

Know Your Audience: The benefits and pitfalls of generating plain language summaries beyond the “general” audience

Tal August
Allen Institute for AI
Seattle, Washington, USA

Kyle Lo
Allen Institute for AI
Seattle, Washington, USA

Noah A. Smith
University of Washington & Allen Institute for AI
Seattle, Washington, USA

Katharina Reinecke
University of Washington
Seattle, Washington, USA

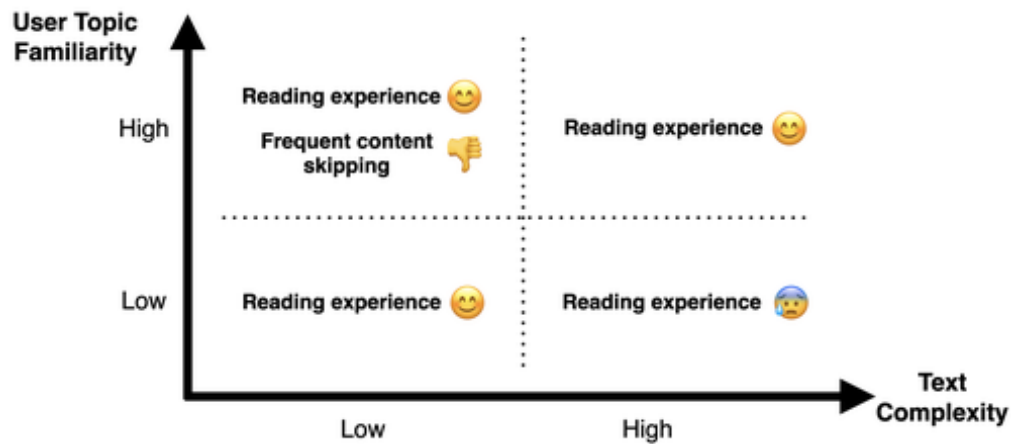


Figure 1: Simpler summaries were often the best reading experience for participants with little to no background in a scientific topic. However, readers with high topic familiarity, even those considered part of the general public (i.e., not a researcher), ignored more information in low complexity summaries while still reporting these simple summaries as equally engaging as high complexity ones. Our results provide guidance on generating plain language summaries for a wider range of general audiences.

ABSTRACT

Language models (LMs) show promise as tools for communicating science to the general public by simplifying and summarizing complex language. Because models can be prompted to generate text for a specific audience (e.g., college-educated adults), LMs might be used to create multiple versions of plain language summaries for people with different familiarities of scientific topics. However, it is not clear what the benefits and pitfalls of adaptive plain language are. When is simplifying necessary, what are the costs in doing so, and do these costs differ for readers with different background knowledge? Through three within-subjects studies in which we surface summaries for different envisioned audiences to participants of different backgrounds, we found that while simpler text led to the best reading experience for readers with little to no familiarity

in a topic, high familiarity readers tended to ignore certain details in overly plain summaries (e.g., study limitations). Our work provides methods and guidance on ways of adapting plain language summaries beyond the single “general” audience.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI; HCI design and evaluation methods.**

KEYWORDS

Language complexity, science communication, LLMs

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0330-0/24/05.
<https://doi.org/10.1145/3613904.3642289>

ACM Reference Format:

Tal August, Kyle Lo, Noah A. Smith, and Katharina Reinecke. 2024. Know Your Audience: The benefits and pitfalls of generating plain language summaries beyond the “general” audience. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 26 pages. <https://doi.org/10.1145/3613904.3642289>

1 INTRODUCTION

A rich body of work in HCI has shown that for many interfaces, one size does not fit all. Adapting interfaces to different users has the potential to improve usability [13, 87], aesthetic judgements [28, 69], and trust [61, 68]. Increasingly, language styles, such as community language norms [26], formality [9], and text complexity [11, 95] have been the focus of adaptable user interfaces. Work has shown that language styles can impact behavior in online experiments [9], counseling conversations [4], online communities [26], and security interfaces [98]. This work has highlighted the benefits of adapting language to people with different backgrounds [11].

With the rise of language models (LMs), interfaces promising adaptable language have progressed rapidly. Models like GPT-4 can ostensibly rewrite language for any reader by prompting the model to generate text for an envisioned audience or persona (e.g., a 5th grader) [52, 81, 104]. This is especially enticing in scholarly and scientific communication, where language styles (e.g., medical jargon) can present major communication barriers [85]. Research has explored using models to adapt scientific papers for non-experts (referred to as general audience readers in this paper) [11, 44], and paid services like Elicit,¹ or Explainpaper² promise to make scientific language easier to read and understand.

While adaptable language interfaces for communicating science are promising, it is not clear when and how to adapt. Most research showing that general audience readers respond positively to simplified language has focused on a single version of a simplified summary and a single general audience [31, 42, 44]. People have different knowledge and topic familiarity (e.g., someone who has read popular science books on a subject compared to someone who has not) that can impact how they respond to scientific information [14, 39, 74], suggesting that a simplified summary may be good for some, while a more complex version may be advantageous for others. However, no work has empirically shown this to be the case. Further, simplified summaries usually convey less information [7] and can unintentionally lead to people being overconfident in their understanding [90]. In contexts where details are important, it may be important to preserve all information, even at the cost of longer or more complex text (e.g., a medical research paper [11]). This gap in research is particularly important for developers of new interactive text interfaces [11, 63] because it is currently not clear what the benefits and pitfalls of adaptive text are: when is simplifying necessary, what are the costs in doing so, and do these costs differ for readers with different background knowledge?

Here we investigate how changes in scientific text affect the reading experience of general audience readers, for the first time taking into account varying levels of complexity in the text and background topic familiarity of the reader. We focus on scientific text complexity, defined as a combination of simple language and information content (§2). We introduce three RQs to understand how changes in complexity and information content affect readers:

RQ1: How do participants of different backgrounds respond to **human-written** scientific text at different complexity levels?

RQ2: How do participants of different backgrounds respond to **machine-generated** scientific text at different complexity levels?

RQ3: How do participants respond to generated scientific summaries at different complexities if they **report similar information**?

We started with studying expert-written summaries (RQ1) to establish what benefit we might expect from using alternative complexity versions, assuming no interference from imperfect text generation tools. We followed up with two studies using machine-generated summaries. In study 2 we used generated summaries with no restriction on information content (RQ2), following prior work on generating scientific summaries for general audience readers [10, 43]. In study 3 we evaluated generated summaries that aimed to preserve information content in lower complexity summaries (i.e., explaining details rather than removing them) (RQ3). We ran within-subjects experiments on Mechanical Turk for each RQ (Study 1: $N = 199$, Study 2: $N = 191$, Study 3: $N = 203$) evaluating whether topic familiarity affected participants' response to summaries written or generated for different envisioned audiences at three levels of complexity.

We found that topic familiarity mattered for determining the ideal summary for a reader. While the lowest complexity summaries were generally better for people with minimal topical knowledge (illustrated in the lower left quadrant of Figure 1), participants with more topic familiarity reported similar reading experiences across the three summary versions. Further, the lowest complexity summaries came with two costs to high familiarity participants. The first was that low complexity summaries in studies 1 and 2 removed details and reported on less information than high complexity summaries, shown with automatic and manual evaluations. This loss of information came with the benefit of improving the reading experience for low familiarity participants, but there was no benefit for high familiarity participants. The second, related cost was that high familiarity participants were more likely to skip sections of lower complexity summaries in all three studies (upper left quadrant of Figure 1). The most commonly skipped text focused on a paper's limitations, highlighting the risk that low complexity summaries have for high familiarity readers.

Our findings provide guidance on when and how to adapt scientific language to general audiences readers. Given our findings, we propose to only use the plainest language when an audience knows very little about a topic. In cases where audiences might have extensive background knowledge (even if they are not researchers themselves), language can be more complex—even drawn from the research paper—in order to convey more information and keep more knowledgeable audiences engaged (§4 & 5). When it is vital to convey complete information, such as in a patient-clinician context, plain language that explains all information can still be beneficial even if it is much longer, but only to those with little knowledge of a scientific topic (§6). Our findings make the following contributions:

- (1) **Shows the effect of text complexity on general audience readers of varying topic familiarity** (e.g., not comparing doctors and patients, but comparing different patients). We found that plain language summaries are better for those with little knowledge of a topic, and complex

¹<https://elicit.org/>

²<https://www.explainpaper.com/>

summaries, even those containing original scientific text, are better for those with more background knowledge.

- (2) **Highlights the benefits and pitfalls of generating plain language summaries.** When plain language summaries matched a reader’s background, readers had better reading experiences (e.g., were more engaged and had an easier time reading); however, plain language summaries often included less information and could lead to increased skipping when readers were more familiar in a topic.
- (3) **Provides guidance on generating plain language for different audiences.** Science communicators and interface designers can use our findings and methodology (§5.1.1 & 6.1.1) to effectively provide multiple summaries of scientific findings to different people and build adaptive text interfaces. We discuss this guidance further in §7.1.

While LMs make it possible to generate language for a wide range of contexts and people, there are also risks of factually incorrect generations [67]. We discuss these risks in the context of science communication (§5.1.2) and the need for expert oversight for generative systems (§7). Our work illustrates ways for automated methods to assist human efforts in communicating scientific information to a wider range of people, going beyond a single general audience.

2 LANGUAGE COMPLEXITY

In this paper we define language complexity based on prior work in readability, plain language summarization, and science communication. Broadly we break down complexity along two dimensions: surface level, textual features of the language (referred to as “plainness” in this paper) and the information conveyed by the language (referred to as “information content”). In this work we realize different language complexities by writing or generating summaries to different potential audiences (e.g., a high-school educated adult).

In most science communication writing, both plainness and information content are varied to produce text suitable for different audiences. This joint variation is reflected in the guidelines for plain language summaries³ and in the strategies science writers use to communicate with interested publics [7]. At the same time, these two dimensions have real-world constraints: there are situations in which technical words must be used to convey specific meaning, or where there is a desire to understand the majority of the details in the original scientific article, such as a patient reading a medical research paper or lab report [11, 76]. In studies 1 and 2, we allow plainness and information content to vary based on the intended audience (§4 & §5). In study 3, we explicitly try to preserve information content by explaining rather than removing details from the high complexity summaries to evaluate the effect longer plain summaries have on readers of different backgrounds (§6).

3 RELATED WORK

Below we cover additional prior work related to language personalization, plain language summaries for science communication, and augmented reading.

3.1 Personalizing language

There is a rich literature on adaptive interfaces and personalization in many domains, including website design [86], advertisement [46, 101], study recruitment [8], journalism [2], and education [35, 38, 77]. Usually personalization focuses on adjusting visual elements, but work has also shown the benefit of adjusting language to different audiences. In the medical domain, Dimarco et al. [34] proposed HealthDoc, a system that generated personalized patient pamphlets according to patient demographic information, education, and health history. Prior work has found that such tailoring of patient pamphlets can improve health outcomes, including smoking behavior and future health complications [66, 96, 100]. In journalism, Adar et al. [2] introduced PersaLog, a system for authoring personalized news articles. Articles authored using PersaLog presented alternative content (e.g., heat estimates for different areas) depending on user traits (e.g., a user’s location). Finkelstein et al. [38] showed that adjusting the dialect of a tutoring system could improve learning outcomes for children using African American English. Also in the education domain, work has shown that adjusting learning environments to learning styles or using personally-relevant examples can improve learning objectives [27, 53]. Past work has also personalized generated news articles [78], scientific definitions [71], recommended articles to read [45], and the amount of text displayed in a website [107].

Previous adaptive language-based interfaces have either relied on experts to author multiple versions of content [2], used rules and templates to automatically adjust content [34, 78], or focused on specialized populations (e.g., researchers [71]). Manually writing versions of text for each possible reader is infeasible, and rule-based approaches are brittle and only applicable to narrow content adaptation. In this paper we evaluate the feasibility of using modern NLP techniques to automatically generate multiple versions of text across a range of language complexities to communicate scientific information to different general audience readers.

3.2 Plain language summaries

Plain language summaries (PLS), also referred to as lay-summaries, patient summaries, or consumer summaries [99] are becoming an increasingly common method for communicating scientific findings with the public. Shailes [93] surveyed ten organizations that produced plain language summaries, finding that while summaries might initially be intended for one audience (e.g., undergraduates), often other people would engage with the summaries [88].

Studies have also explored how plain language summaries should be written based on empirical evidence from readers. Santesso et al. [89] found that using structured headings and narrative flow improved comprehension compared to paragraphs of text explaining results. Ellen et al. [36] interviewed participants about their preferences for plain language summaries, finding that people prefer key message headings and bullets over paragraphs. Silvagnoli et al. [95] explored the preferences of summary text complexity, measured by automated readability formulas, across different age groups. They found that most people preferred a medium complexity, while the lowest complexity was viewed as too simple and the highest complexity as too hard. Other work has studied how to present numerical results in summaries [17], uncertainty in findings [3]

³<https://consumers.cochrane.org/PLEACS>

and how summaries compare to other methods of science outreach, such as infographics [16], press releases [51] and Wikipedia articles [6]. In this paper we investigate if there is a benefit to adjusting the complexity of plain language summaries to different general audience readers.

3.3 Augmenting scientific reading

New interaction techniques have augmented readers' process to improve understanding and engagement, especially for scientific text. Chaudhri et al. [19] introduced *Inquire Biology*, a biology textbook that allows students to view concept definitions and ask open-ended questions about information in the textbook. Work has also developed new interaction techniques for researchers reading papers, including surfacing definitions [47], searching over related work sections [80], providing paper passages that answer natural language queries [109] and navigating concepts within a paper [1, 50]. With the improved performance of LMs like GPT-3, 3.5, and 4 [79], there has been dramatic growth in augmented reading interfaces for scientific papers [63]. For the general public, August et al. [11] introduced *PaperPlain*, a reading interface augmented with NLP to support general audience readers in approaching medical research papers. *PaperPlain* includes a curated set of key questions for guiding readers to the most important information in research papers. Augmented readers have also been released as products. *Explainpaper*⁴ is an LM-powered reading interface that allows users to ask questions over a paper and get simplified summaries.

Recent advances in NLP have also introduced automated methods to augment science communication [29, 44, 103]. Devaraj et al. [31] introduced a dataset of plain language summaries for clinical topics and a trained model for simplifying medical information. Laban et al. [57] constructed a new dataset of simplification edits made on Wikipedia articles, Basu et al. [12] introduced a dataset of simplification edits for medical texts, and Guo et al. [42] introduced a new evaluation suite for plain language summarization. August et al. [10] introduced methods to generate definitions at different levels of complexity. Shaib et al. [92] evaluated simplified summaries of biomedical papers generated by GPT-3, finding that GPT-3 could simplify and summarize from single paper, but it struggled to synthesize information across multiple papers.

Previous work for augmenting or generating scientific text either assumes there is a single ideal summary for all readers, or that adapting language to an individual reader is always useful. To our knowledge, no work has investigated if and when adaptation is important for scientific communication. This is of particular importance to developers of augmented reading interfaces because it is currently not clear when augmentation or adaptation is necessary. For example, do all general audience readers need a reading interface to provide a plain language summary of a scientific paper? If so, should this summary look the same for everyone, or is there measurable improvement in reading experience if the summary matches the background of the reader? In this paper we investigate how general audience readers with different familiarity in a scientific topic respond to scientific text at different complexities to inform the development of augmented reading interfaces for scientific text.

⁴<https://www.explainpaper.com/>

4 STUDY 1 – EXPERT-WRITTEN SUMMARIES

Study 1 focused on expert-written summaries to establish what benefit we might expect from alternative complexity versions. The study answered our first research question:

RQ1: How do participants of different backgrounds respond to **human-written** scientific text at different complexity levels?

Science writers adapt scientific language for general audiences. However, there is rarely a single general audience, and writers may use different strategies to engage different general audiences [7, 84]. Study 1 investigated how adjusting scientific language complexity affected people of different knowledge backgrounds.

4.1 Method

The three studies shared the majority of their procedure, materials, participant recruitment, and analyses (Figure 2). Below we report on the shared portions and those unique to study 1. Later, we report on differences in the methodology of studies 2 (§5) and 3 (§6).

4.1.1 Procedure. Participants answered questions about their scientific background, read summaries of scientific papers at three levels of complexity, and answered questions about the summaries. At the start of each experiment, participants filled out a demographics questionnaire, including questions on their education, STEM experience, and interest in scientific subjects. After the demographic questionnaire, participants read three article summaries, described in §4.1.2. The articles and complexity levels were randomized. Each participant saw one of each complexity in random order.

Summaries were broken down into sections answering key questions about the paper, following prior work showing that sections and headers were preferred by general audience readers [89]. The key questions were based on prior work studying the key information that science communicators focus on in a paper [7, 21] and from questions general audience readers found useful to determine relevant information in research papers [11]. Summaries were displayed as a title and a list of accordions (Figure 3). Participants could open multiple accordions at once. The questions were:

- (1) What did the paper want to find out?
- (2) What did the paper do?
- (3) What did the paper find?
- (4) What are the limitations of the findings?
- (5) What is the real world impact of this work?

Below the summary, participants could check a box requesting the original research paper. If participants checked this box, then a link to the paper was provided at the end of the study. Participants were asked to read the summaries for at least 30 seconds, though they could read for as long as they wanted. If participants clicked the continue button before 30 seconds, they were prompted to read for at least 30 seconds. They could ignore this prompt by clicking the continue button again. Participants on average took 143 seconds per article (std=103 seconds) for study 1, 100 seconds (std=84) for study 2, and 137 seconds (std=78) for study 3. Participants then answered questions on their topic familiarity and reading experience.

4.1.2 Materials.

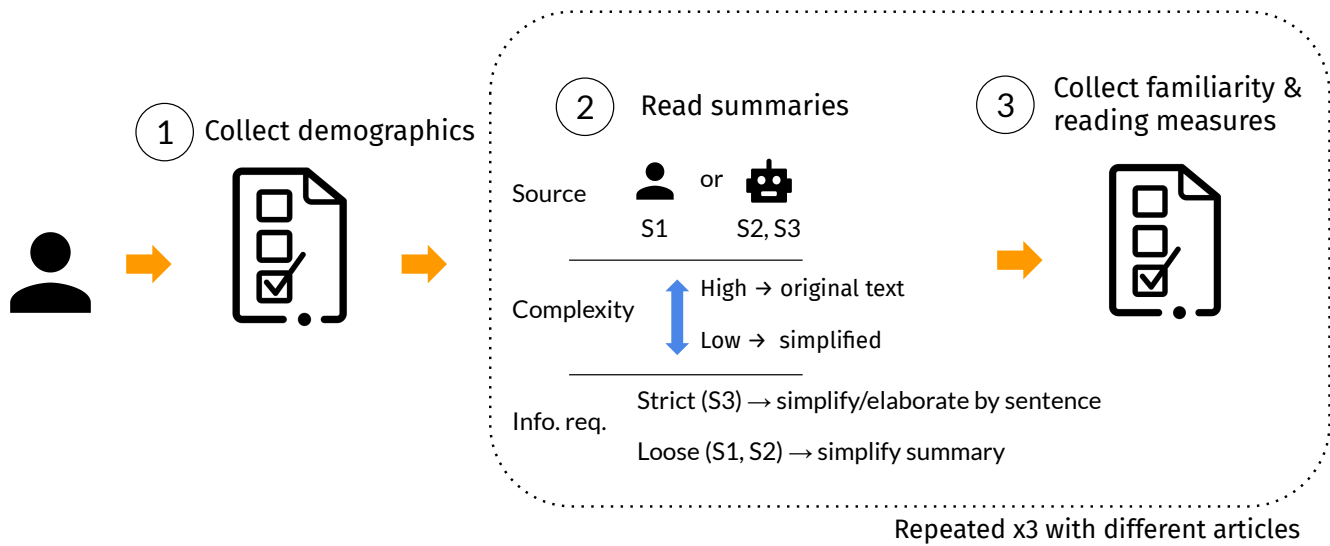


Figure 2: Flowchart of the study method, with shared features of all studies listed once. Ordering of summaries were randomized.

Article selection. We selected research papers that had public appeal by sampling papers posted and widely discussed in the large subreddit *r/science* in 2019. We randomly sampled 10 papers posted on *r/science* that contained a link to a research paper (as opposed to a press release or news article), and that had a score within the top 10% of posts containing research papers. We used the PSAW Python PushShift API for accessing *r/science*.⁵ The papers ranged in topics from public policy to nanotechnology, reflecting the breadth of research papers posted and discussed on *r/science*.

Authoring the summaries. An expert science writer with over 5 years of science communication experience crafted two versions of each summary. Each version was written for a different audience of a certain education level: a high school student or a college educated adult. In addition, the writer extracted sentences from the original paper to answer each key question. This constituted a third complexity aimed at other researchers. We defined these three complexity levels as Low (high school student), Medium (college educated adult), and High (researcher). Because the original paper text used a different voice than the other two versions, we lightly edited the High version by changing “we” to “the researchers.” One author reviewed each summary version and provide feedback to the writer on language complexity between the three versions in four weekly meetings, as well as asynchronously with Google Docs. The rest of the authors reviewed the completed summaries to determine that each version was distinct from the others in language complexity. The writer was paid \$17.22 USD per hour. Table 1 provides examples of the summaries and Table 2 lists word and sentence statistics for all summaries. All summaries are provided in the supplementary.

4.1.3 Measuring language complexity. We additionally report on automated measures of complexity for each summary version in order to see how the generated summaries differ across complexity levels. Table 2 details the measures for each generated version. We report on three automated measures: uncommon words (i.e., English words outside the top 1,000 most common), function word count, and language model perplexity. While these measures do not capture all dimensions of complexity, they are measures for analyzing scientific complexity at scale used in prior work on adjusting language in science communication [10, 42]. Each measure is described in more detail in Appendix A.

Table 2 reports the results of the automated measures for all three studies. The Medium and Low machine generated summaries in studies 2 and 3 had noticeable differences in average number of words, average proportion of uncommon English words outside the top 1,000, average proportion of function words, and language model perplexity. Compared to the expert written summaries, the generated summaries had more differences in the automated complexity measures, especially for generated text in study 2.

4.1.4 Participants. We recruited participants on Amazon Mechanical Turk with the slogan, “Read about interesting scientific findings and answer questions about your experience.” Participants were paid \$2.50. Participants were required to have completed over 1,000 HITs with a minimum approval rating of 95% and be US-based. For studies 1 and 3, participants were required to be master Turkers. This study was approved by our institution’s IRB. We removed participants whose native language was not English (1 in study 1, 2 in study 2, and 3 in study 3) and who indicated in a final self-report survey that they had technical difficulties or were cheating (1, 12, and 0, respectively). After removal, we had 199 participants for study 1, 191 for study 2, and 203 for study 3. Table 3 lists demographics and topic familiarity.

⁵<https://psaw.readthedocs.io/en/latest/>

Progress: 3 / 3

Please read the summary below for at least 30 seconds, though you can read for longer. If you are interested, you can also click the checkbox to get the original paper at the end of the study, though this is not required.

Aggressive Video Games are Not a Risk Factor for Future Aggression in Youth: A Longitudinal Study

What did the paper want to find out?	^
The issue of whether games with aggressive or violent content contribute to aggression or violence in society remains an issue of significant controversy worldwide. A large sample of 3034 youth in Singapore were assessed for links between aggressive game play and seven aggression or prosocial outcomes two years later.	
What did the paper do?	v
What did the paper find?	v
What are the limitations of the findings?	v
What is the real-world impact of this work?	v

Check this box if you would like to receive more information about this article at the end of the study



Figure 3: The study interface for reading the article summaries. The accordions started closed.

Extrinsic motivations like payment can lead participants to maximize pay at the expense of data quality (e.g., by rushing through a study [9, 106]). Studies 1 and 3 used Master Turkers, who have been shown to provide data quality equivalent to intrinsically motivated participants (e.g., participants motivated by supporting science) [106]. After finding comparable results between master and non-master workers in a study 2 pilot, we did not include the masters

requirement for study 2. However, we did have to remove more participants who had reported cheating during study 2.

While participants might have behaved differently (e.g., skipped less sections, §4.1.5) if they were interested in the summaries for their own sake, we did not expect this to bias differences across complexity versions due to the within-subjects nature of the studies. Considering that prior work studying general audience readers of scientific articles has found that readers may skip parts of an article

Source	Complexity	Summary
Expert - Study 1	High	These results demonstrate an unprecedented opportunity for development of these nanorgs as renewable sugar-free microbial factories for the production of biofuels and chemicals.
	Medium	This work is some of the first to examine the <i>feasibility of interfacing nanoscale materials</i> with living cells ... which could have broader implications for diagnostic and therapeutic applications of this technology.
	Low	This work is some of the first to be done investigating the <i>possibility of using nanoscale materials</i> inside living cells ... which has far-ranging applications for medicine.
Machine - Study 2	Medium	The study found that nanorobots ... can be used to <i>externally regulate the cellular function</i> of living cells using electromagnetic stimuli such as light, sound, or magnetic field.
	Low	This study found that nanorogs can be used ... to <i>control living cells using light, sound, or magnetic fields.</i>
Machine - Study 3	Medium	This study shows that <i>nanoscale organisms (nanorgs) can be developed into sustainable, sugar-free factories.</i>
	Low	These findings show a new chance to create <i>tiny organisms (called nanorgs) ... without using sugar, using sunlight in a way that can be reproduced on a larger scale.</i>

Table 1: Examples of the summaries. These summaries were under the heading “What are the real world impacts of the findings?” for the same paper. Bolded purple text indicates examples of changes in information content between the summaries, and italicized blue text indicates changes in plainness. For study 2, there was no information restriction in generated summaries. In study 3, there was information restriction for generated summaries.

Source	Complexity	# Words _{std}	# Sentences	Unc. Words ↑	Func. words ↓	Perplexity ↑
Expert - Study 1	High	483.10 _{107.02}	16.70 _{4.03}	0.55 _{0.04}	0.27 _{0.03}	94.60 _{30.91}
	Medium	369.60 _{82.56}	12.10 _{1.97}	0.47 _{0.05}	0.31 _{0.02}	60.08 _{14.84}
	Low	358.90 _{98.92}	11.50 _{3.21}	0.43 _{0.05}	0.32 _{0.02}	53.68 _{9.96}
Machine - Study 2	Medium	529.20 _{182.48}	20.80 _{7.05}	0.51 _{0.04}	0.31 _{0.03}	64.14 _{24.16}
	Low	259.00 _{43.42}	13.20 _{1.75}	0.28 _{0.05}	0.36 _{0.03}	23.92 _{5.71}
Machine - Study 3	Medium	878.90 _{212.81}	31.00 _{8.06}	0.48 _{0.02}	0.34 _{0.02}	46.26 _{9.90}
	Low	1005.00 _{273.63}	37.70 _{9.91}	0.37 _{0.03}	0.37 _{0.02}	34.40 _{4.83}

Table 2: Average number of words and sentences, along with differences in automated complexity measures between in each summary version. For study 2, there was no information restriction in generated summaries. In study 3, the summaries were generally longer because they included more details from the High summaries (i.e., they had stricter information requirements). Arrows denote expected increase (↑) or decrease (↓) in measure as complexity increases.

[24], we are excited to investigate how our findings generalize to readers motivated simply by interest in a topic.

Topic familiarity. After each summary, participants rated their familiarity with the article’s topic on a 1–5 Likert-style scale based

4.1.5 Measures.

		Study 1	Study 2	Study 3
Age	0-19	0	0	0
	20-29	14	49	9
	30-39	68	87	76
	40-49	71	32	57
	50-59	29	18	29
	60-69	14	4	21
	70-79	3	1	2
	80+	0	0	0
Gender	Male	98	96	93
	Female	99	95	109
	Prefer not to answer	2	0	4
Education	Pre-high school	0	1	0
	High school	58	30	48
	College	117	114	137
	Graduate school	19	40	20
	Professional school	5	6	1
# STEM courses after high school	0	36	21	36
	1-3	89	93	104
	4-6	41	57	32
	7-10	11	9	10
	≥11	22	11	21

(a) Participant demographics

Familiarity	Study 1	Study 2	Study 3
1	359	150	297
2	115	72	134
3	97	132	134
4	26	165	39
5	0	54	5
Total	597	573	609

(b) Topic familiarity based on question “How familiar are you with the topic of this article?” 1=“I have never heard about this topic before”, and 5=“I have written research papers on this topic.”

Table 3: Participant demographics (a) and topic familiarity (b) for all studies

on the question: “How familiar are you with the topic of this article?”⁶ with 1 being “I have never heard about this topic before” and 5 being “I have written research papers on this topic.” Table 3b details the topic familiarity ratings for the three studies.

Reading experience ratings. We collected subjective ratings to understand how the different complexity levels affected participants’ reading experience. Participants completed the ratings after reading each summary. All ratings were based on a 1–5 Likert-style scale. These included:

- (1) **Reading ease:** Participants rated their reading difficulty based on the question: “How easy was it for you to read the article?”
- (2) **Understanding:** Participants rated their confidence understanding the summary based on the question: “How confident do you feel in your understanding of the article?”
- (3) **Interest:** Participants rated how interesting they found a summary based on the question: “How interesting did you find the article?”
- (4) **Value:** Participants rated how valuable they found the information in the summary based on the question: “How much would you agree that this article contained valuable information?”

Skipped sections. We analyzed how many summary sections participants skipped in each complexity condition. As described in

⁶Because participants were only ever presented summaries, not the original paper, in the study the summaries were referred to as ‘articles.’

§4.1.1, each summary was made up of five accordian drop-downs that participants could open. Each accordian section began closed. Participants were not instructed to open all sections. To determine which sections were opened, we logged click events for each accordian section.

Requested articles. A primary goal of science communication is to encourage audiences to engage further with science [74]. We capture the potential for increased engagement with science by analysing how likely participants were to request the original scientific article after reading a summary.

4.1.6 Analysis. We compared measures across the complexity versions using linear mixed-effects models (LMMs). LMMs are commonly used to analyze data in which the same participant provides multiple, possibly correlated, measurements, referred to as repeated measures [62] and have been used as an analysis tool in the behavioral sciences [25] and human-computer interaction [47, 48].

We fit a model for each reading experience rating, number of skipped sections, and original article requests. Each model contained fixed effects for the complexity version, topic familiarity, an interaction term for familiarity and complexity, and random effects for paper and participant IDs. We conducted post-hoc two-sided *t*-tests for pairwise comparisons to examine the differences in measures between pairs of complexity levels estimated by the linear mixed effects models. These pairwise comparisons reveal not only what differences between measures are significant, but the estimated differences *d* between measures. Because *d* is estimated

by the linear mixed-effects model, it represents the expected difference in some measure (e.g., reading ease), when controlling for the participant and paper random effects in the model. For example, if the estimated difference $d^{\text{Low-High}}$ in reading ease between two complexity options Low and High is 0.894, we can interpret this difference as participants rated the Low complexity, on average, 0.894 points higher for reading ease (out of 5) compared to the High complexity when controlling for participant and paper. We report these differences to provide further intuition about the effect of different complexity levels. We also include effect sizes, calculated using Cohen’s d and denoted SMD for standardized mean difference, as an additional measure of effect beyond the estimated pairwise difference.

The reading experience measures used Likert-style scales, making parametric tests potentially not appropriate, we report analogous non-parametric tests in Appendix B, which yield similar p -values and findings. For these analyses we use the PYMER4 Python package for fitting the models and pairwise comparisons. All t -tests were corrected from multiple hypotheses using the Holm-Bonferroni correction. The analysis was equivalent for the three studies. We report all pairwise differences and test statistics in Appendix F.

4.2 Results

Table 8 in the appendix lists all pairwise differences.

4.2.1 Reading experience measures. Figure 4a plots all participants’ ratings across summary complexities for study 1. Overall participants found the Low summaries most appealing. Across all measures there is a greater number of high ratings and fewer low ratings as participants are presented with less complex summaries. Compared to the High summaries, participants rated Low summaries as significantly easier to read ($d_{\text{ease}} = 0.893$, $p < 0.0001$, $SMD = 0.99$), understand ($d_{\text{understand}} = 0.589$, $p < 0.0001$, $SMD = 0.77$), and more interesting ($d_{\text{interest}} = 0.381$, $p = 0.018$, $SMD = 0.55$). Participants also rated the Medium summaries as significantly easier to read and were more confident in their understanding compared to the High summaries ($d_{\text{ease}} = 0.653$, $p < 0.0001$, $SMD = 0.71$; $d_{\text{understand}} = 0.400$, $p = 0.006$, $SMD = 0.59$).

Topic familiarity was a strong indicator of reading experience measures and interacted with summary complexity. Looking at Figure 5, as familiarity increased, ratings across all metrics and complexity levels generally went up (i.e., the orange bars shrink while the dark purple bars grow). Also apparent in Figure 5: at low familiarity, rating distribution are most different across the complexity levels. As familiarity increases, though, there were fewer low ratings and more high ratings for all complexity levels. This effect was also illustrated in the linear mixed effect models. Participants who rated their familiarity with a summary’s topic lowest (1 on a scale of 1–5) rated the Low summaries as being significantly easier to read, understand, more interesting, and containing more valuable information compared to the High summaries in study 1 ($d_{\text{ease}} = 1.490$, $SMD = 1.27$; $d_{\text{understand}} = 1.160$, $SMD = 1.07$; $d_{\text{interest}} = 0.943$, $SMD = 0.80$; $d_{\text{value}} = 0.509$, $SMD = 0.49$; $p < 0.0001$ for all comparisons). Participants who were most familiar with the summary’s topic, though, rated High complexity summaries as similarly easy to read and understand, and equally interesting and valuable as Low

and Medium summaries. Table 8 in the appendix lists all pairwise differences.

4.2.2 Skipped sections. Participants on average skipped 0.113 (std = 0.536) sections (out of 5). Skipped sections were lowest for the High summaries (mean=0.060, std=0.327) compared to the Low (mean = 0.129, std = 0.559) and Medium (mean = 0.149 std = 0.661) summaries. Topic familiarity mattered for determining number of skipped sections. Participants who rated their topic familiarity highest (4 on a 1–5 scale), clicked on significantly fewer sections in the Low summaries compared to the High summaries ($d_{\text{unclicked}} = 0.682$, $p = 0.008$, $SMD = 0.68$). Table 8 in the appendix lists all pairwise differences between skipped sections. Across all studies, the most common section skipped by participants was the paper’s limitations (“What are the limitations of the findings?”, 25% of skipped sections), the least common section was the paper’s goals (“What did the paper want to find out?”, 13%).

4.2.3 Original article requests. Participants on average requested the original article 14.7% of the time. Requests were roughly similar across the complexity levels (Low: mean=14.5%, Medium: mean = 15.5%, High: mean=14.0%). Topic familiarity affected how likely participants were to request the original article depending on complexity level. Participants with the second lowest familiarity (2 out of 5) requested the original article significantly less often in the Low summaries compared to the High summaries ($d_{\text{requests}} = -0.184$, $p = 0.007$, $SMD = -0.47$). Table 8 in the appendix lists all differences.

5 STUDY 2 - MACHINE-GENERATED SUMMARIES WITH NO RESTRICTION ON INFORMATION CONTENT

The results from study 1 suggest that low complexity summaries are best for low familiarity participants, while high familiarity participants were more likely to skip sections in low complexity summaries. We were curious if we would see similar differences in complexity preference with machine-generated summaries. We therefore conducted study 2, answering our second research question:

RQ2: How do participants of different backgrounds respond to **machine-generated** scientific text at different complexity levels?

There are methods to automatically adjust generated language complexity [10], but no work has explored the interaction of generated language complexity and participant background knowledge. Here we follow prior work on automated plain language summarization and allow generated text to vary information content freely [43, 57]. In study 3 we explore methods to preserve information through all complexity levels (§6).

5.1 Method

Below we describe generating summaries for study 2 and assessing their factuality. Please refer to §4.1 for shared methodology of studies 1, 2, and 3.

5.1.1 Materials.

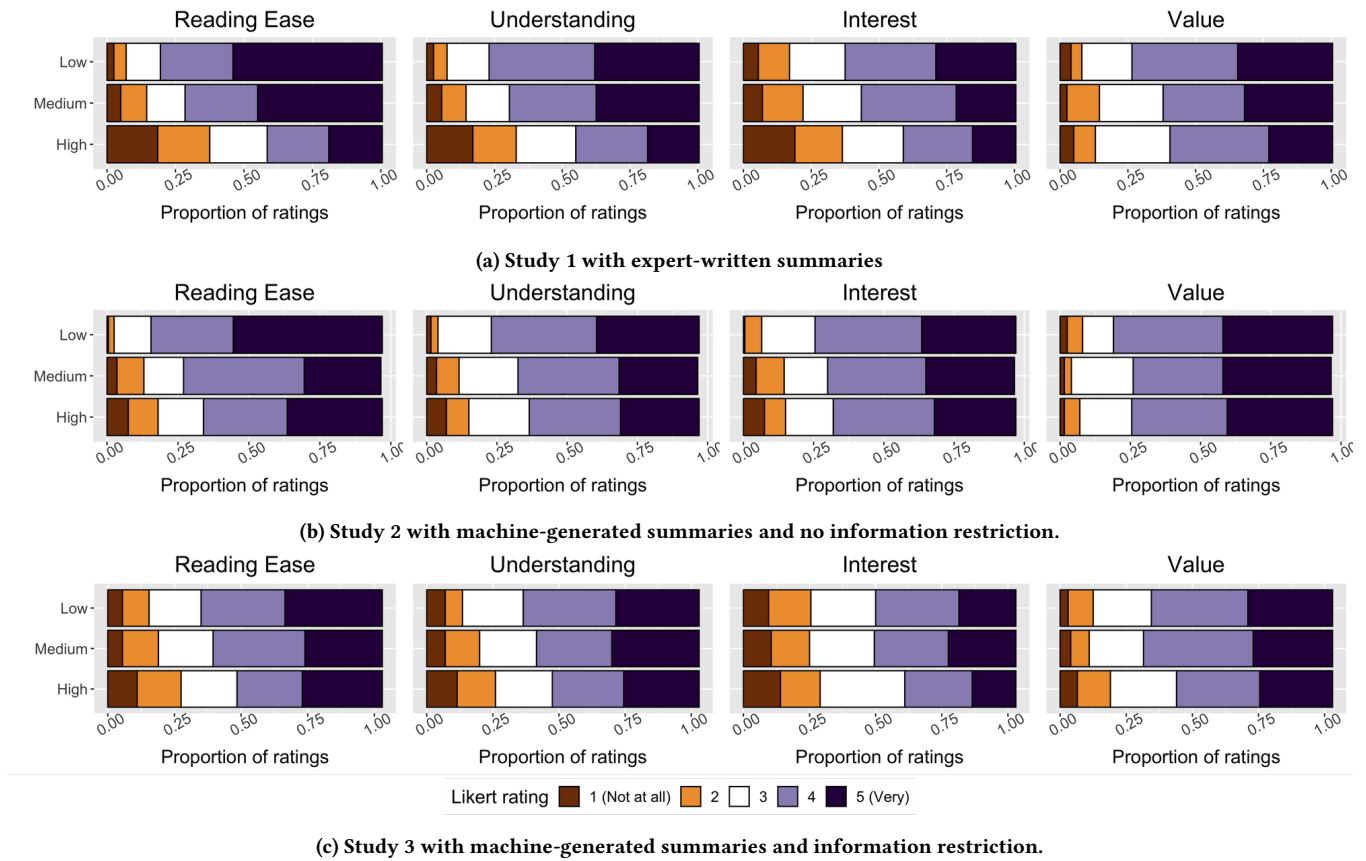


Figure 4: Distribution of ratings for each subjective reading experience measure across complexity levels. The ratings were based on the following questions: Reading ease: “How easy was it for you to read the article?”, Understanding: “How confident do you feel in your understanding of the article?”, Interest: “How interesting did you find the article?”, Value: “How much would you agree that this article contained valuable information?” Notice the greater number of high ratings (purple) and fewer low ratings (orange) as participants are presented with less complex summaries.

Generating the summaries. We generated summaries at different complexities in a two step process. In the first step, we generated candidate summaries using GPT-3. GPT-3 is a language model commonly used in generation tasks, including plain language summarization [15]. We adapted a preset prompt for GPT-3 to generate summaries with varying complexity. The original prompt was “Summarize this for a second-grade student: [TEXT]” Our adapted prompts for GPT-3 were 14 alternate prompts, from “first-grade student” to “twelfth-grade student”, along with “college student” and “college-educated adult.” We used GPT-3 (DAVINCI-003) in July 2022, with temperature set to 0.3 and the rest of the parameters set to default OpenAI API settings. At the time we ran this study, more sophisticated systems like ChatGPT had not been released. We investigate more sophisticated models (i.e., GPT-3.5 Turbo) in Study 3 (§6.1.1).

Because GPT-3 was not designed to explicitly vary text complexity, we additionally used the complexity ranker from August et al. [10] to rank the GPT-3 generations on a gradient of complexity. The complexity ranker was a linear discriminator trained to classify scientific text as either from a news article or research paper.

The ranker used features shown to be predictive of reading difficulty in scientific language, including technical word occurrences, proportion of function words, and text length [10]. After scoring each generation for complexity, we selected the generation with the highest and lowest score for the Low and Medium versions. For the High summaries, we used the original sentences extracted from the paper by the writer in §4.1.2. More details on the GPT-3 generations are in Appendix C.

5.1.2 Assessing factuality in generated summaries. A major limitation of language models is that they can generate text with meaning that was not part of the original input [67], referred to as hallucinations [41, 67]. While there are methods for reducing hallucinations or encouraging factuality [40, 56, 65], no automated method guarantees factual accuracy or fidelity to original text. In the context of science communication, such hallucinations can risk confusing or, worse, misinforming readers. A reader might trust a hallucinated result opposite to what was reported in the original paper [32], or be so confused by the contradictory evidence as to lose trust in the research.

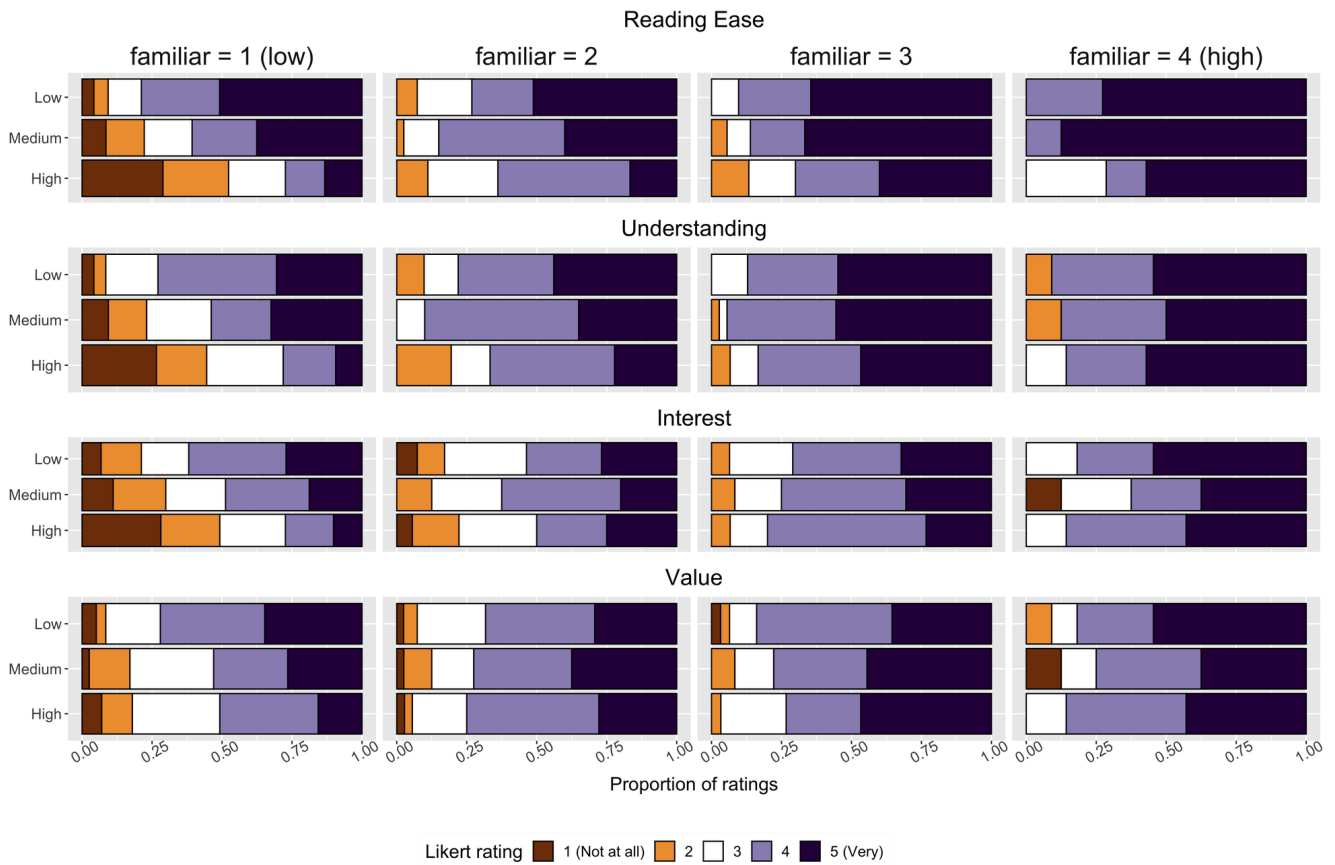


Figure 5: Distribution of ratings for each reading experience measure across complexity and participant topic familiarity for study 1 (expert written summaries).

Because of these risks, we advocate for NLP systems to be used in conjunction with experts. Plain language summaries are often written by researchers, editors, or science writers [93, 99]. Authors could generate multiple versions of a summary and then verify factual accuracy. In this way, we could lessen the workload of writing plain language summaries, make summaries adaptable to different audiences, and protect against factually incorrect generations.

In the context of study 2 and 3, one author selected generations that did not contain factually incorrect information, acting as the expert for checking generated summaries before publishing. In study 2, out of 120 generated summaries (6 sections including the title \times 10 papers \times 2 complexities), 14 generations contained incorrect information. In all 14 cases, a replacement was found by selecting from at most 6 alternative generations. The average number of generations the author looked at to find a replacement was 2.36. For study 3, while there were generations that were ill-formed (e.g., the model asking for clarification on an acronym) there were no factually incorrect generations. This difference in factuality might

be due to improvements between GPT-3 (used in study 2) and GPT-3.5 (used in study 3).⁷ Appendix E contains more information on hallucinations in our generated summaries.

5.2 Results

5.2.1 Reading experience measures. Similar to study 1, participants in study 2 rated Low summaries as significantly easier to read ($d_{ease} = 0.535, p < 0.0001, SMD = 0.56$) and understand ($d_{understand} = 0.323, p = 0.001, SMD = 0.38$) than the High summaries (Figure 4b). However, we observed two different results in this second study. First, while study 1 participants found Medium summaries significantly easier to read and understand than High summaries, study 2 participants did not. Second, while study 1 participants did not rate the Low and Medium summaries as significantly different, study 2 participants *did* rate Low summaries as significantly easier to read and understand than Medium summaries ($d_{ease} = 0.472, p < 0.0001, SMD = 0.53; d_{understand} = 0.279, p = 0.004, SMD = 0.29$).

⁷Because GPT-3.5 is a proprietary system, the full details of which have not been disclosed, we cannot be certain about whether or how factuality was improved.

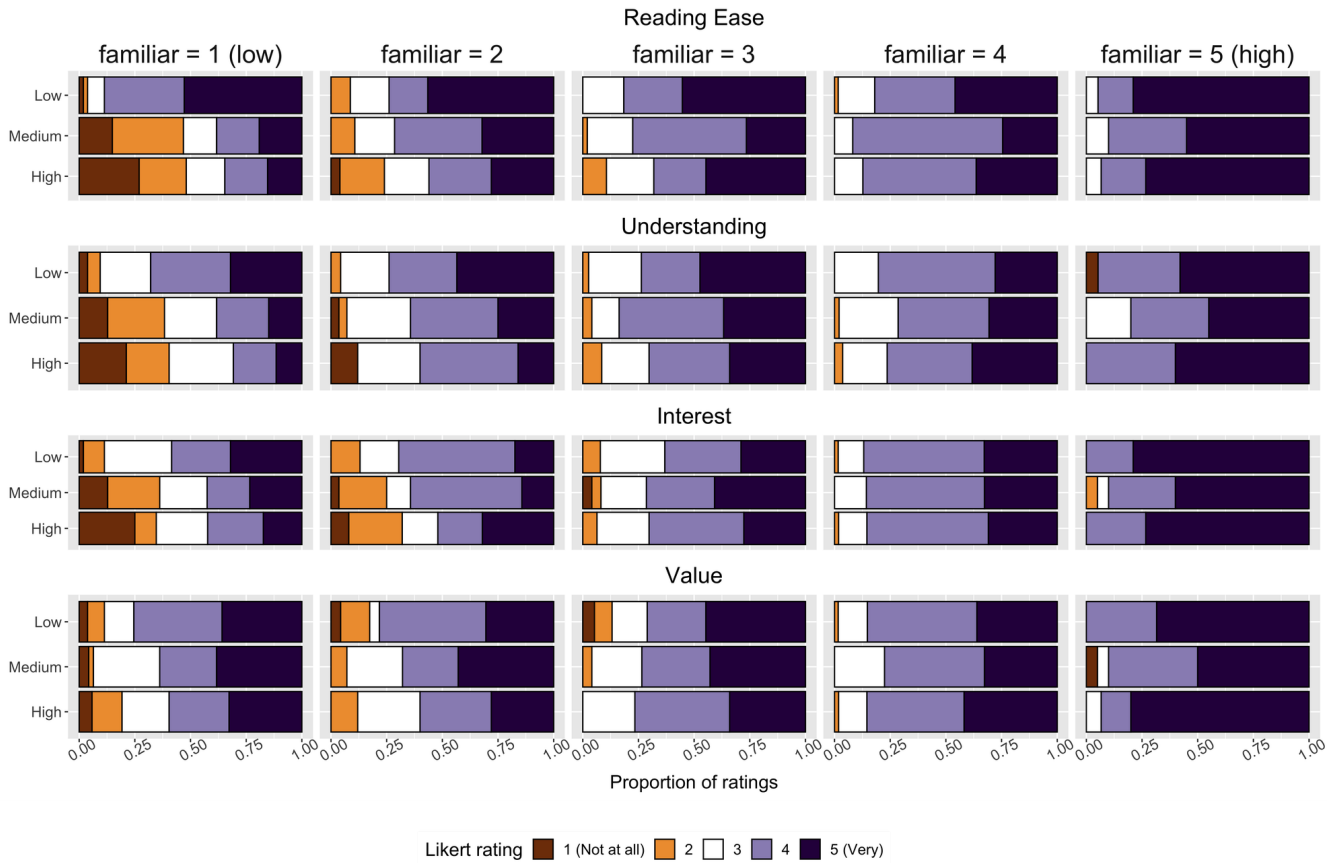


Figure 6: Distribution of ratings for each reading experience measure across complexity and participant topic familiarity for study 2 (machine-generated summaries and no information restriction).

Topic familiarity again interacted with complexity to equalize reading experience measures. Similar to study 1, participants with the lowest familiarity of a summary’s topic rated the Low summaries as being significantly easier to read, understand, more interesting, and containing more valuable information compared to the High summaries ($d_{ease} = 1.642$, $SMD = 1.34$ $d_{understand} = 1.103$, $SMD = 0.94$ $d_{interest} = 0.909$, $SMD = 0.62$ $p < 0.0001$; $d_{value} = 0.407$, $p = 0.031$, $SMD = 0.25$). In contrast, participants with the highest familiarity (5 on a 1–5 scale) rated their reading experience similarly between the complexity versions. Figure 6 plots ratings.

5.2.2 Skipped sections. Participants on average skipped 0.785 (std = 1.621) sections in study 2. While the overall rate of skipped sections was higher than for study 1, the trend of more skipped sections for lower complexity summaries held. Skipped sections were lowest for the High summaries (mean=0.749, std=1.543) compared to the Low (mean=0.849, std=1.710) and Medium (mean=0.759, std=1.613) summaries. Similar to study 1, participants with the highest rated familiarity (5 on a 1–5 scale) skipped significantly more sections in the Low summaries compared to the High summaries ($d_{unlicked} = 0.900$, $p = 0.011$, $SMD = 0.35$). This estimated difference between

skipped sections constitutes close to a full extra section skipped (e.g., skipping all of the summary’s limitations).

5.2.3 Original article requests. Participants on average requested the original article 52.5% of the time. Generally participants requested the original article from the Low summaries more often (mean=55.5%) than either the Medium (mean=48.2%) or High (mean=53.9%). In contrast to study 1, where low familiarity participants requested the original article more for High summaries, participants in study 2 with the second lowest familiarity requested the original article significantly more often in the Low summaries compared to the Medium summaries ($d_{requests} = 0.214$, $p = 0.036$ $SMD = 0.68$). Table 9 in the appendix lists all pairwise differences.

The results from study 2 corroborate and expand on our findings from study 1. Participants with low familiarity preferred generated low complexity summaries, while high familiarity participants again skipped sections of low complexity summaries more often. One contrasting finding from study 2 was that some participants with low familiarity requested the original article more often for low complexity summaries over more complex summaries. Given that we observed similar findings from study 1 with expert-written summaries, the results of study 2 suggest that machine-generated

summaries are a viable method for efficiently adjusting language to different audiences.

6 STUDY 3 - MACHINE-GENERATED SUMMARIES PRESERVING INFORMATION CONTENT

Summaries from studies 1 and 2 had no restriction on what information needed to be included. This followed past work in plain language summarization, where writers or models select some information to explain, and remove other information (e.g., focusing on a single finding or concept for low complexity text) [7, 43, 97]. However, selectively conveying information comes with the risk of removing information a reader might want [11], or giving a reader a false sense of understanding [90]. Emboldened by newer, stronger models being released (e.g., ChatGPT, or GPT-4), we were curious if generated text could preserve details from high complexity summaries in their low complexity counterparts, potentially mitigating the risk of information loss. This motivates our third research question:

RQ3: How do participants respond to generated scientific summaries at different complexities if they report similar information?

6.1 Method

Below we describe our method for generating summaries in study 3. Please refer to §4.1 for shared methodology of studies 1, 2, and 3.

6.1.1 Materials.

Generating detail-preserving summaries. In studies 1 and 2, there was no requirement that summaries preserve information (i.e., it was acceptable if a simpler summary removed some information). For study 3, we sought to generate low complexity summaries that preserved information content (i.e., were plainer but included all details). We did this by leveraging stronger models released after study 2 and developing a prompting technique to simplify each sentence separately, prompting the model to elaborate on details rather than remove them. In simplification literature, both removing and elaborating on details are common tasks [12, 57]. In the context of study 3, we structured model input and prompts to minimize detail removal and maximize elaboration for all details in the original sentence. We used GPT-3.5 Turbo in May 2023 with temperature set to 1.0 and the rest of the parameters set to default OpenAI API settings.

We generated summaries that did not remove and instead elaborated on details by restricting the model input and changing our prompting technique. Rather than input the entire High summary, as in study 2 (§5.1.1), we provided GPT-3.5 with a single sentence at a time and instructed it to explain, rather than remove, any information from the original sentence. To avoid having subsequent sentences repeat themselves, the prompt included the history of previous simplified sentences and instructed the model not to explain a concept it had explained above. In addition to the instructions, the prompt included one example of a scientific sentence and its associated simplified version.

We used two prompts, one for Medium summaries and one for Low. The Medium prompt instructed the model to rewrite the

sentence for someone very familiar with the topic of the sentence, with a target reading level of a college educated adult. For the Low summaries the target user was someone who was not at all familiar with the sentence’s topic, with a target reading level of 5th grade. 5th grade was chosen based on previous work in generating plain language summaries [11], and on our observations that selecting a high school reading level, as we had done for the expert-authored summaries, produced text similar to the Medium prompt. The full prompts are included in Appendix D. Table 1 provides examples of the generated summaries.

6.1.2 Measuring information content in summaries. Before collecting participant response to the summaries, we analyzed how information content differed between the summary versions in the three studies. We used four automatic measures and one manual measure of information content based on previous work studying alignment between scientific text and summaries [37, 43, 55]:

SummaC: Laban et al. [56] introduced a natural language inference (NLI) approach to summary consistency. The method uses an NLI model to score each sentence from a source summary with sentences from a target summary on how much the target sentences follow from the source sentence (i.e., is true given the source sentence). We use the SUMMAC-CONV model using the default settings from the original metric library.⁸

SuperPAL: Ernst et al. [37] introduced a supervised method for scoring alignment between source and target summaries by annotating spans of text representing information units (i.e., a standalone fact). Using these annotated spans, the authors trained a model for the task of identifying information alignment between a source and target summary. In an evaluation of alignment scores for scientific summaries, SuperPAL was found to be the most effective at identifying aligned claims between the source and target [55]. We use the BUI-NLP/SUPERPAL model⁹ with the default settings.

ROUGE-L [60]: ROUGE is a common score for assessing summary quality by scoring the number of n-gram overlaps between source and target summaries. ROUGE has also been used as a baseline approach to aligning sentences between source and target summaries [43]. Following this prior work, we use ROUGE-L, which measures the longest common sub-sequence of tokens between a source and target sentence. We use the Huggingface EVALUATE package for calculating ROUGE-L.¹⁰

BERTScore [108]: BERTScore is a common score for summary evaluation that computes semantic similarity using pre-trained contextual embeddings from the BERT model [33]. We use the Huggingface EVALUATE package for calculating BERTScore and report the F1 score.¹¹

⁸<https://github.com/tingofurro/summac/tree/master>

⁹<https://github.com/martiansideofthemoon/longeval-summarization>

¹⁰<https://github.com/huggingface/evaluate/tree/main>

¹¹<https://github.com/huggingface/evaluate/tree/main>

For each measure we take the average maximum alignment score for sentences in the High summaries with sentences from the Medium and Low summaries. If a sentence in the High summary has low alignment scores for all sentences in the Medium or Low summaries, this would suggest that the information is not reported in the summaries.

In addition to the automatic measures reported above, we ran a manual evaluation of the information content between each of the summary version. We annotate all information units—defined similar to prior work as proposition-level semantically equivalent statements [37]—for the High summaries and count how many of these units appear in the Medium and Low summaries. Annotating information units at this level has been used in prior work for evaluating claims in scientific summaries [55]. In our summaries these units were predominately definitions of terminology, reporting of results, methodological details, and background explanations. Our codes are provided in the supplementary.

Table 4 lists the scores for summaries' information content. Across all measures and versions, the Low summaries score lower than the Medium summaries. The most common information skipped in all the summaries (based on our manual evaluation of information units) was information about the findings from the studies. This aligns with feedback from our writer, who said that in the Low summaries they focused on only the most important finding, while in the Medium summaries they included more details. One reason for the lower scores on most automatic measures for the expert summaries might be due to the writer using fewer overlapping words compared to the models. The same can explain the higher ROUGE-L score for the study 2 Medium summaries, which used many spans verbatim from the original summaries. In comparison to the summaries from studies 1 and 2, though, the summaries in study 3 have consistently higher scores and differences between the Medium and Low versions are within 1.5 standard deviations.

6.2 Results

6.2.1 Reading experience measures. Compared to the first two studies, there were smaller differences in reading experience ratings between the three complexity versions. Figure 4c plots the overall ratings. While participants generally rated Low summaries as easier to read ($d_{ease} = 0.166$, $p = 1.0$, $SMD = 0.29$) and understand ($d_{understand} = 0.734$, $p = 0.051$, $SMD = 0.24$) compared to High summaries, these differences were smaller and not significant.

Participants who had the lowest familiarity of the summary's topic again rated the Low summaries as significantly easier to read and understand than the High summaries ($d_{ease} = 0.362$, $p = 0.019$, $SMD = 0.27$; $d_{understand} = 0.420$, $p = 0.003$, $SMD = 0.28$). Similar to studies 1 and 2, participants with more familiarity rated the three summary versions similarly, with no significant differences between them. Figure 7 plots ratings broken down by familiarity.

6.2.2 Skipped sections. In Study 3, participants on average skipped 0.554 (std=1.220) sections. Similar to studies 1 and 2, skipped sections were lowest for the High summaries (mean=0.490, std=1.134) compared to the Low (mean=0.529, std=1.209) and Medium (mean = 0.642, std=1.310) summaries. Participants who rated their topic familiarity as a 3 out of 5, indicating moderate familiarity, skipped significantly more sections in the Medium summaries compared

to the High summaries ($d_{unclicked} = 0.472$, $p = 0.026$, $SMD = 0.57$) and Low summaries ($d_{unclicked} = 0.583$, $p = 0.004$, $SMD = 0.51$).

6.2.3 Original article requests. Similar to study 2, participants requested the original article from the Low summaries more often (mean=18.7%) than either the Medium (mean=12.8%) or High summaries (mean=12.8%). Also supporting our results from study 2, participants in study 3 with the lowest familiarity requested articles significantly more often after reading the Low summaries compared to the Medium summaries ($d_{requests} = 0.108$, $p = 0.023$, $SMD = 0.39$) and High summaries ($d_{requests} = 0.110$, $p = 0.023$, $SMD = 0.34$). Table 10 in the appendix lists all pairwise differences.

7 DISCUSSION

In this paper we set out to understand how general audience readers with different background knowledge respond to alternative versions of scientific language. We conducted three studies, using both human-written and machine-generated text, investigating the effect of language complexity and topic familiarity on reading experience and behavior. We found that the lowest complexity summaries, both human-written and machine-generated, provided the most benefit to readers with little familiarity of a scientific topic (e.g., those who had never heard of the summary's topic before). Not only did low complexity summaries make it easier for low familiarity participants to read and understand the summaries, but in the case of machine-generated summaries, the low complexity summaries also encouraged them to request the original scientific article more, engaging with the science beyond what was required for the study.

In most cases, though, the benefits of low complexity came at the cost of reduced information content. In our first two studies, low complexity summaries provided less information overall than high complexity summaries, especially in reporting multiple findings (§6.1.2). In our third study, when we encouraged models to generate plain language that preserved details, we found that only readers with the lowest topic familiarity rated the longer plain summaries as easier to read and understand (§6.2). Most science communication text focuses on the most important findings and theories to convey by default [7, 21]. This is because reporting all scientific findings in plain language requires explaining any concepts an audience might not know [105], leading to long explanations that risk reader fatigue and loss of interest. Our findings from study 3 align with this work by showing that conveying complete information in plain language leads to longer summaries that were only easier to read for those who had no background in the summary's topic.

While lower complexity summaries might be ideal for low familiarity readers, they may invite high familiarity readers to ignore information. Across the three studies, participants with higher familiarity skipped sections of low complexity summaries significantly more than high complexity summaries. This could potentially be due to lack of interest, or feeling like the summary was talking down to them [95]. In some cases, the difference in number of skipped sections was close to one section out of five. While not all information may be necessary to convey, the skipped information was often the most risky to skip: the study's limitations.

Our findings are the first to illustrate the benefits and drawbacks of simplification for general audience readers with varying background knowledge. Prior work developing systems to support

Source	Complexity	SummaC	SuperPAL	ROUGE-L	BERTScore	Info. Units
Expert - Study 1	Medium	0.086 _{.188}	0.227 _{.012}	0.211 _{.115}	0.879 _{0.026}	0.557 _{.255}
	Low	0.098 _{.194}	0.225 _{.001}	0.197 _{.102}	0.878 _{0.024}	0.478 _{.241}
Machine - Study 2	Medium	0.782 _{.359}	0.673 _{.161}	0.730 _{.308}	0.957 _{0.047}	0.810 _{.317}
	Low	0.290 _{.360}	0.384 _{.250}	0.203 _{.128}	0.882 _{0.027}	0.418 _{.264}
Machine - Study 3	Medium	0.839 _{.263}	0.722 _{.042}	0.439 _{.124}	0.926 _{0.022}	0.998 _{.014}
	Low	0.750 _{.302}	0.684 _{.077}	0.313 _{.112}	0.910 _{0.022}	0.977 _{.062}

Table 4: Differences in automated information content measures between summary versions.



Figure 7: Distribution of ratings for each reading experience measure across complexity and participant topic familiarity for study 3 (machine generated summaries with information restriction).

science communication has predominantly focused on providing a single version of simplified language and treated general audience readers as a single, monolithic group [31, 42]. While science communicators have a strong intuition that adapting language to different audiences is important [7], no work has taken the step of showing that such adaptation can provide measurable benefits. In our three studies, we show that the simplest summaries benefit readers with the least knowledge of a topic the most, and that more

complex summaries are best for those with greater background knowledge.

7.1 Guidance on adaptive plain language

This paper provides guidance on designing generated language for both science communicators and interface designers. Based on our findings we make the following suggestions:

- **Low complexity for low familiarity/information** The least complex plain language summaries are better when

one or both of the following is true: there is no requirement to convey complete information (§4 & 5), or the reader has little to no familiarity in the topic (in this case longer, plain summaries can be used, §6).

- **High complexity for high familiarity** More complex summaries—even with text drawn from the research paper—are better when audiences have more background knowledge (even if they are not experts in a field), in order to convey more information and keep readers engaged (§4 & 5).
- **Plain language for high information, when necessary** LMs can be used to generate plain language summaries that preserve details (§6.1.1); however, these summaries only benefit those with little knowledge of a scientific topic (§6) and should be used only when necessary because it leads to much longer text that risks losing readers that have even moderate topic familiarity.

Science communicators can use our findings to guide their efforts when reaching different audiences. If an article is intended for readers with no familiarity in a topic, a science writer could meet these needs by generating and editing a very plain summary or by assessing their own writing with automatic complexity measures (§4.1.3). In contrast, if a science communicator is worried about losing the engagement of readers with more topic familiarity, they could focus on a more complex summary, either generated or written. Further, a writer could create multiple alternative versions of a summary suited for different audiences quickly using our generation techniques (§5.1.1 & 6.1.1).

Interface designers can also leverage the techniques we illustrate in our studies to create interactive and adaptive reading interfaces. For example, a reading interface could generate a new summary on-the-fly based on the reader, or allow readers to interactively select different versions as they read. Short user surveys could be used to determine the ideal adaptation [102], similar to the method employed in this paper (§4.1.5). A complementary method would be to model users through behavioral signals, a common approach in the education literature [30, 54]. We observed that participants with higher familiarity were more likely to skip sections when the complexity was too low. A system that adapts scientific complexity could monitor how much skipping a reader engages in, increasing complexity with increased skipping. Another approach to modeling a user in this context is to analyze past reading or writing behavior [5]. A system could predict an ideal complexity based on the terminology and concepts contained in documents a user already knows. We recommend some level of user control for adaptive language. While users might not always know the ideal level of complexity for themselves, an adaptive language system could also include a knob or dial that allows a reader to scan through possible versions if the current adaption is not ideal.

One major hurdle in deploying systems using language models is the risk of hallucinations. We argue that such hallucinations necessitate human expert involvement. Rather than expert involvement being a limitation, though, we envision it improving human-human communication across the barriers that scientific language can impose. Science communication is ideally a conversation, not only a transmission of information [74]. Our hope is that requiring expert oversight will help science communicators quickly create

summaries that serve diverse audiences while also encouraging communicators to think deeply about the audiences they are reaching with their work.

8 CONCLUSION

In this paper, we investigate how general audience readers respond to scientific summaries written or generated at different levels of complexity. Across our three studies, using expert-written and machine-generated summaries, we show that the ideal text is based on a participant’s familiarity of a topic. Low familiarity participants rated the low complexity summaries as easiest to engage with. High familiarity participants rated the summaries equally regardless of complexity, while skipping more sections of low complexity summaries. We also find that using traditional generation or science communication techniques often leads to loss in information as language becomes less complex, but that new generative models are capable of generating plain text while explaining complex topics, retaining much of the information of higher complexity summaries. Our findings highlight the tradeoffs in adapting language complexity for different audiences and provide a path forward for communicating scientific information to a wider range of people.

ACKNOWLEDGMENTS

We thank Sarah Kahle for authoring the summaries and providing guidance on science communication practices and the Mechanical Turk annotators for their work on the project. We also thank the anonymous reviewers and members of the UWNLP and DUB community for their helpful feedback. This work was supported in part by the Office of Naval Research under MURI grant N00014-18-1-2670 and by a Twitch Research Fellowship.

REFERENCES

- [1] Takeshi Abekawa and Akiko Aizawa. 2016. SideNoter: Scholarly Paper Browsing System based on PDF Restructuring and Text Annotation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, Hideo Watanabe (Ed.). The COLING 2016 Organizing Committee, Osaka, Japan, 136–140. <https://aclanthology.org/C16-2029>
- [2] Eytan Adar, Carolyn Gearig, Ayshwarya Balasubramanian, and Jessica Hullman. 2017. PersaLog: Personalization of News Article Content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 3188–3200. <https://doi.org/10.1145/3025453.3025631>
- [3] Fiona Alderdice, Jenny McNeill, Toby J Lasserson, Elaine Beller, Margaret Carroll, Nina Smith, Paula M. Cuccaro, Efrat K. Gabay, Julie A. Boom, Roger W. Schvanveldt, and Cui Tao. 2020. Mining HPV Vaccine Knowledge Structures of Young Adults From Reddit Using Distributional Semantics and Pathfinder Networks. *Cancer Control : Journal of the Moffitt Cancer Center* 27 (2020).
- [4] Harrison Anzinger, Sarah Alexandra Elliott, and Lisa Hartling. 2020. Comparative Usability Analysis and Parental Preferences of Three Web-Based Knowledge Translation Tools: Multimethod Study. *Journal of Medical Internet Research* 22 (2020).
- [5] Tal August, Lauren Kim, Katharina Reinecke, and Noah A Smith. 2020. Writing Strategies for Science Communication: Data and Computational Analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*. 5327–5344.

- [8] Tal August, Nigini Oliveira, Chenhao Tan, Noah Smith, and Katharina Reinecke. 2018. Framing effects: Choice of slogans used to advertise online experiments can boost recruitment and lead to sample biases. *Proceedings of the ACM on Human-Computer Interaction 2*, CSCW (2018), 1–19.
- [9] Tal August and Katharina Reinecke. 2019. Pay attention, please: Formal language improves attention in volunteer and paid online experiments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–11.
- [10] Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating Scientific Definitions with Controllable Complexity. In *Association for Computational Linguistics*.
- [11] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 74 (sep 2023), 38 pages. <https://doi.org/10.1145/3589955>
- [12] Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. MedEASi: Finely Annotated Dataset and Models for Controllable Simplification of Medical Texts. *arXiv:2302.09155*
- [13] Amanda Baughan, Tal August, Naomi Yamashita, and Katharina Reinecke. 2020. Keep it Simple: How Visual Complexity and Preferences Impact Search Efficiency on Websites. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).
- [14] Angela Collier Bliss. 2019. Adult Science-Based Learning: The Intersection of Digital, Science, and Information Literacies. *Adult Learning* 30, 3 (2019), 128–137.
- [15] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *ArXiv abs/2005.14165* (2020).
- [16] Ivan Buljan, Mario Malički, Elizabeth Wager, Livia Puljak, Darko Hren, Frances Kellie, Helen West, Žarko Alfirević, and Ana Marušić. 2018. No difference in knowledge obtained from infographic or plain language summary of a Cochrane systematic review: three randomized controlled trials. *Journal of clinical epidemiology* 97 (2018), 86–94.
- [17] Ivan Buljan, Ružica Tokalić, Marija Roguljić, Irena Zakarija-Grković, Davorka Vrdoljak, Petra Milić, Livia Puljak, and Ana Marušić. 2020. Framing the numerical findings of Cochrane plain language summaries: two randomized controlled trials. *BMC Medical Research Methodology* 20 (2020).
- [18] Mengyao Cao, Yue Dong, and Jackie Chi Kit Cheung. 2022. Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization. In *Association for Computational Linguistics*.
- [19] Vinay K. Chaudhri, Britte Haugan Cheng, Adam Overholzer, Jeremy Roschelle, Aaron Spaulding, Peter E. Clark, Mark T. Greaves, and David Gunning. 2013. Inquire Biology: A Textbook that Answers Questions. *AI Mag.* 34 (2013), 55–72.
- [20] Rune Christensen. 2018. Cumulative Link Models for Ordinal Regression with the R Package ordinal. https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf R package version 2022.11-16. Data accessed: September 1, 2021.
- [21] Cochrane. 2021. New Standards for Plain Language Summaries. <https://consumers.cochrane.org/PLEACS> Date accessed: July 1, 2022.
- [22] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained Language Models for Sequential Sentence Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3693–3699. <https://doi.org/10.18653/v1/D19-1383>
- [23] Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics* (2014), 97–135.
- [24] Matthew Conlen, Alex Kale, and Jeffrey Heer. 2019. Capture & Analysis of Active Reading Behaviors for Interactive Articles on the Web. *Computer Graphics Forum* 38 (2019).
- [25] Robert Cudeck. 1996. Mixed-effects Models in the Study of Individual Differences with Repeated Measures Data. *Multivariate behavioral research* 31 3 (1996), 371–403.
- [26] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*. 307–318.
- [27] Judy Davis-Dorsey, Steven M. Ross, and Gary R. Morrison. 1991. The Role of Rewording and Context Personalization in the Solving of Mathematical Word Problems. *Journal of Educational Psychology* 83 (1991), 61–68.
- [28] Adriano Luiz de Souza Lima and Christiane Gresse von Wangenheim. 2022. Assessing the Visual Esthetics of User Interfaces: A Ten-Year Systematic Mapping. *International Journal of Human-Computer Interaction* 38 (2022), 144 – 164.
- [29] Dina Demner-Fushman and Noémie Elhadad. 2016. Aspiring to Unintended Consequences of Natural Language Processing: A Review of Recent Developments in Clinical and Consumer-Generated Text Processing. *Yearbook of medical informatics* 1 (2016), 224–233.
- [30] Michel C. Desmarais and R. Baker. 2012. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction* 22 (2012), 9–38.
- [31] Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level Simplification of Medical Texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4972–4984. <https://doi.org/10.18653/v1/2021.naacl-main.395>
- [32] Ashwin Devaraj, William Sheffield, Byron C. Wallace, and Junyi Jessy Li. 2022. Evaluating Factuality in Text Simplification. In *Association for Computational Linguistics*.
- [33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [34] Chrysanthe Dimarco, Graeme Hirst, Leo Wanner, and John Wilkinson. 2007. HealthDoc: Customizing patient information and health education by medical condition and personal characteristics.
- [35] Peter Dolog and Wolfgang Nejdl. 2003. Personalisation in Elena: How to cope with personalisation in distributed eLearning Networks. In *Proceedings of International Conference on Worldwide Coherent Workforce, Satisfied Users-New Services For Scientific Information*.
- [36] Moriah E. Ellen, John N Lavis, Michael G. Wilson, Jeremy M. Grimshaw, R. Brian Haynes, Mathieu Ouimet, Parminder Raina, and Russell Lindsay Gruen. 2014. Health System Decision Makers' Feedback on Summaries and Tools Supporting the Use of Systematic Reviews: A Qualitative Study. *Evidence & Policy: A Journal of Research, Debate and Practice* 10 (2014), 337–359.
- [37] Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. 2021. Summary-Source Proposition-level Alignment: Task, Datasets and Supervised Baseline. In *Proceedings of the 25th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Online, 310–322. <https://doi.org/10.18653/v1/2021.conll-1.25>
- [38] Samantha L. Finkelstein, Evelyn Yarzebinski, Callie Vaughn, Amy E. Ogan, and Justine Cassell. 2013. The Effects of Culturally Congruent Educational Technologies on Student Achievement. In *Artificial Intelligence in Education*.
- [39] Elena Forzani. 2016. *Individual Differences in Evaluating the Credibility of Online Information in Science: Contributions of Prior Knowledge, Gender, Socioeconomic Status, and Offline Reading Ability*. Ph.D. Dissertation. University of Connecticut.
- [40] Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Asli Çelikil-maz, and Yejin Choi. 2021. Discourse Understanding and Factual Consistency in Abstractive Summarization. In *European Chapter of the Association for Computational Linguistics*.
- [41] Tanya Goyal and Greg Durrett. 2021. Annotating and Modeling Fine-grained Factuality in Summarization. *ArXiv abs/2104.04302* (2021).
- [42] Yue Guo, Tal August, Gondy Leroy, Trevor A. Cohen, and Lucy Lu Wang. 2023. APPLS: A Meta-evaluation Testbed for Plain Language Summarization. *ArXiv abs/2305.14341* (2023).
- [43] Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2023. Retrieval augmentation of large language models for lay language generation. *Journal of Biomedical Informatics*, (2023).
- [44] Yue Guo, Weijian Qiu, Yizhong Wang, and Trevor A. Cohen. 2021. Automated Lay Language Summarization of Biomedical Scientific Reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [45] Choochart Haruechaiyasak and Chaianun Damrongrat. 2008. Article Recommendation Based on a Topic Model for Wikipedia Selection for Schools. In *International Conference on Asia-Pacific Digital Libraries*.
- [46] John R. Hauser, Glen L. Urban, Guilherme Liberali, and Michael Braun. 2009. Website Morphing. *Marketing Science* 28, 2 (mar 2009), 202–223.
- [47] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2021. Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021).
- [48] Marti A. Hearst, Emily Pedersen, Lekha Priya Patil, Elsie Lee, Paul Laskowski, and Steven L. Franconeri. 2020. An Evaluation of Semantically Grouped Word Cloud Designs. *IEEE Transactions on Visualization and Computer Graphics* 26, 9 (2020), 2748–2761.

- [49] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6 (1979), 65–70.
- [50] Abhinav Jain, Nitin Gupta, Shashank Mujumdar, Sameep Mehta, and Rishi Madhok. 2018. Content Driven Enrichment of Formal Text using Concept Definitions and Applications. *Proceedings of the 29th on Hypertext and Social Media* (2018).
- [51] Jasna Karacic, Pierpaolo Dondio, Ivan Buljan, Darko Hren, and Ana Marušić. 2019. Languages for different health information readers: multitrait-multimethod content analysis of Cochrane systematic reviews textual summary formats. *BMC Medical Research Methodology* 19 (2019).
- [52] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453* (2023).
- [53] Aleksandra Klačnja-Milićević, Boban Vesin, Mirjana Ivanović, and Zoran Budimac. 2011. E-Learning personalization based on hybrid recommendation strategy and learning style identification. *Computer Education* 56 (2011), 885–899.
- [54] Iulia Kotseruba and John K. Tsotsos. 2018. 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review* 53 (2018), 17–94.
- [55] Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization. In *European Chapter of the Association for Computational Linguistics*.
- [56] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics* 10 (2022), 163–177. https://doi.org/10.1162/tacl_a_00453
- [57] Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. SWiPE: A Dataset for Document-Level Simplification of Wikipedia Pages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Toronto, Canada, 10674–10695. <https://aclanthology.org/2023.acl-long.596>
- [58] GONDY Leroy, Stephen Helmreich, and James R. Cowie. 2010. The influence of text characteristics on perceived and actual difficulty of health information. *International journal of medical informatics* (2010), 438–449.
- [59] GONDY Leroy, Stephen Helmreich, James R. Cowie, Trudi Miller, and Wei Zheng. 2008. Evaluating online health information: Beyond readability formulas. In *American Medical Informatics Association Annual Symposium Proceedings*. American Medical Informatics Association, 394.
- [60] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. <https://aclanthology.org/W04-1013>
- [61] Gitte Lindgaard, Cathy Dudek, Devjani Sen, Livia Sumegi, and Patrick S. Noonan. 2011. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Transactions on Computer-Human Interaction (TOCHI)* 18 (2011), 1:1–1:30.
- [62] Magnus Lindstrom and Douglas M. Bates. 1990. Nonlinear mixed effects models for repeated measures data. *Biometrics* 46 3 (1990), 673–87.
- [63] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie (Yu-Yen) Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, F.Q. Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Michael Kinney, Aniket Kittur, Hyeonsu B Kang, Egor Klevak, Bailey Kuehl, Michael Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita R Rao, Paul L Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline M Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2023. The Semantic Reader Project: Augmenting Scholarly Documents through AI-Powered Interactive Reading Interfaces. *Commun. ACM* (2023).
- [64] Patrice Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *European Conference on Digital Libraries*.
- [65] Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic Decoding: (Un)supervised Neural Text Generation with Predicate Logic Constraints. In *North American Chapter of the Association for Computational Linguistics*.
- [66] Chrysanne Di Marco, Peter Bray, H. Dominic Covvey, Donald D. Cowan, Vic Di Ciccio, Eduard H. Hovy, Joan Lipa, and C. Yang. 2006. Authoring and Generation of Individualized Patient Education Materials. *American Medical Informatics Association Annual Symposium proceedings* (2006), 195–9.
- [67] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1906–1919. <https://doi.org/10.18653/v1/2020.acl-main.173>
- [68] Morten Moshagen and Meinald T. Thielsch. 2010. Facets of visual aesthetics. *International Journal of Human Computer Studies* 68 (2010), 689–709.
- [69] Mati Möttöus and David Jose Ribeiro Lamas. 2015. Aesthetics of Interaction Design: A Literature Review. In *Machine Intelligence and Digital Interaction '15*.
- [70] Randall Munroe. 2017. *Thing explainer complicated stuff in simple words*. John Murray.
- [71] Sonia Murthy, Kyle Lo, Daniel King, Chandra Bhagavatula, Bailey Kuehl, Sophie Johnson, Jonathan Borchardt, Daniel Weld, Tom Hope, and Doug Downey. 2022. ACCoRD: A Multi-Document Approach to Generating Diverse Descriptions of Scientific Concepts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Abu Dhabi, UAE, 200–213. <https://doi.org/10.18653/v1/2022.emnlp-demos.20>
- [72] Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 2727–2733. <https://doi.org/10.18653/v1/2021.eacl-main.235>
- [73] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy, 319–327.
- [74] Matthew C Nisbet and Dietram A Scheufele. 2009. What’s next for science communication? Promising directions and lingering distractions. *American journal of botany* 96, 10 (2009), 1767–1778.
- [75] Geoff Norman. 2010. Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education* 15 (2010), 625–632.
- [76] Emily Nunn and Stephen Pinfield. 2014. Lay summaries of open access journal articles: engaging with the general public on medical research. *Learned Publishing* 27, 3 (2014), 173–184.
- [77] Amy E. Ogan, Evelyn Yarzebinski, Roberto De Roock, Cristina E. Dum Dumaya, Michelle P. Banawan, and Ma. Mercedes T. Rodrigo. 2017. Proficiency and Preference Using Local Language with a Teachable Agent. In *International Journal of Artificial Intelligence in Education*.
- [78] Changhoon Oh, Jinhan Choi, Sungwoo Lee, Sohyun Park, Daeryong Kim, Jungwoo Song, Dongwhan Kim, Joonhwan Lee, and Bongwon Suh. 2020. Understanding User Perception of Automated News Generation System. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).
- [79] OpenAI. 2023. GPT-4 Technical Report. *ArXiv abs/2303.08774* (2023).
- [80] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 742, 20 pages. <https://doi.org/10.1145/3544548.3580841>
- [81] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (Bend, OR, USA) (UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 74, 18 pages. <https://doi.org/10.1145/3526113.3545616>
- [82] Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, 186–195.
- [83] Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training.
- [84] Mathieu Ranger and Karen Bultitude. 2016. “The kind of mildly curious sort of science interested person like me”: Science bloggers’ practices relating to audience recruitment. *Public Understanding of Science* 25, 3 (2016), 361–378.
- [85] Habeeb Ibrahim Abdul Razack, Sam T. Mathew, Fathinul Fikri Ahmad Saad, and Saleh A. Alqahtani. 2021. Artificial intelligence-assisted tools for redefining the communication landscape of the scholarly world. *Journal of Science Communication*.
- [86] Katharina Reinecke and Krzysztof Z Gajos. 2014. Quantifying visual preferences around the world. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014).
- [87] Katharina Reinecke, Patrick Minder, and Abraham Bernstein. 2011. MOCCA - a system that learns and recommends visual preferences based on cultural similarity. In *Intelligent User Interfaces*.
- [88] Nathan Sanders. 2013. Astrobites: Students Making Astrophysics Accessible. <https://blogs.scientificamerican.com/incubator/astrobites-students-making-astrophysics-accessible/> Date Accessed: August 1, 2021.
- [89] Nancy Santesso, Tamara Rader, Elin Strømme Nilsen, Claire Glenton, Sarah E. Rosenbaum, Agustín Ciapponi, Lorenzo Moja, Jordi Pardo Pardo, Qi Zhou, and Holger J. Schünemann. 2015. A summary to communicate evidence from

- systematic reviews to the public improved understanding and accessibility of information: a randomized controlled trial. *Journal of clinical epidemiology* 68 2 (2015), 182–90.
- [90] Lisa Scharrer, Rainer Bromme, Mary Anne Britt, and Marc Stadler. 2012. The seduction of easiness: How science depictions influence laypeople's reliance on their own evaluation of scientific information. *Learning and Instruction* 22 (2012), 231–243.
- [91] Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers Unite! Unsupervised Metrics for Reinforced Summarization Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3246–3256. <https://doi.org/10.18653/v1/D19-1320>
- [92] Chantal Shaib, Millicent Li, Sebastian Pathiyamattom Joseph, Iain James Marshall, Junyi Jessy Li, and Byron Wallace. 2023. Summarizing, Simplifying, and Synthesizing Medical Evidence using GPT-3 (with Varying Success). In *Annual Meeting of the Association for Computational Linguistics*.
- [93] Sarah Shailes. 2017. Plain-language Summaries of Research: Something for everyone. *eLife* 6 (mar 2017).
- [94] Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S Weld, and Doug Downey. 2022. VILA: Improving structured content extraction from scientific PDFs using visual layout groups. *Transactions of the Association for Computational Linguistics* 10 (2022), 376–392.
- [95] Leia Martinez Silvagnoli, Caroline Shepherd, James Pritchett, and Jason Gardner. 2022. Optimizing Readability and Format of Plain Language Summaries for Medical Research Articles: Cross-sectional Survey Study. *Journal of Medical Internet Research* 24 (2022).
- [96] Celette Sugg Skinner, Victor J. Strecher, and Harm J. Hospers. 1994. Physicians' recommendations for mammography: do tailored messages make a difference? *American journal of public health* 84 1 (1994), 43–9.
- [97] Neha Srikanth and Junyi Jessy Li. 2020. Elaborative Simplification: Content Addition and Explanation Generation in Text Simplification. In *Findings of the Association for Computational Linguistics*.
- [98] Jackson Stokes, Tal August, Robert A Marver, Alexei Czeskis, Franziska Roesner, Tadayoshi Kohno, and Katharina Reinecke. 2023. How language formality in security and privacy interfaces impacts intended compliance. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [99] Marlene Stoll, Martin Kerwer, Klaus Lieb, and Anita Chasiotis. 2022. Plain language summaries: A systematic review of theory, guidelines and empirical research. *PLoS ONE* 17 (2022).
- [100] Victor J. Strecher, Matthew W Kreuter, D.J. den Boer, Sarah C Kobrin, Harm J. Hospers, and Celette Sugg Skinner. 1994. The effects of computer-tailored smoking cessation messages in family practice settings. *The Journal of family practice* 39 3 (1994), 262–70.
- [101] Glen L. Urban, Guilherme (Gui) Liberali, Erin MacDonald, Robert Bordley, and John R. Hauser. 2014. Morphing Banner Advertising. *Marketing Science* 33, 1 (jan 2014), 27–46.
- [102] Shaun Wallace, Zoya Bylinskii, Jonathan Dobres, Bernard Kerr, Sam Berlow, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Dave B. Miller, Jeff Huang, and Ben D. Sawyer. 2022. Towards Individuated Reading Experiences: Different Fonts Increase Reading Speed for Different Individuals. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29 (2022), 1 – 56.
- [103] Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023. Pre-trained Language Models in Biomedical Domain: A Systematic Survey. *ACM Comput. Surv.* 56, 3, Article 55 (oct 2023), 52 pages. <https://doi.org/10.1145/3611651>
- [104] Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large Language Models are Diverse Role-Players for Summarization Evaluation. In *Natural Language Processing and Chinese Computing*. Springer Nature Switzerland, Cham, 695–707.
- [105] Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023. Elaborative Simplification as Implicit Questions Under Discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 5525–5537. <https://doi.org/10.18653/v1/2023.emnlp-main.336>
- [106] Teng Ye, Katharina Reinecke, and Lionel P. Robert. 2017. Personalized Feedback Versus Money: The Effect on Reliability of Subjective Data in Online Experimental Platforms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (Portland, Oregon, USA) (CSCW '17 Companion)*. Association for Computing Machinery, New York, NY, USA, 343–346. <https://doi.org/10.1145/3022198.3026339>
- [107] Chen-Hsiang Yu and Robert C. Miller. 2010. Enhancing web page readability for non-native readers. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010).
- [108] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=SkeHuCVFDr>
- [109] Tiancheng Zhao and Kyusong Lee. 2020. Talk to Papers: Bringing Neural Question Answering to Academic Search. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 30–36. <https://doi.org/10.18653/v1/2020.acl-demos.5>

A AUTOMATED COMPLEXITY MEASURES

Below we describe in more details the automated complexity measures used in §4.1.2.

Thing Explainer out-of-vocabulary (TE): We count the ratio of words outside the top 1,000 most common words in English. The words are based on Wiktionary’s contemporary fiction frequency list.¹² This method was popularized by the popular book *Thing Explainer*, which explains scientific concepts using only the 1,000 most frequent words in English [70].

Function words In medical communication, the proportion of function words (e.g., prepositions, auxiliary or verbs) was found to be positively correlated with perceived and actual readability [58, 59]. We measure the proportion of function words in a sentence using `scispacy` [73].

Language model perplexity (GPT ppl.) Language models are systems for predicting words in a sequence. The perplexity of the model is a measure of how different a sequence of text is from the language the model was trained on. Perplexity has been found to correlate with perceived and actual reading difficulty [23, 82]. We use the GPT model [83] to measure language model perplexity, as it was trained on common English (as opposed to scientific text).

B ORDINAL REGRESSION FOR LIKERT-SCALE VARIABLES

As our reading experience measures were measured on a Likert-style scale, the linear mixed effects model (LMM) estimates could be ill-suited for analysis, especially if these measures were not sufficiently normally distributed. As an alternative, we additionally fit analogous cumulative link mixed-effects models (CLMM) from the `ORDINAL` R package [20] and conducted likelihood ratio tests, which are similar to *F*-tests but more conservative, on the interaction term of complexity level of article familiarity.

To accurately identify the effect complexity has on our measures and its interaction with topic familiarity, we define two models for each measure. Each model includes the same random effects of paper ID and participant ID to control for variation among papers and participants.

- (1) LMM_{full} : Containing fixed effects for the complexity version, topic familiarity, an interaction term for familiarity and complexity, and random effects for paper and participant IDs.
- (2) LMM_{none} : Containing a fixed effect for topic familiarity and random effects for paper and participant IDs.

With these models we evaluate how complexity affects reading measures (e.g., reading ease) by comparing the model goodness-of-fit between LMM_{full} and LMM_{none} using the χ^2 likelihood-ratio test. If LMM_{full} has a significantly stronger fit, this suggests that complexity has a significant effect on that reading measure.

Table 5 lists the *p*-values for the likelihood ratio tests on the CLMM and LMM models. The *p*-values are similar across the two

methods, with the one exception being a significant difference in understanding for Study 3. To confirm our findings of differences across complexity measures, we additionally ran Mann–Whitney *U*-tests on the reading experience ratings. While the studies were within-subjects, we treated the data as unpaired because familiarity ratings differed across the same participant, and therefore were not grouped together. While these tests did not control for participant or paper random effects (as the post-hoc *t*-tests we report in the results do), the findings remained similar to those reported in Tables 8, 9 and 10. Following prior work [11, 47, 75], we report results from the parametric tests (i.e., LMMs and pairwise difference *t*-tests) in the paper.

C GENERATING SUMMARIES - STUDY 2

GPT-3 was not designed to explicitly vary text complexity, so while generations might vary naturally in complexity due to the changes in prompt, there is no guarantee that prompts will align with complexity (i.e., prompting GPT-3 with “Summarize for a first grade student” will not necessarily lead to lower complexity than prompting with “tenth grade student”). In a preliminary analysis of the summaries, we found that the summaries, while tending toward simpler with lower grades, could still be quite complex in the first grade prompted version and much simpler at higher grade levels. Table 6 provides examples of generations and associated prompts.

There are automatic methods for scientific information extraction [22] and PDF parsing [64, 94] that could in the future be used to extract information directly from a research paper PDF. We leave such extensions to future work, as our goal was to explore the feasibility of automatically adjusting language complexity. Any errors introduced by other automated methods (e.g., incorrect text from PDF parsing) could muddy our ability to identify how alternate complexity levels perform in our envisioned context.

D GENERATING SUMMARIES - STUDY 3

The full prompts were:

- **Low:** You are a helpful assistant who will rewrite 5-10 scientific sentences for a reader who is not at all familiar with the sentence’s topic. You will be given one sentence after another. For each sentence, define any necessary terms and provide any background knowledge that a reader who is not at all familiar with this topic might need. Your target grade level is 5th grade. You do not need to explain a term or concept you have already explained. Your goal is to help a lay reader understand all the information contained in the original sentence. Below is an example.
Input sentence: The accretion of terrestrial planets from the solar nebula has been a subject of extensive theoretical modeling and simulation.
Simplified output: The growth of rocky planets like the Earth by adding particles, called accretion, from the cloud of dust and gas surrounding the early Sun, called the solar nebula, has been studied a lot using theoretical calculations and computer simulations. Theoretical calculations refer to math that is used to make predictions about how things happen based on a particular theory. Computer simulations are programs that model how a system works.

¹²https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Contemporary_fiction

Measure	Model	Study 1	<i>p</i>	Study 2	<i>p</i>	Study 3	<i>p</i>
Reading Ease	CLMM	168.28	<0.001	94.39	<0.001	14.04	<0.01
	LMM	181.04	<0.001	108.33	<0.001	15.92	<0.05
Understanding	CLMM	118.23	<0.001	51.22	<0.001	9.41	0.116
	LMM	134.63	<0.001	55.08	<0.001	12.39	<0.05
Interest	CLMM	57.38	<0.001	18.22	<0.01	10.11	0.116
	LMM	61.64	<0.001	26.88	<0.001	9.32	0.107
Value	CLMM	18.66	<0.001	11.09	<0.05	7.08	0.132
	LMM	19.64	<0.001	10.65	<0.01	8.82	0.107

Table 5: Likelihood ratio test statistics and *p*-values for likelihood ratio test of cumulative link (CLMM) and linear (LMM) mixed-effects models. Significant values are bolded. *p*-values are adjusted using Holm-correction.

Prompt Level	Summary
Second-grade student	In this study, people who could do more than 40 pushups were associated with a reduced risk of having a heart problem in the next 10 years, compared to those who could do fewer than 10. The study also found that people who could do more than 10 pushups had a reduced risk of having a heart problem, and that the risk decreased as the number of pushups increased.
Seventh-grade student	This study found that people who could do more than 40 pushups were less likely to have a heart problem in the next 10 years. The study also found that the more pushups someone could do, the less likely they were to have a heart problem.
College student	This study found that people who could do more than 40 pushups had a significantly lower risk of having a cardiovascular disease event, compared to those who could do 10 or fewer. This may be because those who could do more pushups also had better baseline cardiovascular disease risk factors. They also found that people who could do 11 or more pushups at the start of the study had a lower risk of having a cardiovascular disease event during the study.

Table 6: Examples of the generated summaries with different prompts for study 2 using GPT-3. Note that the prompts were not used to select complexity levels. This part of the summaries was under the heading “What did the paper find?” Notice how the second grade prompt is slightly longer and uses larger words (e.g., “associated with reduced risk” compared to “less likely to”) than the seventh grade prompt. At the same time, the college student prompt uses more complex language (e.g., “cardiovascular disease event”) compared to both other generations.

- Medium:** You are a helpful assistant who will rewrite 5-10 scientific sentences for a reader who is very familiar with the sentence’s topic. You will be given one sentence after another. For each sentence, define any necessary terms and provide any background knowledge that a reader who is very familiar with this topic might need. Your target grade level is a college-educated adult. You do not need to explain a term or concept you have already explained or that the reader is likely to know. Your goal is to help the reader understand all the information contained in the original sentence. Below is an example.

Input sentence: The accretion of terrestrial planets from the solar nebula has been a subject of extensive theoretical modeling and simulation.

Simplified output: The formation of terrestrial planets through accumulating dust, gas, and debris, called accretion, from the solar nebula, has been studied extensively using theoretical calculations and computer simulations.

E FACTUALITY IN GENERATED SUMMARIES

Out of 120 generated summaries in study 2 (6 sections × 10 papers × 2 complexities), 22 were labelled as containing any hallucinated content. The labels were mutually exclusive. There were three types of hallucinations we identified: correct information not from the original text, incorrect information not from the original text and reversing the direction of findings. Table 7 includes examples of these three hallucinations.

The extent and kind of hallucinations in our summaries can tell us what risk such hallucinations pose and how much effort an expert must invest to make the summaries publishable. For example, if the majority of hallucinations are new but correct information (a common type of hallucination [18]), then they pose less of a risk and require less expert knowledge to fix than if the hallucinations instead reverse the direction of a found effect (another type of hallucination [32]). We generated summaries with no restriction on hallucinated content. After generation, one author labelled all generations for hallucinated content.

Including correct information not from the original text occurred in 3 hallucinations. Usually these hallucinations included text about the study findings with no associated text from the original source text, or else hallucinated the existence of graphs from additional studies (e.g., “This chart shows the probation rates of the US population ...”). These hallucinations reported correct information, even though the information was not reported in the source text.

9 hallucinations included incorrect information not from the original text. These hallucinations added unrelated findings to the summary that were not reported in the study. Examples include hallucinating an association between asthma and nut intake, while the original article reported on nut intake and neuropsychological development.

Including correct and incorrect information not from the original text are similar to *extrinsic* hallucinations in the summarization literature [41], or *information insertion* in the simplification literature [32]. Both refer to hallucinations adding information not found in the original source.

Reversing the direction of findings occurred in 5 hallucinations. These hallucinations reported the exact opposite result than was reported in the original study. These hallucinations are considered *intrinsic* hallucinations, or *information substitution* which are hallucinations that include information in direct contrast to the original source [32, 67].

These three types of hallucinations are well-documented in literature studying generative model hallucinations [18, 32, 41, 67]. We add to this previous literature by showing how such hallucinations occur in this reading context.

We also explored using automated methods to identify hallucinations. We tried two commonly used automated measures for hallucinations, SummaQA [91] and entity-level F1 [72]. SummaQA uses a BERT-based question answering model to answer questions extracted from the source text with the summary text. We use the original extracted sentences as the source text. Entity-level F1 measures the number of entities that occur in a generated summary compared to the ground truth summary. We use *scispacy* [73] to extract entities. We observed no significant differences in either score between generated summaries with or without hallucinations (two-sided *t*-test $t_{118} = 0.04$, $p = 0.972$ for SummaQA F-score, $t_{118} = 1.90$, $p = 0.119$ for entity-level F1 after Holm correction). When inspecting the scores of generations, we also observed that both scores skewed positively (i.e., measured less hallucinated content) towards summaries that had language more similar to the original. This led to the scores negatively impacting the lower complexity summaries since they used language more distinct from the original researcher version. Based on these results, we did not use any automated factuality scores to curate the summaries.

F PAIRWISE TEST STATISTICS

Below we report all test statistics for pairwise comparisons in the three studies.

Hallucination type	Example	Reason	% Generations
Incorrect additional information	The study found that the babies of women who ate nuts during pregnancy were less likely to have certain health problems.	Nothing in study about health problems	7.5%
Correct additional information	These cells work together to make sure that we feel pain when we are hurt. This is important because it helps us to avoid getting hurt again.	Nothing in original article about the importance of pain sensation	2.5%
Reverse direction of findings	This study found that spending more time playing video games can lead to more aggressive behavior.	Finding was that time spent playing video games did not lead to more aggressive behavior	4.2%

Table 7: Three types of hallucinations encountered in our generated summaries in study 2 (with GPT-3).

	Familiarity	d^{Lo-Me}	p	d^{Lo-Hi}	p	d^{Me-Hi}	p
Reading Ease	1	0.554	<0.0001	1.490	<0.0001	0.936	<0.0001
	2	0.103	0.621	0.782	0.001	0.679	0.003
	3	0.197	0.391	0.695	0.013	0.498	0.059
	4	0.101	0.817	0.609	0.544	0.508	0.588
	All	0.238	0.069	0.894	<0.0001	0.655	<0.0001
Understanding	1	0.458	<0.0001	1.160	<0.0001	0.701	<0.0001
	2	0.022	0.910	0.693	0.002	0.671	0.002
	3	0.172	0.597	0.391	0.240	0.219	0.597
	4	0.160	1.0	0.127	1.0	-0.033	1.0
	All	0.203	0.094	0.593	<0.0001	0.390	0.006
Interest	1	0.296	0.021	0.943	<0.0001	0.647	<0.0001
	2	-0.007	0.975	0.298	0.593	0.305	0.593
	3	0.024	1.0	-0.009	1.0	-0.033	1.0
	4	0.864	0.220	0.261	0.603	-0.603	0.520
	All	0.294	0.085	0.373	0.042	0.079	0.613
Value	1	0.314	0.020	0.509	<0.0001	0.195	0.104
	2	-0.012	1.0	0.009	1.0	0.021	1.0
	3	-0.087	1.0	-0.099	1.0	-0.012	1.0
	4	0.329	1.0	-0.123	1.0	-0.451	1.0
	All	0.136	0.996	0.074	1.00	-0.062	1.00
Skipped Sections	1	0.041	0.994	0.051	0.994	0.009	0.994
	2	-0.107	0.813	-0.007	0.941	0.099	0.813
	3	-0.252	0.056	-0.008	0.943	0.244	0.056
	4	0.285	0.202	0.682	0.008	0.398	0.202
	All	-0.008	0.892	0.179	0.020	0.188	0.020
Article Requests	1	0.009	0.768	0.056	0.206	0.047	0.270
	2	-0.026	0.659	-0.184	0.007	-0.159	0.017
	3	-0.078	0.439	0.027	0.685	0.105	0.287
	4	0.063	1.0	0.018	1.0	-0.045	1.0
	All	-0.008	1.0	-0.021	1.0	-0.013	1.0

Table 8: Post-hoc (two-sided) tests for pairwise differences in fixed-effects estimates between complexity versions and across all participant topic familiarities for study 1 with expert-written summaries. ‘All’ topic familiarity refers to pairwise differences across complexity levels without a topic familiarity subgroup (e.g., average difference across complexity levels.) This table reports the difference in fixed-effects estimates $i - j$ and Holm-Bonferroni-corrected p -values [49] under our mixed-effects model, where i and j correspond to complexity options. — $Lo = Low$, $Me = Medium$, and $Hi = High$. Statistically significant p -values are bold. For example, in Table 8 in the column for d^{Lo-Hi} and row for “Reading Ease,” and “1” in topic familiarity we can interpret the result as participants with a 1 topic familiarity rated the Low complexity, on average, 1.490 points higher for reading ease (out of 5) compared to the High complexity when controlling for participant and paper.

	Familiarity	d^{Lo-Me}	p	d^{Lo-Hi}	p	d^{Me-Hi}	p
Reading Ease	1	1.385	<0.0001	1.645	<0.0001	0.260	0.120
	2	0.310	0.274	0.660	0.024	0.350	0.274
	3	0.392	0.101	0.321	0.161	-0.071	0.683
	4	0.057	1.0	-0.045	1.0	-0.102	1.0
	5	0.216	1.0	0.093	1.0	-0.122	1.0
	All	0.472	<0.0001	0.535	<0.0001	0.063	0.455
Understanding	1	0.836	<0.0001	1.103	<0.0001	0.267	0.110
	2	0.369	0.267	0.630	0.035	0.262	0.269
	3	0.035	0.850	0.223	0.678	0.188	0.678
	4	0.030	1.0	-0.077	1.0	-0.107	1.0
	5	0.127	0.702	-0.266	0.702	-0.394	0.514
	All	0.279	0.004	0.323	0.001	0.043	0.622
Interest	1	0.590	0.001	0.909	<0.0001	0.319	0.055
	2	0.134	1.0	0.125	1.0	-0.009	1.0
	3	-0.047	1.0	-0.064	1.0	-0.018	1.0
	4	-0.004	1.0	-0.009	1.0	-0.006	1.0
	5	0.251	1.0	0.052	1.0	-0.199	1.0
	All	0.185	0.077	0.202	0.077	0.017	0.818
Value	1	0.069	0.674	0.407	0.031	0.339	0.079
	2	-0.220	0.970	0.009	0.970	0.229	0.970
	3	-0.173	1.0	-0.161	1.0	0.012	1.0
	4	0.151	0.665	-0.085	0.665	-0.236	0.427
	5	0.270	0.616	-0.101	0.720	-0.371	0.570
	All	0.020	1.0	0.014	1.0	-0.006	1.0
Skipped Sections	1	0.083	1.0	-0.058	1.0	-0.142	1.0
	2	-0.333	0.645	-0.143	0.924	0.190	0.924
	3	0.100	1.0	-0.006	1.0	-0.106	1.0
	4	0.367	0.123	0.277	0.234	-0.090	0.631
	5	0.235	0.424	0.900	0.011	0.665	0.066
	All	0.090	0.553	0.194	0.138	0.103	0.553
Article Requests	1	0.095	0.299	0.057	0.603	-0.038	0.603
	2	0.214	0.036	0.001	0.991	-0.213	0.036
	3	0.001	1.0	-0.046	1.0	-0.047	1.0
	4	0.048	1.0	0.006	1.0	-0.042	1.0
	5	0.069	1.0	0.042	1.0	-0.026	1.0
	All	0.085	0.015	0.012	0.699	-0.073	0.035

Table 9: Study 2 with machine-generated summaries and no restriction on information content. See Table 8 for examples of pairwise comparison interpretation.

	Familiarity	d^{Lo-Me}	p	d^{Lo-Hi}	p	d^{Me-Hi}	p
Reading Ease	1	0.149	0.260	0.362	0.019	0.213	0.182
	2	0.340	0.165	0.669	0.002	0.330	0.165
	3	-0.134	1.0	-0.075	1.0	0.059	1.0
	4	0.235	1.0	0.201	1.0	-0.034	1.0
	5	0.319	1.0	-0.327	1.0	-0.646	1.0
	All	0.182	1.0	0.166	1.0	-0.016	1.0
Understanding	1	0.186	0.147	0.420	0.003	0.234	0.111
	2	0.033	1.0	0.174	1.0	0.141	1.0
	3	-0.169	1.0	-0.062	1.0	0.107	1.0
	4	0.298	0.859	0.523	0.527	0.225	0.859
	5	-0.228	1.0	-0.456	1.0	-0.228	1.0
	All	0.024	1.0	0.120	1.0	0.096	1.0
Interest	1	-0.018	0.902	0.295	0.091	0.313	0.080
	2	0.141	0.777	0.342	0.373	0.201	0.777
	3	-0.291	0.613	-0.173	0.853	0.117	0.853
	4	-0.162	1.0	0.103	1.0	0.265	1.0
	5	0.834	1.0	0.383	1.0	-0.450	1.0
	All	0.101	1.0	0.190	1.0	0.089	1.0
Value	1	-0.014	0.922	0.213	0.269	0.226	0.269
	2	-0.025	1.0	0.165	1.0	0.190	1.0
	3	-0.380	0.180	0.037	0.856	0.417	0.180
	4	0.494	0.454	0.870	0.116	0.376	0.454
	5	2.245	0.177	2.385	0.200	0.139	0.906
	All	0.464	0.139	0.734	0.051	0.270	0.284
Skipped Sections	1	0.023	1.0	0.110	1.0	0.087	1.0
	2	-0.062	1.0	-0.015	1.0	0.047	1.0
	3	-0.583	0.004	-0.110	0.529	0.472	0.026
	4	0.167	1.0	0.074	1.0	-0.093	1.0
	5	-0.055	1.0	-0.203	1.0	-0.148	1.0
	All	-0.102	1.0	-0.029	1.0	0.073	1.0
Article Requests	1	0.108	0.023	0.110	0.023	0.002	0.963
	2	-0.015	0.805	-0.079	0.618	-0.064	0.652
	3	0.101	0.347	0.082	0.352	-0.018	0.783
	4	-0.135	0.683	-0.000	1.0	0.135	0.683
	5	-0.100	1.0	0.039	1.0	0.138	1.0
	All	-0.008	1.0	0.030	1.0	0.039	1.0

Table 10: Study 3 with machine-generated summaries and restriction on information content. See Table 8 for examples of pairwise comparison interpretation.