(a_t|s_t)}(A_t - \beta_{\text{KL}}\text{KL}_t),\right.$$
$$\left.\text{clip}\left(\frac{\pi_{\theta_{\text{online}}}(a_t|s_t)}{\pi_{\theta_{\text{offline}}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon\right)(A_t - \beta_{\text{KL}}\text{KL}_t)\right) \quad (4)$$

11: **end for**

12: Update the value model by minimizing a square-error objective:

$$\rho \leftarrow \arg\min_{\rho} \frac{1}{n}\sum_{n=1}^{n}\frac{1}{T}\sum_{t=1}^{T}(V_\rho(s_t) - V_{\text{targ}}(s_t))^2 \quad (5)$$

13: **end for**

14: **Output:** π_θ

Adaptive Kullback–Leibler controller

Following Ziegler et al. (2019, Sec. 2.2), we dynamically adjust β_{KL} to target a specific KL divergence value, $\text{KL}_{\text{target}}$, using a log-space proportional controller.

The update rule is:

$$\beta_{\text{KL}_{t+1}} = \beta_{\text{KL}_t}\left(\left(1 + K_\beta.\,\text{clip}\left(\frac{\text{KL}(\pi_\theta,\ \pi_{\theta_{\text{SFT}}}) - \text{KL}_{\text{target}}}{\text{KL}_{\text{target}}}, -0.2, 0.2\right)\right)\right)$$

where K_β is the proportional gain, set to 0.01.

Reward function

The reward function evaluates the overall quality of the output (s_T). After the initial failures of testing a traditional readability measure (i.e., ARI) as the criterion, we decided to use a balance of two accessibility measures: average sentence length in words and word accessibility, adopted from Riddell and Igarashi (2021), to guide the optimisation.

Word accessibility reward A word's accessibility is approximated by how frequently it appears in a large reference corpus. We chose the English Wikipedia corpus due to its domain similarity and applied a Moses tokenizer, yielding a vocabulary of 14.6 million types from a total of 3.6 billion tokens. If a token is among the most common 100,000 types, we report its frequency per billion tokens as its accessibility measure. Otherwise, we estimate its frequency using ridge regression with an $l2$-norm coefficient equal to 1.0. This model allows us to make serviceable estimates of the frequency of arbitrary tokens, including tokens that do not appear in the reference corpus. This model takes as input the token's length in Unicode code points, its byte unigrams, byte bigrams, and byte trigrams. The model estimates the token's log frequency per 1 billion tokens. We used the natural logarithm of frequencies per billion tokens as the measure of word accessibility. For example, the accessibility score for 'big' is 11.8, while 'colossal' scores 7.3. Despite being comparable in meaning, the model's production of the latter will receive fewer rewards. Coefficient β_{wa} is to control the scale of the credit given for word accessibility.

We have faithfully followed the experiment of Riddell and Igarashi (2021) with three differences. First, our reference corpus is the English Wikipedia, whereas the original study used the Common Crawl News corpus. Second, we did not discard duplicated sentences as Riddell and Igarashi (2021) did, because we found that sentence duplication is not common in Wikipedia. Third, the original study reported word *inaccessibility* scores by negating the logarithm of frequency per billion. We report *accessibility*, without negation, because it is more naturally suited to serve as a reward. Refer to Riddell and Igarashi (2021, pp. 1186–1187) for the training of the ridge regression.

Sentence length reward Sentence length is also determined by a Moses tokeniser, which preserves hyphenation and splits contractions. For the Moses rule-based tokeniser, we use the *sacremoses* Python package. We negate the value of sentence length for intuitive calculation of the rewards for optimisation.

Experiment setup

Training

We initialised the policy models π_θ by adopting the Gemma-2B checkpoint reported in Wang and Clark (2024) ($\pi_{\theta_{SFT}}$). The original Gemma-2B was trained on three trillion tokens, consisting of publicly available data as well as proprietary datasets comprising 'primarily English data from web documents' (Mesnard et al., 2024). The specific checkpoint we adopted was fine-tuned using the SASS corpus in a straightforward manner. It was chosen for its strong performance for simplification quality, faithfulness, and relatively compact size.

The two accessibility rewards were weighted as follows: the word accessibility reward was set to β_{WA} = 4.0, and we report two models with different β_{SL} values of 0.05 and 0.2. Because word accessibility is in logarithmic space, we subtracted 10 from the average word accessibility of the output and reset any values lower than 10 to 0 to keep them within a reasonable range. For adaptive control of the per-token semantic reward, we started with an initial β_{KL} = 0.2 and targeted a KL divergence of 8.0 nats during the training course, capping it in the range between 0.15 and 0.25. We used a micro batch size of 4, with each sequence used to run the PPO algorithm for 4 epochs using importance sampling, with gradient accumulation steps set to 4. We used a clip range of 0.2 for the policy gradient and value function estimation to ensure stability. The value function coefficient was set to 0.1. The optimisation used standard AdamW optimiser parameters (β_1 = 0.9, β_2 = 0.999, = 1×10^{-8}) (Kingma and Ba, 2015; Loshchilov and Hutter, 2017). The learning rate was fixed at 1×10^{-6}. The training was conducted on 2 H100 (80GB) GPUs using mixed precision training in bfloat16. The sampling temperature was set to 0.7 in the rollout phase. Following Huang et al. (2024), we assigned unfinished rollouts (indicated by the missing end of sequence token) with a fixed low score to encourage the model to generate complete simplified narratives. We also report a model trained as dictated by ARI (i.e., RLARI; described in Section 1.2), with all other parameters kept the same as those guided by the accessibility measures. We tested other hyperparameters for guiding the optimisation of RLARI in pilot studies but obtained similar results.

We performed multiple runs for each reinforcement learning process and selected the checkpoint to report based on a balance of semantic retention and ARI score obtained on the validation set. We observed that the readability of the generated text on the validation set often began to decline rapidly after plateauing for a while, typically accompanied by a surge in the standard deviation of average word accessibility among sentences. This signals that the language model was sacrificing language quality and semantic relevance for extra improvements in readability. Therefore, we reported the checkpoint immediately before such instability occurred.

Evaluation

We evaluated the simplified texts generated by the language models trained with the reinforcement learning framework using 200 abstracts from the SASS corpus test set for

simplification. Though advanced decoding methods might further refine the quality of the outputs, we used multinomial sampling with the temperature set to zero to intentionally produce the most deterministic outputs. This approach helped us better understand the modelling of accessible language and made it easier for us to identify potential quirks. We assessed the quality of the generated simplified abstracts both quantitatively and qualitatively. Quantitatively, we measured the generated texts based on their semantic retention and accessibility using established relevance measures as well as readability and accessibility metrics.

Semantic retention

BERTScore calculates the cosine similarity between each token in the candidate sentence and each token in the reference sentence using contextual embeddings from a pre-trained language model; its results align well with human judgement on semantic similarity evaluation (Zhang et al., 2019). It is not directly influenced by lexical overlap, making it more suitable for evaluating simplification systems than are metrics that rely on matching words, such as BLEU (Papineni et al., 2002). For our evaluation, we used embeddings from the 18th layer of a BERT-large-uncased model and reported the F1 score. This choice is based on prior findings indicating that the 18th layer yields a strong Pearson correlation (0.72) on the WMT16 To-English benchmark (Zhang et al., 2019).

Simplification and accessibility

Accessibility can be measured with respect to the overall simplification quality (SARI); readability (ARI and Flesch-Kincaid); and other straightforward document complexity measures, including average sentence length, word accessibility (i.e., log frequency per 1 billion tokens found in English Wikipedia), the log ratio of its proportion of VOA Special English words (1,517 types in total), and average word length.

SARI (System output Against References and against the Input sentence) is specifically designed to evaluate text simplification (Xu et al., 2016). It aims to measure how well a simplified text retains the original meaning while improving readability. SARI provides a balanced measure of how well a text simplification system performs by focusing on the necessary operations of adding, deleting, and retaining words.

ARI and Flesch-Kincaid readability tests assign a numerical score to text that reflects the US grade level required for comprehension. Lower scores (1–13) indicate content suitable for kindergarten through twelfth grade, with each score corresponding to a subsequent grade level. Scores in the range of 14–18 suggest college-level readability, ranging from first- to senior-year content appropriateness. Higher scores (19 and above) are associated with advanced college education. Both measures use average sentence length. Flesch-Kincaid uses syllables per word, while ARI uses characters per word for its linear combination with sentence length.

We harvested VOA Special English vocabulary comprising 1,517 unique words (VOA1500). We included VOA Special English Word Book Sections A-Z, Science Programs, and Organs of the Body hosted on Wikipedia (Wikimedia Foundation, 2024). We calculate the ratio of words that appear in the VOA1500 to those that do not, then report the natural logarithm of this ratio for each generated sample. Values above 0 indicate that the text contains more Special English words than non-Special English words, and a higher value indicates a greater presence of ‘easy’ words.

Language quality, faithfulness, and completeness

We manually examined 5% of all generated samples, corresponding to a randomly chosen subset of the test set from the SASS corpus. Each generation is annotated with respect to language quality, faithfulness, and completeness using a rubric of Good, Acceptable, and Poor. We focused on fluency and grammaticality and hand-picked both good and problematic examples when evaluating language quality. For the evaluation of faithfulness, we conduct close readings to assess the extent to which a simplified abstract remains factually faithful to the original narrative. If

uncertainty arises, we consult the corresponding manuscript, as our abstract simplification system must avoid producing misinformation. Completeness is also a key consideration, as it is essential to include the main findings and implications of the research for the general public, since this is the primary goal of scientific dissemination.

Findings and discussion

Quantitative assessment

Table 2 summarises the performance of Gemma-2B, tuned in different ways, when evaluated on the test set of the SASS corpus. The first scenario is the supervised fine-tuned baseline (SFT), which performs next-token prediction on the SASS corpus training set.

The second and third scenarios are reinforcement learning through PPO guided by ARI (RLARI) or accessibility measures (RLAM). We assessed the generation quality by considering both semantic retention and simplification, specifically using BERT score (BS), SARI, ARI, Flesch-Kincaid readability test (F-K), the log ratio of words in the VOA1500 vocabulary (VOA), sentence length (SL), word accessibility (WA), and word length (WL). A one-tailed paired t-test was conducted for each metric to compare observations between the reinforcement learning and supervised fine-tuning baselines, assuming improvement in document readability. Bonferroni correction was applied to each set of tests to maintain a family-wise significance level of 0.05.

Model	β_{SL}	β_{WA}	ARI ↓	F-K ↓	SARI	VOA	SL ↓	WA	WL ↓	BS
SFT	-	-	15.5 (3.0)	16.5 (2.6)	39.1 (5.0)	-0.26 (0.30)	20.6 (4.1)	11.9 (0.5)	5.2 (0.4)	0.64 (0.06)
RLARI	-	-	12.6* (2.9)	14.3* (2.5)	40.1* (4.8)	-0.17* (0.31)	16.4* (3.7)	12.0 (0.5)	5.0* (0.4)	0.64 (0.05)
RLAM	0.05	4.0	13.5* (2.8)	14.8* (2.4)	39.8 (5.1)	0.08* (0.29)	21.0 (4.2)	12.7* (0.5)	4.8* (0.4)	0.62 (0.06)
RLAM	0.2	4.0	12.5* (2.9)	14.0* (2.5)	39.8 (5.0)	-0.01* (0.32)	17.7* (3.4)	12.4* (0.5)	4.9* (0.4)	0.63 (0.05)

Table 2. Comparison of Gemma-2B's performance across three approaches: the supervised fine-tuned baseline (SFT), reinforcement learning guided by ARI (using an intermediate checkpoint before significant policy gradient instability was observed, RLARI), and reinforcement learning guided by two accessibility measures (RLAM). SFT was fine-tuned using the Scientific Abstract-Significance Statement (SASS) corpus. The columns labelled β_{WA} and β_{SL} pertain specifically to RLAM, where the rewards for average word accessibility and sentence length are balanced. The inference on the test split from SASS uses multinomial sampling. Metrics ARI, F-K, SARI, VOA, SL, WA, WL, and BS stand for Automated

Readability Index, Flesch-Kincaid readability test, log ratio of VOA1500 vocabulary, sentence length, word accessibility, word length, and BERTScore (F1), respectively. Measures followed by a down arrow symbol (↓) indicate that lower values are better. Numeric values in parentheses are the corresponding standard deviations. A paired two-tailed t-test was performed on observations of each measure between each model and the original abstracts. At a model-wise p-value of 0.05, measures that differ significantly from the SFT baseline are marked with an asterisk.

We observe that reinforcement learning models trained with different rewards exhibit a notable reduction in reading level, bringing abstracts down to high school levels. The model directly guided by ARI (RLARI) achieves an ARI of 12.6, while the most performant model guided by accessibility measures (RLAM, β_{SL} = 4.0 and β_{WA} = 0.2) reaches 12.5, both aligning with the readability level expected for individuals who have completed K-12 education (approximately ARI 13). However,

RLARI and RLAM models achieve these readability improvements in different ways. For RLAM models, better readability is achieved mostly through improved token-level accessibility. RLAM models show an increase in word accessibility from 0.5 to 0.8 compared to the supervised baseline. This increase in the natural logarithm suggests that words generated by RLAM models are, on average, 1.6 to 2.2 times more frequent in the English Wikipedia corpus than those generated by the SFT model. In comparison, RLARI's 0.1-unit increase in word accessibility, although observed, does not result in a statistically significant change in word frequency compared to the SFT model. Similarly, the log ratio of the VOA1500 vocabulary in the RLAM models shows a significant improvement, with log ratios ranging from −0.02 to

0.08. This implies that for every 100 non-VOA1500 (or more complex) words generated, RLAM models can produce approximately 99 to 106 VOA1500 basic words. In contrast, the SFT and RLARI models exhibit VOA log ratios of −0.26 and −0.17, respectively, indicating that for every 100 non-VOA1500 words generated, these models produce only around 77 (84) VOA1500 words. The average word length in characters for RLAM models ranges from 4.8 to 4.9, slightly shorter than RLARI's and outperforming SFT. Overall, the above evidence suggests that RLAM models achieve better readability by using more common, simpler, and shorter words.

On the other hand, the RLARI model achieves better readability by producing much shorter sentences, with only a marginal boost in word-level accessibility. The RLARI model has the shortest average sentence length of 16.4 words, significantly outperforming the SFT model. In comparison, a significantly shorter average sentence length is only observed when RLAM's sentence length reward coefficient (β_{SL}) exceeds 0.08. We also observed that, in pilot studies, if β_{SL} is set to a higher value, such as 0.5, the model's optimisation will collapse after only a few hundred steps, similar to what we consistently observed in the optimisation of RLARI models. The improvement in the RLARI model's word-level accessibility is inconsistent: while we observed significant gains in the VOA log ratio and word length, the word accessibility measure did not show statistically significant improvement after a Bonferroni correction. That said, although RLARI uses more VOA basic words and shorter words, their frequency in the English Wikipedia corpus is not high enough to result in a significant increase in word accessibility compared to the SFT baseline.

For semantic coherence with the human-written significance statements, the BERT scores from the reinforcement learning models are not significantly worse than those of the SFT model. This finding suggests that the generations from the reinforcement learning models are likely to remain semantically faithful and that the reinforcement learning process does not significantly degrade language quality, at least up to the point before unstable optimisation signals were observed. SARI scores from the reinforcement learning models are significantly, yet only marginally, better than those of the SFT baseline. Since SARI is a measurement of a combination of deletions, additions, and retention operations, a straightforward explanation is not available. However, the higher SARI scores confirm the greater simplification quality of the reinforcement learning models, which is likely due to a combination of simpler words and shorter sentences.

Qualitative analysis

We annotated 5% of the generated simplified abstracts from reinforcement learning models guided by ARI (RLARI) and accessibility measures (RLAM, with $\beta_{WA} = 4.0$ and varying β_{SL} values) with respect to language quality, faithfulness, and completeness, as shown in Figure 2. The test abstracts included three from biological sciences and one each from chemistry; mathematics; evolutionary biology; environmental sciences; ecology; economic sciences; and earth, atmospheric, and planetary sciences.

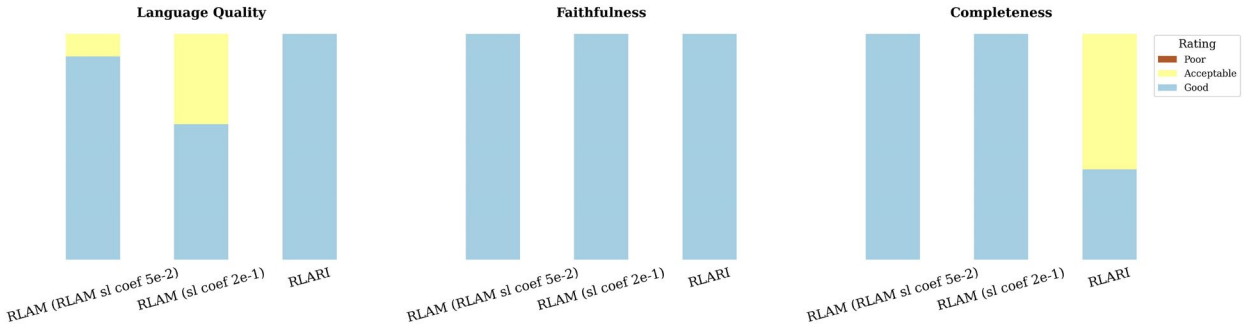

Figure 2. We annotated 5% of the generated outputs from reinforcement learning models guided by ARI (RLARI) and accessibility measures (RLAM, with $\beta_{WA} = 4.0$ and two different β_{SL} values).

Regarding overall quality, we found that reinforcement learning-trained models generally produced high-quality language. Compared to the SFT generations, RLAM outputs are often shorter and more semantically complete, due to the imposed token budget for newly generated content (241 tokens, the length of the longest significance statement in the training set). The main shortcoming is the presence of small trailing phrases that often deflate readability scores, such as '(PsycINFO Database Record),' '(show more),' and 'All rights reserved.' The first artefact is also found in the SFT model generations and is hypothesised to be carried over from the corpora that Gemma-2B was previously exposed to, as this pattern does not appear in the SASS corpus. An informal review of the remaining generations suggests that this phenomenon is amplified, appearing even in samples without such trailers in the SFT generations. The latter two artefacts were newly observed and are caused by the reinforcement learning processes. Where any of these issues appear, we annotate them as 'Acceptable,' even though the generations are otherwise fluent and grammatical. We also found that the proportion of these artefacts steadily increases as training continues, and they are usually found in only a subset of samples across different experiments.

In assessments of faithfulness, we did not find any models hallucinating in the generations we examined. However, in informal examinations, we did find that reinforcement learning checkpoints may hallucinate by generating simple but overly hedging or short expressions when over-optimised. Generations from RLARI-trained models often remain unfinished, but they retain the main gist of the abstract. Although we do not observe trailer phrases in the RLARI-generated texts like those found in the RLAM models, this issue frequently arises in other RLARI runs. Subsequent checkpoints often exhibit similar problems and tend to deteriorate rapidly once the model starts cutting corners. This typically occurs shortly before the training process completely fails. However, the reported checkpoint happens to miss this characteristic.

Conclusion

To improve the accessibility of scientific literature to the general public, we implemented reinforcement learning techniques to guide language models, extending beyond the traditional cross-entropy objective. Our study demonstrates that carefully balancing accessibility measures at the word and sentence levels can effectively guide Gemma-2B in simplifying scholarly abstracts, outperforming the supervised fine-tuning baseline by a large margin. This approach achieves these improvements without compromising language quality or faithfulness and mitigates the supervised fine-tuning model's tendency to overemphasise research implications. The best model trained using our method successfully adjusts the readability level of scholarly abstracts by approximately six US grade levels in other words, from a postgraduate to a high school level. Compared to the supervised fine-tuning model, the words generated by the model trained via our

approach are proven to be more common (1.6 to 2.2 times more frequent), easier (with more VOA basic words), and shorter (by 0.3 to 0.4 characters). This improvement addresses a key limitation of existing corpora, in which the target distribution (i.e., significance statements) often does not adequately prioritise the accessibility of word choice. We hope this work contributes to bridging the gap between scholarship and a broader audience, advancing the understanding and development of better simplification systems, and ultimately fostering a more informed and engaged society.

Acknowledgements

We gratefully acknowledge the support of the Institute of Museum and Library Services (No. RE-246450-OLS-20) and the National Social Science Fund of China (No. 23&ZD221). We also thank the organisers of the LIS Education and Data Science Integrated Network Group (LEADING), including Jane Greenberg, Erik Mitchell, Kenning Arlitsch, Jonathan Wheeler, and Samantha Grabus. Additionally, we are thankful to Coltran Hophan-Nichols and Alexander Salois from the University Information Technology Research Cyberinfrastructure at Montana State University for providing computational resources on the Tempest High Performance Computing System, Doralyn Rossmann for research support, and Deanna Zarrillo for early involvement in the project.

About the authors

Haining Wang is a doctoral candidate in Information Science at Indiana University Bloomington. His research focuses on natural language processing with applications in the humanities, social sciences, and biomedical sciences. He can be reached at hw56@iu.edu.

Jason A. Clark is a Professor and Head of Research Optimisation, Analytics, and Data Services (ROADS) at Montana State University Library. His research interests include machine learning, digital libraries, and the intersection of artificial intelligence with user experience, including algorithmic literacy and machine learning patterns. He can be contacted at jaclark@montana.edu.

Hannah McKelvey is an Associate Professor at Montana State University Library. Her research interests are focused on practice-based librarianship, focusing on electronic resource management, user discovery behaviour, and collection assessment through usage analytics. She can be contacted at hannah.mckelvey@montana.edu.

Leila Sterman is an Associate Professor at Montana State University Library. Her research interests focus on scholarly communication, including open access publishing, copyright, and institutional repositories. She can be contacted at leila.sterman@montana.edu.

Zheng Gao is currently a senior algorithm engineer at Ant Group. His research interests encompass machine learning, deep learning, and large language models, including LLM post-training and evaluation. He can be contacted at gao27@alumni.iu.edu.

Zuoyu Tian is an Assistant Professor at Macalester College. His research interests lie in computational linguistics and natural language processing, with a particular focus on applying computationally intensive methods to study language variation and change. He can be contacted at ztian@macalester.edu.

Xiaozhong Liu is an Associate Professor in Computer Science and Data Science at Worcester Polytechnic Institute. His research interests include natural language processing, text/graph mining, information retrieval/recommendation, metadata, and computational social science. He can be contacted at xliu14@wpi.edu.

References

Berenbaum, M. R. (2021). On COVID-19, cognitive bias, and open access.

Calleja, N., AbdAllah, A., Abad, N., Ahmed, N., Albarracin, D., Altieri, E., Anoko, J. N., Arcos, R., Azlan, A. A., Bayer, J., Bechmann, A., Bezbaruah, S., Briand, S. C., Brooks, I., Bucci, L. M., Burzo, S., Czerniak, C., De Domenico, M., Dunn, A. G., Ecker, U. K. H., Espinosa, L., Francois, C., Gradon, K., Gruzd, A., Gülgün, B. S., Haydarov, R., Hurley, C., Astuti, S. I., Ishizumi, A., Johnson, N., Johnson Restrepo, D., Kajimoto, M., Koyuncu, A., Kulkarni, S., Lamichhane, J., Lewis, R., Mahajan, A., Mandil, A., McAweeney, E., Messer, M., Moy, W., Ndumbi Ngamala, P., Nguyen, T., Nunn, M., Omer, S. B., Pagliari, C., Patel, P., Phuong, L., Prybylski, D., Rashidian, A., Rempel, E., Rubinelli, S., Sacco, P., Schneider, A., Shu, K., Smith, M., Sufehmi, H., Tangcharoensathien, V., Terry, R., Thacker, N., Trewinnard, T., Turner, S., Tworek, H., Uakkas, S., Vraga, E., Wardle, C., Wasserman, H., Wilhelm, E., Würz, A., Yau, B., Zhou, L., and Purnat, T. D. (2021). A public health research agenda for managing infodemics: Methods and results of the first who infodemiology conference. *JMIR Infodemiology*, 1(1): e30979.

DuBay, W. H. (2004). The principles of readability. Technical report, Impact Information, Costa Mesa, CA.

Flesch, R. (1946). *The Art of Plain Talk*. Harper & Row, New York, first edition.

Goldsack, T., Zhang, Z., Lin, C., and Scarton, C. (2022). Making science simple: Corpora for the lay summarisation of scientific literature. *arXiv preprint arXiv*:2210.09932.

Huang, S., Noukhovitch, M., Hosseini, A., Rasul, K., Wang, W., and Tunstall, L. (2024). The N+ implementation details of RLHF with PPO: A case study on TL; DR summarization. arXiv preprint arXiv:2403.17031.

Islam, A. N., Laato, S., Talukder, S., and Sutinen, E. (2020). Misinformation sharing and social media fatigue during COVID-19: An affordance and cognitive load perspective. Technological forecasting and social change, 159:120201.

Joseph, S., Kazanas, K., Reina, K., Ramanathan, V. J., Xu, W., Wallace, B. C., and Li, J. J. (2023). Multilingual simplification of medical texts. arXiv preprint arXiv:2305.12532.

Kerwer, M., Chasiotis, A., Stricker, J., Günther, A., and Rosman, T. (2021). Straight From the Scientist's Mouth—Plain Language Summaries Promote Laypeople's Comprehension and Knowledge Acquisition When Reading About Individual Research Findings in Psychology. Collabra: Psychology, 7(1):18898.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. In International Conference on Learning Representations.

Mazur, B. (2000). Revisiting plain language. Technical communication, 47(2):205–205.

Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Petelin, R. (2010). Considering plain language: Issues and initiatives. Corporate Communications: An International Journal, 15(2):205–216.

Plavén-Sigray, P., Matheson, G. J., Schiffler, B. C., and Thompson, W. H. (2017). Research: The readability of scientific texts is decreasing over time. eLife, 6: e27725.

Pool, J., Fatehi, F., and Akhlaghpour, S. (2021). Infodemic, misinformation and disinformation in pandemics: Scientific landscape and the road ahead for public health informatics research. In Public Health and Informatics, pages 764–768. IOS Press.

Ramamurthy, R., Ammanabrolu, P., Brantley, K., Hessel, J., Sifa, R., Bauckhage, C., Hajishirzi, H., and Choi, Y. (2023). Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In The Eleventh International Conference on Learning Representations.

Riddell, A. and Igarashi, Y. (2021). Varieties of plain language. In Mitkov, R. and Angelova, G., editors, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pages 1180–1187.

Rothwell, J. (2020). Assessing the economic gains of eradicating illiteracy nationally and regionally in the United States. Technical report, Barbara Bush Foundation for Family Literacy and Gallup. Accessed: 2024-07-12.

Schriver, K. A. (2017). Plain language in the US gains momentum: 1940–2015. IEEE Transactions on Professional Communication, 60(4):343–383.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33:3008–3021.

Wang, H. and Clark, J. (2024). Simplifying scholarly abstracts for accessible digital libraries using language models. In 2024 ACM/IEEE Joint Conference on Digital Libraries, JCDL '24, page 8, Hong Kong, China. ACM.

Wang, Y., McKee, M., Torbica, A., and Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. Social science & medicine, 240:112552.

Wikimedia Foundation (2024). VOA Special English Word Book. [Online; accessed 9January-2024].

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in current text simplification research: new data can help. Transactions of the Association for Computational Linguistics, 3:283–297.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. Transactions of the Association for Computational Linguistics, 4:401–415.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.