

Measuring the Impact of LLM-Generated Scientific Summaries for Neurodivergent Readers

Abraham Ur Rahman¹, Skyler Koba², Audrey Michal³, Steven R. Wilson³

¹Clark University, ²Penn State University, ³University of Michigan-Flint
arahman@clarku.edu, skoba.ux@gmail.com, {almichal, steverw}@umich.edu

Abstract

Language technologies increasingly mediate access to scientific information, yet their impact depends on whether they improve real reading experiences for diverse audiences. We study large language model (LLM)-generated scientific summaries as an accessibility intervention for neurodivergent adults, focusing on readers with ADHD and/or autism. Using a mixed-methods design, we first conducted a formative qualitative study with 10 neurodivergent adults to identify design needs. We then ran a quantitative experiment with 200 neurodivergent adults comparing five summary styles across lower- and higher-complexity scientific texts, measuring comprehension, trust, engagement, and cognitive load. Results show that writing-style effects depend on text complexity: structured LLM-generated summaries improved comprehension for higher-complexity articles but reduced comprehension for lower-complexity articles relative to a baseline human-generated extractive summary. Trust increased for simpler LLM summary styles, while engagement and cognitive load showed fewer robust writing-style effects. Our findings suggest that accessible NLP systems should adapt explanatory support to reader needs and text difficulty, rather than offering a single “best” summary style.

1 Introduction

“Eighty-five percent of medical research goes to waste, partly because it is not appropriately communicated to stakeholders. This represents a critical issue ... because individuals who are impacted by research should be able to readily access that research.” – Gasparini et al. (2025)

Scientific knowledge can support informed decision-making about health, education, climate, policy, and everyday life. However, making research papers publicly available does not guarantee that readers can understand, trust, or use them. Bar-



Figure 1: We present one baseline (extractive) and four LLM-generated summary styles to neurodivergent adults in order to measure how summary style impacts comprehension, trust, cognitive load, and engagement.

riers such as technical jargon, dense structure, ambiguous tone, and long-form presentation can make scientific communication inaccessible, particularly for readers whose information-processing needs are not well supported by conventional styles.

Large language models (LLMs) are increasingly used to summarize scientific documents (Liu et al., 2024; Zhang et al., 2025). This creates an opportu-

nity for positive social impact: NLP systems could help more people engage with research that affects their lives. Yet this opportunity also raises an evaluation challenge: how can we measure whether an NLP system is actually beneficial for a community? Standard summarization metrics such as factual consistency, conciseness, and coverage are necessary but insufficient. For accessibility-oriented applications, impact also depends on reader comprehension, cognitive load, trust, engagement, autonomy, and the ability to verify information.

In this paper, we study LLM-generated summaries of scientific articles for neurodivergent readers, focusing on adults with ADHD and/or autism. Prior work suggests that autistic readers may experience difficulty with ambiguous meaning, tone, and long syntactic structures, while readers with ADHD may encounter challenges with sustained attention, task structure, and executive-function demands during reading (Hadden et al., 2023; Yaneva et al., 2019; Parks et al., 2022; Miranda et al., 2017). These experiences vary widely across individuals, and many neurodivergent readers develop their own strategies for navigating dense text. Designing accessible NLP systems therefore requires both community-informed design and empirical measurement. By making scientific information more accessible, our work aligns with broader social good goals of Education and Inclusion (Karamolegkou et al., 2026).

Our central research question in this work is: *How does the style of LLM-generated scientific summaries affect neurodivergent readers’ comprehension, trust, engagement, and cognitive load?* While working to answer this, we make the following contributions: (1) We present a formative study with neurodivergent adults that identifies design considerations for LLM-generated scientific summary styles; (2) we operationalize these considerations through five controlled styles for scientific-document summarization; and (3) we evaluate these styles in a randomized study with 200 neurodivergent adults, measuring comprehension, trust, engagement, and cognitive load for lower- and higher-complexity scientific texts.

We find that changes in summary format led to modest, outcome-dependent effects rather than uniform improvements. For complex scientific texts, summaries that added clearer structure and terminology support improved objective comprehension relative to an abstract-style baseline. For lower-complexity source-texts, however, these added sup-

ports were not beneficial and were sometimes associated with lower comprehension. Trust was higher for simplified and general-purpose LLM-generated summaries, whereas engagement and cognitive load showed few robust differences across formats after covariate adjustment and multiple-comparisons correction. Overall, these findings suggest that evaluating LLM-generated scientific summaries for accessibility requires accounting for source-text complexity and reader-centered outcomes, not only summary fluency or informativeness.

2 Related Work

2.1 NLP for Accessible Scientific Communication

Automatic summarization has long aimed to condense information while preserving important content. In scientific summarization, systems often target abstracts (Cohan et al., 2018), lay summaries for nonexpert readers (Goldsack et al., 2022), or evidence-synthesis summaries that aggregate findings across multiple studies (DeYoung et al., 2021; Ramprasad et al., 2023). However, a summary that is accurate and concise may still be inaccessible if it assumes background knowledge, lacks structure, or uses a tone that undermines reader trust. Prior work on accessible research summaries argues that communication should be designed with affected communities rather than merely simplified for them (Gasparini et al., 2025). This is especially important when scientific information concerns health, education, environmental risk, or other socially consequential domains.

LLMs expand the design space for accessible summaries because they can generate multiple styles from the same source: question-answer summaries (Tachioka, 2024), summaries paired with automatically generated glossaries (Syed et al., 2024), persona-tailored summaries (Mullick et al., 2024), and summaries at controlled reading levels (Ribeiro et al., 2023). At the same time, LLMs can produce incorrect or misleading content (Xu et al., 2024; Ryser et al., 2025), and readers may distrust AI-generated information (Passi and Vorvoreanu, 2022). The accessibility of LLM summarization therefore cannot be inferred from fluency alone.

2.2 Neurodivergent Readers and Information Processing

Neurodivergent readers are not a monolithic group. Autistic readers may differ in how they process tone, ambiguity, figurative language, and long sentences (Hadden et al., 2023; Yaneva et al., 2019). Written text lacks many of the facial, gestural, and prosodic cues available in face-to-face interaction, which can complicate interpretation for some readers. Readers with ADHD may experience difficulties related to sustained attention, working memory, reading speed, and executive function (Parks et al., 2022; Miranda et al., 2017). Such challenges may be intensified by long texts, low structure, and high cognitive demand.

Prior research also emphasizes that reading difficulties are not solely located within individuals. External factors—including text style, task demands, motivation, and available scaffolding—shape reading experiences (Parks et al., 2022). This perspective motivates our focus on summary style as an accessibility intervention. Rather than asking whether neurodivergent readers can adapt to conventional scientific writing, we ask how NLP systems might adapt scientific communication to diverse reading needs.

2.3 Self-Diagnosis and Community-Informed Research

Many adults identify as autistic or as having ADHD without a formal clinical diagnosis. Barriers include cost, access to specialists, prior misdiagnosis, stigma, and the perception that diagnosis may be too late or not useful (Routledge, 2024; Sarrett, 2016; Eagle and Ringland, 2023). These barriers are especially salient for women and other groups historically underdiagnosed or misdiagnosed. In this work, we therefore include both clinically diagnosed and self-diagnosed participants, while treating diagnostic status as self-reported rather than clinically verified.

3 Formative Qualitative Study

We first conducted a formative qualitative study to identify neurodivergent readers' needs and expectations for scientific article summaries. The aim was not to estimate how common these preferences were across the broader population, but to inform the design of summary styles and survey measures for the larger experiment.

3.1 Participants and Recruitment

We recruited 10 neurodivergent adults through Prolific using prescreeners to identify participants who were at least 18 years old, fluent in English, and self-identified or were clinically diagnosed with ADHD, autism, or both. Participants signed up using Prolific IDs as pseudonyms, received informed-consent materials through Prolific messaging, and completed sessions over Zoom. Participants were not recruited for expertise in the source-paper domains; the focus remained on general neurodivergent readers who might use summarization tools to access general scientific information.

3.2 Procedure & Formative Materials

Each session lasted approximately 60 minutes and included a reading activity, interviews, and a facilitated design activity. Participants first read a standard, neutral-toned AI-generated summary of a simple scientific article (Tommasi et al., 2025) and then discussed their reading strategies, points of confusion, trust judgments, and preferences for scientific summaries.

Participants then completed a facilitated design activity addressing personal reading challenges and participated in a semi-structured interview (see Appendix C). As part of this activity, participants were also shown sample excerpts of two additional prototype AI-generated summary styles of the same paper (Tommasi et al., 2025). In this work, we focus on the interview findings, including participants' reactions to these two prototype styles:

1. **LLM-LOGOS:** A Q&A-style summary emphasizing information grouping, definitions, citations, and selected source quotations.
2. **LLM-PATHOS:** A Q&A-style summary using a professional but conversational persona designed to support engagement and comprehension.

Prompt Generation Details. Both prototype summaries were generated using GPT-4o with standardized system prompts designed to isolate the impact of structural scaffolding versus conversational tone (see Appendix A for the full prompt templates). Explicitly defining these prompt structures is critical, as prompt design directly dictates summary formatting, reading load, and tone. These variables directly influenced participant feedback in this phase. Specifically, feedback influenced prompt refinements described in Section 3.4, including moderat-

ing persona tone and varying structural scaffolding, further finalized into the five summary-style conditions used in the quantitative experiment (Section 4.3).

3.3 Qualitative Analysis

Interview transcripts were analyzed using reflexive thematic analysis from a constructivist perspective (Braun and Clarke, 2021), treating themes as interpretive patterns developed through researchers' engagement with the data rather than as objectively fixed categories (Byrne, 2022).

Two authors of the paper first independently reviewed the transcripts and developed an initial coding framework focusing on participants' experiences of engagement, comprehension, cognitive load, trust, and summary presentation. Related codes were then grouped into candidate themes, considered in relation to participants' reported neurotypes, and refined through collective discussion. As the analysis followed a reflexive thematic framework, inter-rater reliability was not calculated.

3.4 Qualitative Findings

The analysis identified three themes that most directly informed the subsequent quantitative experiment: tone and engagement; comprehension, cognitive load, and structure; and trust and source credibility.

Tone and engagement. Participants' reactions to tone were mixed. Professional language was generally seen as fitting the scientific material, although some participants found it stale or too similar to the original article. In contrast, lighter, conversational language brought a sense of ease to the presentation, with participants describing it as "very catchy" and "attention-grabbing." More pronounced persona-driven language, however, made the presentation feel noticeably different from conventional scientific communication. One participant described a summary as "too casual, too informal," while another felt that the material was "presented in a silly way."

Comprehension, cognitive load, and structure. Participants frequently connected comprehension with cognitive effort. Undefined terminology and jargon-heavy passages led to rereading, guessing, or skipping; as one participant explained, "If I had a definition, then I could follow the rest." Participants also described different ways of managing this effort. Some ADHD participants skimmed for

key points to conserve energy, with one noting, "I just want to know the important bits. Bullet points are great for that kind of thing." In contrast, some autistic participants described reading dense sections fully and rereading when connections were unclear. Participants did not uniformly reject detail; rather, they wanted it to be well structured and chunked. One participant emphasized, "I love categorization. And putting things into boxes," suggesting that organization could make additional detail easier to manage.

Trust and source credibility. Trust depended on both content and presentation. Some participants felt that casual summaries of serious scientific topics created a credibility mismatch; one participant noted feeling "less confident" in material presented with an overly casual tone. Others expressed broader concerns about AI-generated information, particularly "how often it hallucinates and makes stuff up." Participants therefore wanted summaries to clearly signal where information came from, with one describing "the lack of source material" as a major factor in their trust. Together, these responses suggest that credibility depended not only on accuracy, but also on whether the summary appeared professional, traceable, and appropriately cautious.

Implications for the experiment. The qualitative findings informed the development of the quantitative experiment rather than serving as generalizable conclusions. Participant feedback guided several design decisions, including adjustments to persona style in response to concerns about overly casual tone, as well as variations in summary structure and jargon level to examine their effects on comprehension and cognitive load. Participants' emphasis on clear scaffolding also motivated the inclusion of multiple summary structures. Finally, the recurring relationships participants described among engagement, comprehension, cognitive load, and trust informed the selection of these constructs as primary outcome measures in the quantitative study.

4 Quantitative Experiment

4.1 Design

We conducted a $2 \times 2 \times 5$ mixed factorial experiment. The within-subjects factors were source-article complexity (lower- vs. higher-complexity) and article section (Part 1 vs. Part 2). The between-

Article Title	Field	Reference	Complexity
Effects of the Biomedical Bleeding Process on the Behavior of the American Horseshoe Crab, <i>Limulus polyphemus</i> , in Its Natural Habitat	Marine Biology	(Owings et al., 2019)	Lower
Sleep and Sleep Ecology in the First 3 Years: A Web-Based Study	Developmental Psychology	(Sadeh et al., 2009)	Lower
I Keep My Mind on My Money and My Money on My Mind: Trait Machiavellianism in Business Majors	Psychology	(Gruda and McCleskey, 2024)	Lower
Diseases of Rubber Trees: Malaysia as a Case Study	Plant Science	(Mahyudin et al., 2023)	Lower
Geographic Information System	Computing and Geography	(Chang, 2019)	Higher
Applications of 3D Printed Bone Tissue Engineering Scaffolds in the Stem Cell Field	Biomedical Engineering	(Su et al., 2021)	Higher
Rosemary as an Adjunctive Treatment in Patients with Major Depressive Disorder: A Randomized, Double-Blind, Placebo-Controlled Trial	Psychology and Medicine	(Azizi et al., 2022)	Higher
Searching for “the Political” in Environmental Politics	Political Science	(Kenis et al., 2014)	Higher

Table 1: Articles included in the dataset, organized by field and excerpt complexity.

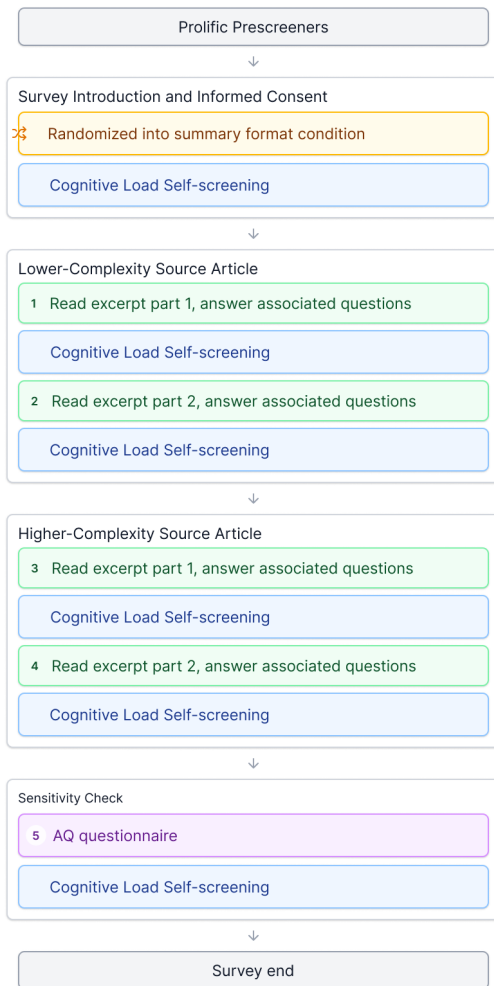


Figure 2: Participant flow through the quantitative survey. Participants completed consent and screening, were randomized to a summary-style condition, read lower- and higher-complexity article summaries in two parts, and completed repeated measures of trust, comprehension, engagement, and cognitive load. (TCE = Trust, Comprehension, Engagement).

subjects factor was summary style. Each participant was randomly assigned to one of five summary

styles: Extractive, LLM-Q&A, LLM-Abstractive, LLM-LOGOS, or LLM-PATHOS. Participants read summaries of two scientific articles, one drawn from the lower-complexity source-article condition and one drawn from the higher-complexity source-article condition, presented in random order. Figure 2 summarizes the full survey sequence.

4.2 Participants

We recruited 200 neurodivergent adults through Prolific. Eligibility required being at least 18 years old, fluent in English, and self-identifying or being clinically diagnosed with ADHD, autism, or both. Participants completed prescreening questions and the Autism Spectrum Quotient (AQ), which we used descriptively and as a covariate in sensitivity analyses. The AQ is not a diagnostic instrument, but it provides a self-report measure of autism-related traits. As in the formative study, participants were not recruited as domain experts; the target population was general neurodivergent readers who may encounter scientific information through AI-generated summaries.

4.3 Materials

Article content was drawn from eight scientific sources (Table 1). We used source-article complexity, rather than summary complexity, as one experimental factor. Four source articles were selected to represent lower-complexity scientific texts, and four to represent higher-complexity scientific texts. This distinction was based on features of the source articles, including technical density, background knowledge demands, terminology, and conceptual complexity. Each source article was then transformed into one human-generated style and four LLM-generated summary styles: Extractive, LLM-Q&A, LLM-Abstractive, LLM-LOGOS, and LLM-

PATHOS. Thus, the low/high-complexity distinction refers to the difficulty of the original scientific article, while the summary-style factor refers to how that article was rewritten for participants.

Summary construction. Extractive summaries were generated by authors of this paper by condensing main sections of source text to a specific word-count range. All LLM summaries were generated from the same source articles using GPT-4o. Generated outputs were used without substantive editing beyond generation error. Full prompts and instructions for each condition are provided in Appendix A. Examples of summary styles are included briefly in Figure 1 and more fully in Appendix B.

Extractive. A human-generated summary extracting and condensing source text from the introduction, methods, results, and conclusion.

LLM-Abstractive. A neutral LLM-generated summary that condensed information while retaining the structure of the source summary.

LLM-Q&A. A Q&A-style summary that condensed and grouped information similar to the LLM-Abstractive style.

LLM-LOGOS. A Q&A-style summary emphasizing information grouping, definitions, citations, and selected source quotations.

LLM-PATHOS. A Q&A-style summary using a professional but conversational persona designed to support engagement and comprehension.

4.4 Measures

Comprehension. After each article part, participants answered multiple-choice questions testing retention of key concepts (e.g., “What was the main objective of the study summarized in the excerpt”, with four multiple choice answers). Correct responses were summed and converted to a proportion correct score.

Cognitive load. Cognitive load was measured using NASA-TLX mental-demand items (Hart and Staveland, 1988). Ratings were collected before reading, after each reading section, and after the AQ questionnaire. The primary variable was end-of-section mental demand; baseline mental demand was included as a covariate.

Trust and engagement. Participants rated trust in the information (“I would trust this source to

inform others about the topics covered in the complete summary (both excerpts)”), perceived accuracy (“I believe the information found in the excerpt was accurate and logical”), interest (from “no interest at all” to “extremely interested”), and engagement while reading the excerpt of the summary (e.g., “The language used in the provided reading was not engaging” [reverse coded], “I remained engaged while reading the text”, and “The layout of the text contributed to keeping my attention”) using Likert-style items. Items were averaged into composite trust and engagement scores.

4.5 Analysis

We fit linear mixed-effects models with participant-level random intercepts (Figure 3). For each dependent variable—comprehension, trust, engagement, and cognitive load—we included fixed effects for summary style, article complexity, and their interaction. The cognitive-load model additionally included baseline NASA-TLX mental demand. Sensitivity analyses included AQ score as a covariate. We also computed exploratory participant-level correlations among comprehension, trust, engagement, and cognitive load. Full model coefficients, standard errors, and *p*-values are reported in Appendix D, Table 2.

5 Results

5.1 Manipulation Check

Across 200 participants, excerpts from higher-complexity articles elicited higher NASA-TLX mental-demand ratings ($M = 5.19$) than excerpts from lower-complexity articles ($M = 4.84$), $t(199) = 3.79$, $p = .0002$, $d_z = .27$ (see Figure 3). This confirms that increased source article complexity corresponded to greater perceived effort when reading summaries of those articles.

5.2 Main Effects of Summary Style

Comprehension. Comprehension depended on summary style and article complexity. For higher-complexity articles, LLM-LOGOS improved comprehension relative to the Extractive baseline ($\beta = 0.149$, $p = .005$), and LLM-Abstractive also showed a significant benefit ($\beta = 0.112$, $p = .010$) (see Figures 3, 5 & 6). LLM-PATHOS showed a non-significant trend ($\beta = 0.093$, $p = .080$), while LLM-Q&A did not differ from Extractive. For lower-complexity articles, the pattern reversed: LLM-Abstractive, LLM-LOGOS, and

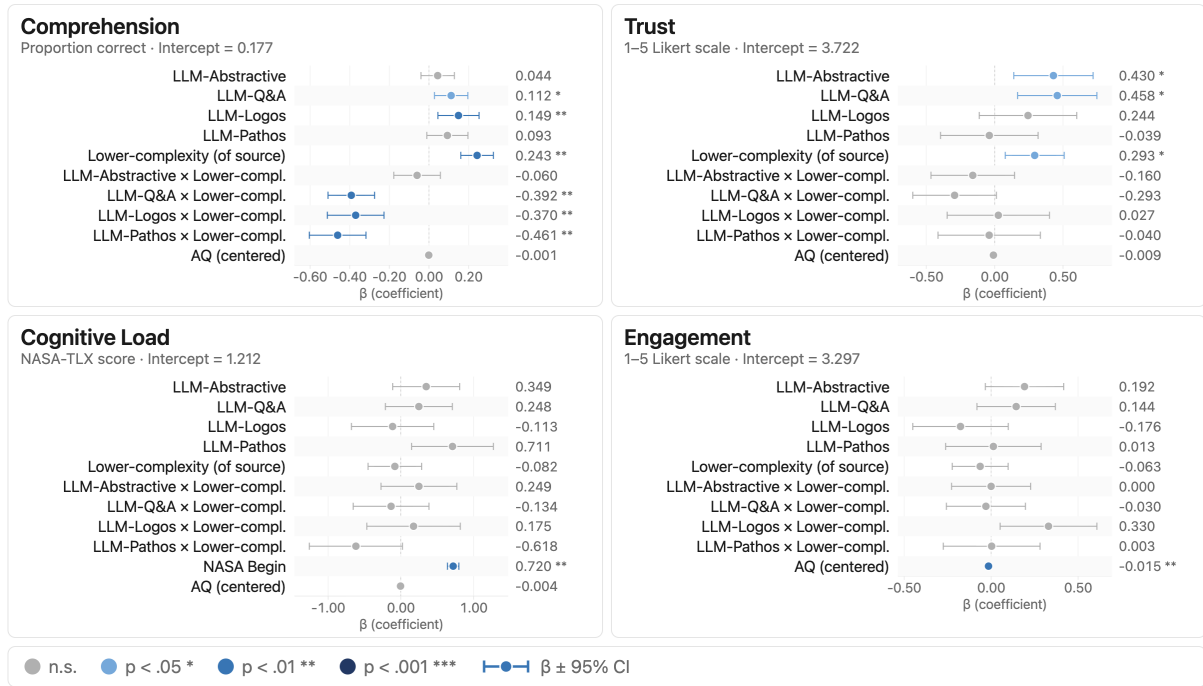


Figure 3: Mixed-effects model results across four outcome variables. Coefficients are fixed-effect estimates from models with participant-level random intercepts. Reference categories are the Extractive summary style and higher-complexity source article. Asterisks indicate significance after within-model FDR correction using the Benjamini–Hochberg procedure; because error bars are unadjusted, CI width and asterisks may not always agree. Full tabular results are in Appendix D.

LLM-PATHOS all showed lower comprehension relative to the Extractive baseline, as indicated by significant style-by-complexity interactions. This crossover suggests that added structure can help readers extract meaning from difficult material but may introduce unnecessary complexity when the source text is already accessible.

Trust. Trust showed a different pattern. LLM-Q&A and LLM-Abstractive summaries were trusted more than the Extractive baseline across conditions (see Figure 3). For example, during open-ended sections of the survey, participants who read LLM-Q&A excerpts attributed higher scores of trust to the text sounding, “like it could be true but [I] can’t be certain”, and “the chances of inaccurate information being published about the [topic] are really low”. Comparatively, participants who read the Extractive baseline excerpts stated that these texts seemed to be “missing details in the information that requires prior knowledge”, and attributed a lower trust score to them seeming, “confusing for a lay person”. LLM-LOGOS and LLM-PATHOS did not significantly differ from Extractive, and style-by-complexity interactions were not significant. This suggests that perceived credibility

may be shaped by clarity and presentation style even when comprehension benefits are conditional.

Cognitive load. Cognitive load showed limited evidence of robust style effects. End-of-section load was predicted by baseline mental demand. LLM-PATHOS showed a nominal association with higher load for higher-complexity articles, but this effect did not survive correction. Overall, perceived effort appeared to be shaped more by initial mental state and text complexity than by summary style alone.

Engagement. Engagement was relatively stable across conditions. No main effects of style or complexity survived correction. LLM-LOGOS showed a nominal engagement increase for lower-complexity articles, but without a corresponding comprehension benefit. However, this finding did not survive correction. AQ score showed a small negative association with engagement, but did not materially alter the main style findings.

5.3 Exploratory Correlations

Figure 4 summarizes the relationships among the four outcome measures. Trust and engagement were moderately positively correlated ($r = .37$).

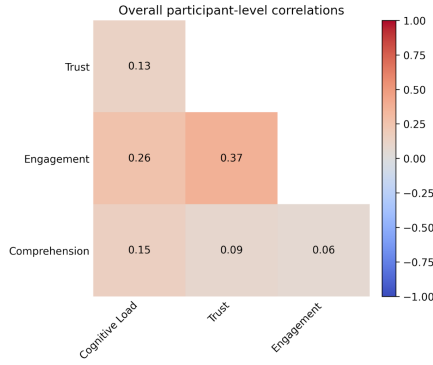


Figure 4: Pearson correlations among participant-level averages for comprehension, trust, engagement, and cognitive load.

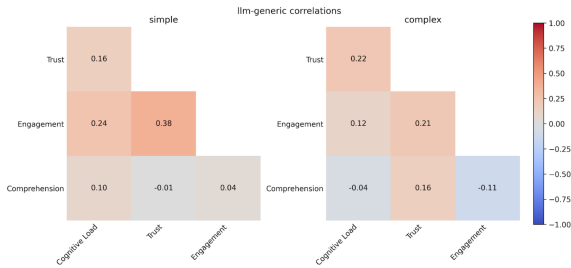


Figure 5: Pearson correlation between metrics for LLM-Generative summary style condition for lower- (left) and higher-complexity (right) articles.

Comprehension had weak correlations with trust ($r = .09$) and engagement ($r = .07$). Cognitive load showed a modest positive relation with engagement ($r = .26$), suggesting that engagement does not always correspond to reduced effort. These exploratory findings reinforce the need to measure multiple dimensions of impact. Condition-specific correlation patterns for the LLM-Generative and LLM-LOGOS summaries are presented in Figures 5 and 6, respectively.

6 Discussion

Our findings complicate a simple narrative in which LLM-generated summaries automatically make science more accessible. Although summary style did impact survey results, specific effects depended on what target area was measured (e.g., comprehension, engagement, trust, cognitive load) and the complexity of the source text.

Positive impact is multidimensional. For accessibility-oriented NLP, improvement cannot be reduced to one metric. For example, a summary style may increase trust without improving comprehension, or increase engagement without reducing



Figure 6: Pearson correlation between metrics for LLM-LOGOS summary style condition for lower- (left) and higher-complexity (right) articles.

cognitive load. In our study, comprehension benefits appeared primarily for higher-complexity articles, while trust benefits appeared for LLM-Q&A and LLM-Abstractive summaries more broadly. Engagement was comparatively stable. These differences show why evaluations of social impact should include user-centered outcomes in addition to standard NLP metrics.

Scaffolding helps when it matches the task.

The strongest relationship was the interaction between summary style and complexity of the source text. LLM-LOGOS and LLM-Abstractive summaries improved comprehension for higher-complexity scientific articles but harmed comprehension for lower-complexity ones. This suggests that scaffolding is not universally beneficial. Definitions, citations, Q&A structures, and added contextual material can help when readers face dense content, but may become distracting or burdensome when the material is already straightforward. Positive impact may therefore require adaptive support rather than maximal support.

Trust and comprehension diverged. Participants trusted LLM-Q&A and LLM-Abstractive summaries more than Extractive summaries, even though these styles did not uniformly improve comprehension. Qualitative responses suggest that trust judgments were influenced by tone, clarity, and perceived seriousness, as well as concerns about AI hallucination. For scientific communication, accessibility should support understanding without encouraging unwarranted confidence; summaries should preserve uncertainty and make source information easy to trace.

Implications for NLP system design. Our results point toward adaptive summarization systems that estimate both text difficulty and user preferences. Such systems could offer optional glossary

overlays, adjustable detail levels, collapsible explanations, annotation tools, and citations linked to source passages. Rather than generating a single summary, future systems might allow readers to move between concise, structured, and elaborated versions. This approach better respects neurodivergent readers' diverse needs and avoids treating accessibility as one-size-fits-all.

7 Conclusion

LLM-generated summaries have the potential to broaden access to scientific knowledge, but their positive impact depends on how they are designed and evaluated. In a mixed-methods study with neurodivergent adults, we found that structured LLM summaries improved comprehension for higher-complexity scientific texts but could reduce comprehension for lower-complexity texts. Trust, engagement, and cognitive load followed different patterns, underscoring that accessibility cannot be captured by a single metric. These findings support adaptive, user-controlled summarization systems that match explanatory scaffolding to text difficulty and reader needs. In the context of NLP for Positive Impact, social benefit should be measured not only by model output quality, but by whether systems help people understand, evaluate, and use information on their own terms.

Limitations

Several limitations temper our conclusions. First, participants were recruited online and diagnoses were self-reported rather than independently verified. This choice reflects real barriers to diagnosis, but limits clinical specificity. We also did not systematically screen for dyslexia or other co-occurring conditions that may affect reading, attention, or comprehension. Second, our participant characterization was limited: although we included the Autism Spectrum Quotient (AQ) descriptively and in sensitivity analyses, we did not include an equivalent ADHD-specific trait measure due to the lack thereof. A supplementary sensitivity analysis using participants' self-reported diagnosis category (ADHD-only, ASD-only, or both) found no significant associations with comprehension, trust, engagement, or cognitive load, though this relied on self-report and excluded participants reporting neither category. Third, our summary styles were generated from a limited set of prompts, representing only a small portion of the LLM summary-

design space. To this, hallucinations have the potential to exist, though comprehension questions were assessed on the generated summaries rather than their fidelity to the source article. Fourth, stipulations of fluency in the English language excluded neurodivergent adults of various linguistic backgrounds, narrowing the sampling pool as well as the descriptive lens regarding participants' experiences. Similarly, differentiating cultural attitudes towards neurodivergence also potentially affected these experiences. Fifth, comprehension was measured immediately after short excerpts, so results may not generalize to long-term retention or real-world decision-making. Sixth, the survey design did not allow participants to refer back to summaries while answering comprehension questions, which may have affected interpretation. Seventh, cognitive load was measured through self-report rather than behavioral or physiological measures such as reading time, eye tracking, or interaction logs. Finally, the qualitative study informed design but was not intended to yield generalizable claims.

Ethical and Societal Implications

This work is motivated by the positive-impact goal of making scientific information more accessible to neurodivergent readers. However, LLM-generated summaries also carry risks. If summaries are inaccurate, oversimplified, or overly persuasive, they may mislead readers or create inappropriate trust in AI-generated content. These risks are especially important for scientific and health-related information, where errors can affect personal decisions.

We therefore argue that accessible summarization systems should include source traceability, uncertainty cues, and reader control. Systems should not replace original documents or expert advice; instead, they should help readers decide what to inspect more closely. Designers should also avoid assuming that neurodivergent readers need uniformly simplified content. Our findings show that many readers want detail, but want it organized, contextualized, and adjustable.

We define positive impact in this work as improving access to scientific knowledge while preserving autonomy, accuracy, and appropriate skepticism. Measuring this impact requires evidence about how people actually use and experience NLP systems. Our study contributes one such evaluation framework by jointly measuring comprehension, trust, engagement, and cognitive load.

Positionality Statement

Members of the research team belong to the user group studied in this paper. Two team members are clinically diagnosed with and medicated for ADHD, and one self-identifies with autistic traits. The team also has familial and social relationships with neurodivergent adults. These experiences shaped our interest in the research questions and our attention to accessibility, autonomy, and participant-centered design. We also recognize that our perspectives do not represent all neurodivergent experiences.

Acknowledgements

This material is based upon work supported by the U.S. National Science Foundation through the Research Experiences for Undergraduates (REU) program under Award No. 2426230.

References

- Saeed Azizi, Neda Mohamadi, Fariba Shariffar, Gholamreza Dehghannoudeh, Farzaneh Jahanbakhsh, and Fatemeh Dabaghzadeh. 2022. [Rosemary as an adjunctive treatment in patients with major depressive disorder: A randomized, double-blind, placebo-controlled trial](#). *Complementary Therapies in Clinical Practice*, 49:101685.
- Virginia Braun and Victoria Clarke. 2021. One size fits all? what counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology*, 18(3):328–352.
- David Byrne. 2022. A worked example of braun and clarke’s approach to reflexive thematic analysis. *Quality & quantity*, 56(3):1391–1412.
- Kang-Tsung Chang. 2019. [Geographic Information System](#). In *International Encyclopedia of Geography*, pages 1–10. John Wiley & Sons, Ltd.
- Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. [A discourse-aware attention model for abstractive summarization of long documents](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.
- Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. [MS²: Multi-document summarization of medical studies](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tessa Eagle and Kathryn E. Ringland. 2023. [“you can’t possibly have adhd”: Exploring validation and tensions around diagnosis within unbounded adhd social media communities](#). In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS ’23*, New York, NY, USA. Association for Computing Machinery.
- Loretta Gasparini, Shaun Ziegenfusz, Natalie Turner, Suze Leitão, Michelle C. St Clair, and Emily Jackson. 2025. [How to create accessible research summaries for the developmental language disorder community](#). *International Journal of Language & Communication Disorders*.
- Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. [Making science simple: Corpora for the lay summarisation of scientific literature](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Dritjon Gruda and Jim A. McCleskey. 2024. [I keep my mind on my money and my money on my mind: Trait Machiavellianism in business majors](#). *Acta Psychologica*, 250:104567.
- Sydney Hadden, Kathryn Pruitt, Karen Adams, and Tyler Peterson. 2023. [How do neurodivergent people engage with tone in digital spaces? a study on the written expression of tone and intent](#). Arizona State University.
- Sandra G Hart and Lowell E Staveland. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier.
- Antonia Karamolegkou, Angana Borah, Eunjung Cho, Sagnik Ray Choudhury, Martina Galletti, Pranav Gupta, Oana Ignat, Priyanka Kargupta, Neema Kotonya, Hemank Lamba, Sun-Joo Lee, Arushi Mangla, Ishani Mondal, Fatima Zahra Moudakir, Deniz Nazar, Poli Nemkova, Dina Pisarevskaya, Naqee Rizwan, Nazanin Sabri, and 14 others. 2026. [NLP for social good: A survey and outlook of challenges, opportunities and responsible deployment](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5110–5170, Rabat, Morocco. Association for Computational Linguistics.
- Anneleen Kenis, , and Matthias Lievens. 2014. [Searching for ‘the political’ in environmental politics](#). *Environmental Politics*, 23(4):531–548.
- Ran Liu, Ming Liu, Min Yu, He Zhang, Jianguo Jiang, Gang Li, and Weiqing Huang. 2024. Sumsurvey: An abstractive dataset of scientific survey papers for long document summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9632–9651.

- Murnita Mohmad Mahyudin, Aizat Shamin Noran, Mohamad Zamir Hadi Ismail, and Khairulmazmi Ahmad. 2023. [Diseases of rubber trees: Malaysia as a case study](#). In *Forest Microbiology*, pages 401–414. Elsevier.
- Ana Miranda, Jessica Mercader, M. Inmaculada Fernández, and Carla Colomer. 2017. [Reading performance of young adults with adhd diagnosed in childhood: Relations with executive functioning](#). *Journal of Attention Disorders*, 21(4):294–304. PMID: 24149941.
- Ankan Mullick, Sombit Bose, Rounak Saha, Ayan Bhowmick, Pawan Goyal, Niloy Ganguly, Prasenjit Dey, and Ravi Kokku. 2024. [On the persona-based summarization of domain-specific documents](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14291–14307, Bangkok, Thailand. Association for Computational Linguistics.
- Meghan Owings, Christopher Chabot, and Winsor Watson. 2019. [Effects of the Biomedical Bleeding Process on the Behavior of the American Horseshoe Crab, *Limulus polyphemus*, in Its Natural Habitat](#). *The Biological Bulletin*, 236(3):207–223.
- Kaitlyn M.A. Parks, Christine N. Moreau, Kara E. Hannah, Leah Brainin, and Marc F. Joanisse. 2022. [The task matters: A scoping review on reading comprehension abilities in adhd](#). *Journal of Attention Disorders*, 26(10):1304–1324. PMID: 34961391.
- Samir Passi and Mihaela Vorvoreanu. 2022. [Overreliance on ai: Literature review](#). Technical Report MSR-TR-2022-12, Microsoft.
- Sanjana Ramprasad, Jered Mcinerney, Iain Marshall, and Byron Wallace. 2023. [Automatically summarizing evidence from clinical trials: A prototype highlighting current challenges](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 236–247, Dubrovnik, Croatia. Association for Computational Linguistics.
- Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. [Generating summaries with controllable readability levels](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.
- Francis Routledge. 2024. [Exploring the self-diagnosis experiences of selfdiagnosed autistic women](#). Rehabilitation Sciences Institute University of Toronto.
- Adrian Ryser, Florian Allwein, and Tim Schlippe. 2025. [Calibrated trust in dealing with llm hallucinations: A qualitative study](#). In *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, pages 413–421.
- Avi Sadeh, Jodi A. Mindell, Kathryn Luedtke, and Benjamin Wiegand. 2009. [Sleep and sleep ecology in the first 3 years: a web-based study](#). *Journal of Sleep Research*, 18(1):60–73.
- Jennifer C. Sarrett. 2016. [Biocertification and neurodiversity: the role and implications of self-diagnosis in autistic communities](#). *Neuroethics*.
- Xin Su, Ting Wang, and Shu Guo. 2021. [Applications of 3D printed bone tissue engineering scaffolds in the stem cell field](#). *Regenerative Therapy*, 16:63–72.
- Shahbaz Syed, Khalid Al Khatib, and Martin Potthast. 2024. [TL;DR progress: Multi-faceted literature exploration in text summarization](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 195–206, St. Julians, Malta. Association for Computational Linguistics.
- Yuuki Tachioka. 2024. [Question answer summary generation from unstructured texts by using llms](#). In *International Conference on Database Systems for Advanced Applications*, pages 261–268. Springer.
- Marco Tommasi, Simone Arnò, Chiara di Marcantonio, Laura Picconi, Maria Rita Sergi, and Aristide Saggino. 2025. [Tasting through the lens of the mind: The impact of personality and mental health on wine sensory and psychoactive effects](#). *Current Research in Food Science*, 10:101033.
- Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. [Hallucination is Inevitable: An Innate Limitation of Large Language Models](#). *arXiv preprint*. Version Number: 2.
- Victoria Yaneva, Constantin Orasan, Le An Ha, and Natalia Ponomareva. 2019. [A survey of the perceived text adaptation needs of adults with autism](#). In *Proceedings of Recent Advances in Natural Language Processing*. Research Institute in Information and Language Processing.
- Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. 2025. [A systematic survey of text summarization: From statistical methods to large language models](#). *ACM Comput. Surv.*, 57(11).

A Appendix: Summary Prompts

This section contains the instructions followed by authors to construct the extractive baseline summaries, and the prompts used, along with the entire source article, to generate the summaries used in the quantitative study.

A.1 Extractive Generation Instructions

1. Copy and paste text from the Introduction/Background, Methodology, Results, and Conclusion sections of the source text into a text editor.
2. Go section by section, cutting out sentences and words that are beyond the scope of the average reader, while retaining crucial info (i.e.,

participant count, relevant context, limiting factors, result interpretation) until the word count enters a range of 300-400 words.

3. Review the reduced text carefully to ensure that the article maintains clarity comparable to source text.
4. Iterate process if any errors are discovered.

A.2 LLM-Q&A Prompt

Summarize the attached article in a Q&A format. The summary must be split into two parts (1) Introduction and background (2) results, discussion, and conclusion. Each of the two parts should contain a minimum of 350 words.

A.3 LLM-Abstractive Prompt

Summarize the attached article. The summary must be split into two parts (1) Introduction and background (2) results, discussion, and conclusion. Each of the two parts should contain a minimum of 350 words.

A.4 LLM-PATHOS Prompt

Summarize the attached article in a Q&A style using a distinct, engaging, professional persona of your choosing. The tone should be professional, engaging, and dialogue-heavy while avoiding robotic or overly formal language. Structure the summary in two parts: Introduction and Background; Results, Discussion, and Conclusion. Each part should be 250–350 words. Include metaphors, side comments, or relatable examples where helpful, while preserving important study details. Include natural moments of skepticism or counterpoints embedded within the persona's voice.

A.5 LLM-LOGOS Prompt

Summarize the attached article in a Q&A style by breaking information into main sections: Study Overview; Key Points and Findings; Term Definitions; and Topics to Explore. The summary should be 500–750 words. Include 5–10 relevant term definitions, 2–3 quotations integrated into the explanation, necessary citations, and a short list of related articles. The goal is to make the summary comprehensive and accurate without overwhelming the reader.

B Summary Style Examples

This section contains the first paragraph of actual example summaries used in the quantitative study,

included here to demonstrate the differences in the summary styles themselves.

B.1 Extractive Example

Source Text: (Azizi et al., 2022)

Rosemary was purchased from the market and was used after scientific name confirmation. The leaves were cleaned and dried. For standardization, the plant extract was prepared by warm maceration method with 80% distilled ethanol. Rosemary oral capsules were formulated using 350 mg of the dried, milled, and sieved plant per capsule which was mixed with lactose monohydrate and starch powder. Hard gelatin capsules were filled with prepared granules using a hand operated capsule filling machine. The study was conducted as a randomized, double-blind, placebo-controlled clinical trial in Besat Clinic affiliated to Kerman University of Medical Sciences, Kerman, Iran, from April to November 2020.

B.2 LLM-Abstractive Example

Source Text: (Chang, 2019)

Geospatial data in GIS can be represented in two formats: vector (points, lines, polygons) and raster (gridded cells). Each format serves different types of information—vector is suited for discrete features like roads or land parcels, while raster works well for continuous data like temperature or elevation. GIS operations include acquiring and managing data, querying attributes, analyzing spatial relationships, and producing maps.

B.3 LLM-Q&A Example

Source Text: (Mahyudin et al., 2023)

Q: What are the recent findings on White Root Disease in Malaysia?

A: The disease can kill trees within months and affects both young and mature trees. It spreads rapidly through root contact. A 2021 genetic study revealed

high intra-population diversity and distinct geographical clustering of *R. microporus* isolates in Malaysia, indicating barriers like the South China Sea and the Titiwangsa mountain range influence gene flow.

B.4 LLM-LOGOS Example

Source Text: (Owings et al., 2019)

Q: What were the key results?

A:

- **Mating Behavior:** Bled females showed reduced spawning behavior compared to controls. For example, control females mated 4.8 times in the first week post-release, while bled females mated only 2.0 times (Owings et al., 2019, p. 211).
- **Depth:** Bled animals stayed in deeper waters during the spawning season (p. 213).

B.5 LLM-PATHOS Example

Source Text: (Su et al., 2021)

Q: Were there any noted challenges or limitations?

A: Definitely. While the potential is huge, we're not without hurdles. These include high costs, limited scalability, material toxicity, and ethical considerations surrounding stem cells and tissue printing. Furthermore, controlling degradation rates and mechanical strength remains tricky—especially for load-bearing applications.

C Formative Qualitative Interview Guide

This section includes the semi-structured interview guide used by researchers during the formative qualitative study.

C.1 Interview Part One

1. Can you explain your thought process while reading this article? For example, what sorts of struggles, concerns, or questions arose while reading?

Of these, which do you think you'll want to incorporate into your solution?

2. When you read a scientific article, what areas do you emphasize and what sections do you tend to skip or struggle with?

Do you sometimes find yourself zoning out?

(Y) why do you think this happens sometimes?

3. When you're confused about a concept—whether it's in an online, academic or casual setting—what set of steps do you take to clear your confusion?

(if confused: provide examples)

4. Have you used AI chatbots like ChatGPT, Gemini, or DeepSeek?

(Y) What kinds of situations do you usually use them for?

What situations do you avoid using them for?

C.2 Interview Part Two

C.2.1 LLM-PATHOS Section

(Introduce example part one)

If you continued having the article [the summary is] referencing explained to you in this way, would it:

1. Make the information easier or harder to understand?

Can you explain why you think this is?

2. Would you feel more or less confident in the explanation if it came in this format?

What about it makes you feel this way?

3. Do you find her way of explaining things more engaging in your understanding of the study?

(Y) What aspects of her explanation made it engaging?

(N) What aspects of the explanation style did not make it engaging?

(Introduce example part two)

If [the summary] continued to explain the counterarguments:

1. Would having access to both perspectives affect your understanding of the concept?

How?

(a) Comprehension				(b) Trust			
Predictor	β	SE	p	Predictor	β	SE	p
Intercept	0.177	0.031	< .001	Intercept	3.722	0.105	< .001
LLM-Q&A	0.044	0.043	.312	LLM-Q&A	0.430	0.148	.004
LLM-Abstractive	0.112	0.043	.010	LLM-Abstractive	0.458	0.148	.002
LLM-LOGOS	0.149	0.053	.005	LLM-LOGOS	0.244	0.182	.181
LLM-PATHOS	0.093	0.053	.080	LLM-PATHOS	-0.039	0.182	.830
Lower-complexity (of source article)	0.243	0.042	< .001	Lower-complexity (of source article)	0.293	0.110	.008
LLM-Q&A $\times$ Lower-compl.	-0.060	0.060	.316	LLM-Q&A $\times$ Lower-compl.	-0.160	0.156	.304
LLM-Abstractive $\times$ Lower-compl.	-0.392	0.060	< .001	LLM-Abstractive $\times$ Lower-compl.	-0.293	0.156	.059
LLM-LOGOS $\times$ Lower-compl.	-0.370	0.073	< .001	LLM-LOGOS $\times$ Lower-compl.	0.027	0.191	.889
LLM-PATHOS $\times$ Lower-compl.	-0.461	0.073	< .001	LLM-PATHOS $\times$ Lower-compl.	-0.040	0.191	.834
AQ	-0.001	0.001	.390	AQ	-0.009	0.006	.115

(c) Cognitive Load				(d) Engagement			
Predictor	β	SE	p	Predictor	β	SE	p
Intercept	1.212	0.264	< .001	Intercept	3.297	0.081	< .001
LLM-Q&A	0.349	0.235	.137	LLM-Q&A	0.192	0.115	.094
LLM-Abstractive	0.248	0.235	.290	LLM-Abstractive	0.144	0.115	.209
LLM-LOGOS	-0.113	0.289	.695	LLM-LOGOS	-0.176	0.140	.209
LLM-PATHOS	0.711	0.287	.013	LLM-PATHOS	0.013	0.140	.927
Lower-complexity (of source article)	-0.082	0.188	.664	Lower-complexity (of source article)	-0.063	0.082	.441
LLM-Q&A $\times$ Lower-compl.	0.249	0.266	.350	LLM-Q&A $\times$ Lower-compl.	-0.000	0.116	1.000
LLM-Abstractive $\times$ Lower-compl.	-0.134	0.266	.616	LLM-Abstractive $\times$ Lower-compl.	-0.030	0.116	.796
LLM-LOGOS $\times$ Lower-compl.	0.175	0.328	.593	LLM-LOGOS $\times$ Lower-compl.	0.330	0.142	.020
LLM-PATHOS $\times$ Lower-compl.	-0.618	0.327	.059	LLM-PATHOS $\times$ Lower-compl.	0.003	0.142	.981
NASA Begin	0.720	0.040	< .001	AQ	-0.015	0.005	.001
AQ	-0.004	0.009	.642				

Table 2: Mixed-effects model results across four outcome variables (a–d). Coefficients are fixed-effect estimates from models with participant-level random intercepts. Reference categories are the Extractive summary style and higher-complexity source article.

- Does this context make the summarized information more engaging or more overwhelming?

Can you explain why you think this is?

C.2.2 LLM-LOGOS Section

(Introduce example part one)

- What contexts would you think to use this format in?
- How would your perception change if these quotations were grabbed from other articles on the same topic?

(Introduce example part two)

- What contexts would you think to use this format in?
- Do you see yourself gravitating more to this format, or the previous one?

Why?

(Introduce example part three)

- How would you react to that kind of section?
- Is it relevant to what you're looking for in a summary?

D Mixed-effects Model Results

Table 2 presents the mixed-effects model results in a table format to more clearly show the exact standard errors and p-values.