Exploring Interactive Level of Complexity in Large Language Model Output for Information Seeking

Indu Panigrahi
indup2@illinois.edu
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Tal August
taugust@illinois.edu
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Abstract

Large language models (LLMs) are increasingly being used for information seeking, especially in knowledge-intensive domains. However, LLMs provide responses with some default amount of complexity, and it has been shown that different people need different levels of complexity based on their background and goals. Combining interactivity with the ability of LLMs to simplify text has been an effective way for users to adjust complexity. However, it is not clear how to define levels of complexity to choose from. We seek to explore this gap by firstly evaluating how LLMs generate text of varying complexity, and secondly, conducting task-based interviews to understand how different end-users leverage and perceive text complexity. In this working paper, we present preliminary findings from the model evaluation followed by an outline of our user study design. The contributions from this work will provide valuable insights for designing reading support tools, and in turn, help lower the barrier of information-seeking for people of different levels of expertise.

CCS Concepts

• **Human-centered computing** → **Natural language interfaces**; **User studies**; **Empirical studies in HCI**.

Keywords

Human-Centered AI, Interactivity, Natural Language Processing

1 Introduction

Many people have begun to use large language models (LLMs) for information seeking [25, 36, 40]. We often see this in knowledge-intensive domains, particularly in the context of research [25]. For example, LLMs have been used at the ideation and literature review stages to sift through and summarize information across documents [25, 29, 30], often using chat interfaces [32, 33, 36].

However, LLMs respond with a default amount of complexity that may not meet the varying informational needs of different users [4, 21]. It has been shown in the context of paper reading that different people need different levels of complexity based on their background and goals [4, 6, 15, 19, 20, 31]. Interactivity has been identified as an effective way for users to customize the information that they see [11, 27, 37]. Additionally, many LLM-based methods have been developed to streamline text simplification [7, 9, 18, 21]. Past work has combined interactivity and LLM-powered simplification to give users the agency to choose whether or not to simplify a given text [6, 11, 12].

However, there are no clear criteria for *defining* levels of complexity. The effectiveness of text simplification methods has been quantified with established metrics that are often indicative of complexity (e.g., length of text [14], readability [10], and human ratings [21]). While we can use these metrics to verify that an LLM follows the expected empirical trends (e.g., increasing length with increasing complexity), there are no guidelines about how models should follow the expected trends (e.g., linearly increasing length, exponentially increasing length) or how those trends affect users.

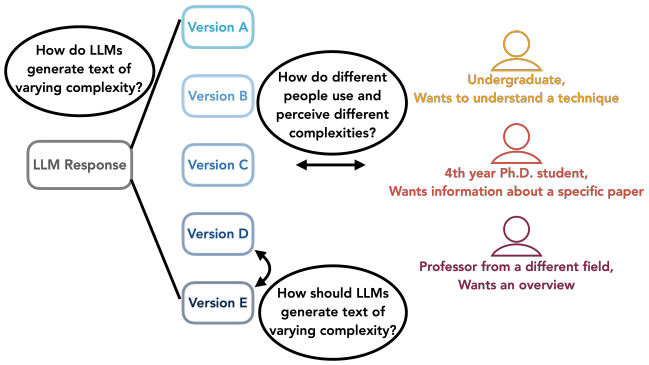

Figure 1: Overview of our research questions.

In this work, we explore this gap by considering the following research questions (Figure 1): **(1)** What trends exist in the characteristics of the different levels of complexity that LLMs generate? **(2)** How does level of complexity and its trends affect how researchers explore a topic? **(3)** How do researchers leverage and perceive interactivity when adjusting complexity? **(4)** What similarities and differences exist between how junior-level and senior-level researchers use interactive level of complexity when exploring a topic?

To address RQ1, we evaluate a variety of LLMs using metrics related to complexity; we present the preliminary takeaways from this evaluation in Section 3.2. To address the remaining RQs, we plan to conduct task-based interviews, the design of which we outline in Section 4. The contributions from these components will provide valuable insights for designing reading support tools that can assist people based on their background and goals. At a higher level, this work will enhance our understanding of how to lower information barriers for people from different levels of expertise, broadening involvement in knowledge-intensive fields.

2 Related Work

LLMs for Information Seeking. With the rising capabilities of LLMs being able to quickly produce different versions of text [4, 21, 23, 39], many LLM-powered interfaces have been developed to help people search for papers [30], skim papers [12], aggregate information across multiple documents [36, 38], and understand content [6, 11, 12, 34, 35]. A popular option has been conversation-based, question-answering systems, such as Elicit [38], Ai2's ASTA [36], and OpenAI's Deep Research [32]. Because of this, we propose a chatbot setting for the user study portion of this work. Additionally, in line with the motivation of information seeking in knowledge-intensive domains, we use STEM topics in our model evaluation and study tasks.

Adapting Text Complexity. It has been shown in the context of paper reading that different people need different levels of complexity based on their background and goals [4, 6, 15, 19, 20, 31]. There have been two primary ways to adapt complexity to users: interactivity (i.e., the *user decides* what to see) [9, 17, 27, 37] and personalization (i.e., the *system decides* what the user sees) [1, 2, 5, 22]. We focus on allowing users to interactively choose between levels of complexity rather than personalization because it is difficult to perfectly model users' preferences, and users' needs can be situational (e.g., a professor wanting a simple definition of a term in their field) [16].

However, deciding how to define different levels of complexity to choose from remains unclear. Past work has focused on evaluating whether or not LLMs could simplify text, often through prompt engineering [7, 9, 18, 21]. These works have not explored empirical trends, such as how much of a difference should exist between the simplified and unsimplified text. To address this gap, we evaluate how LLMs produce text of different levels of complexity when prompted. Additionally, to explore user behavior and preferences towards different levels of complexity, we propose conducting task-based interviews using a chatbot prototype.

3 Model Evaluation

3.1 Set-up

We begin by evaluating how a set of popular and relatively recent LLMs stratify a given input text into 5 levels of complexity. We choose to use 5 levels because it allows for some nuance between the versions of text, and a popular video series called "5 Levels" by WIRED[1] focuses on communicating specialized concepts to 5 different audiences which aligns well with the motivation for this work. We choose the models to account for different parameter counts, open and closed models, non-reasoning and reasoning abilities, and a variety of model families (Table 1).

For inputs, we provide each model with queries and their corresponding, expert-written answer reports from the ScholarQABench benchmark [3]. ScholarQABench contains STEM queries, specifically spanning Computer Science (e.g., *Could you please provide some references to work on multi-document summarization?*), Physics (e.g., *What physical quantities can be precisely measured using a levitated nanosphere?*), and Biology (e.g., *What are the biochemical analytical tools to assess the integrity and stability of LNPs?*).

[1] https://www.wired.com/video/series/5-levels/

Model	Large	Open	Reasoning
GPT-5.1	✓	✗	✗
GPT-5 mini	✗	✗	✓
Claude Sonnet 4.5	✗	✗	✗
Claude Sonnet 4.5 + Thinking	✗	✗	✓
Llama 4 Maverick	✗	✓	✓
Deepseek V3.1	✓	✓	✓

Table 1: Characteristics of LLMs included in evaluation.

For each input, we prompted each model to produce 5 versions of the input, ranging from low to high complexity. We experimented with different characteristics of prompts: specifying audience vs. generic levels, single vs. multi-prompt, and defining endpoints of complexity vs. defining all 5 complexity levels. We decided to use a single prompt that asked for 5 levels of complexity with audiences defined as College student, Junior Ph.D. student, Senior Ph.D. student, Postdoctoral researcher, and Senior researcher. Levels of writing are often labeled with different audiences, a common stratification being stage of education[4, 13, 21]. We chose these audience categories because the wording of the questions in ScholarQABench implied that the inquirer had at least a college education. We also instruct the model to only use information from the query and report provided in the input.

To quantify the complexity of text, past work employs multiple metrics together because no one metric alone captures complexity. Accordingly, we used a set of metrics inspired by past work that capture lexical, syntactic, and informational characteristics: Flesch Reading Ease Score [9, 10, 21, 42], Number of Sentences [14, 21], Average Sentence Length [5, 14, 24], Number of Information Points [4], and % Familiar Words [4, 14]. Flesch Reading Ease Score is a commonly-used scale for rating complexity that is calculated based on lexical and syntactic characteristics [10]. For Number of Information Points, we query GPT-4.1 to identify independent facts, similar to Min et al. [28]; since this metric relies on an LLM, we spot-check a few examples to verify its performance. Lastly, % Familiar Words represents the percent of words in the text that is in the Dale-Chall Word List, which contains around 3,000 familiar English words [8].

3.2 Preliminary Results

In this section, we present two preliminary takeaways from the model evaluation. Based on prior work [4, 14, 21], the intended empirical trend is that as complexity ↑, Flesch Reading Ease Score and % Familiar Words ↓, while the remaining metrics ↑. To directly examine this, we plot the *change* in each metric between consecutive versions that each model generates (Figure 2).

Firstly, models do not strictly increase complexity when prompted for different versions. In all metrics, all models can vary between increasing and decreasing complexity when prompted to generate multiple versions. In Figure 2, this can be seen when a single distribution stretches above and below 0. For example, for Δ % Familiar Words, the distributions for all models cover positive and negative values, particularly towards the later transitions. In other words, for the same transition, models can increase the complexity

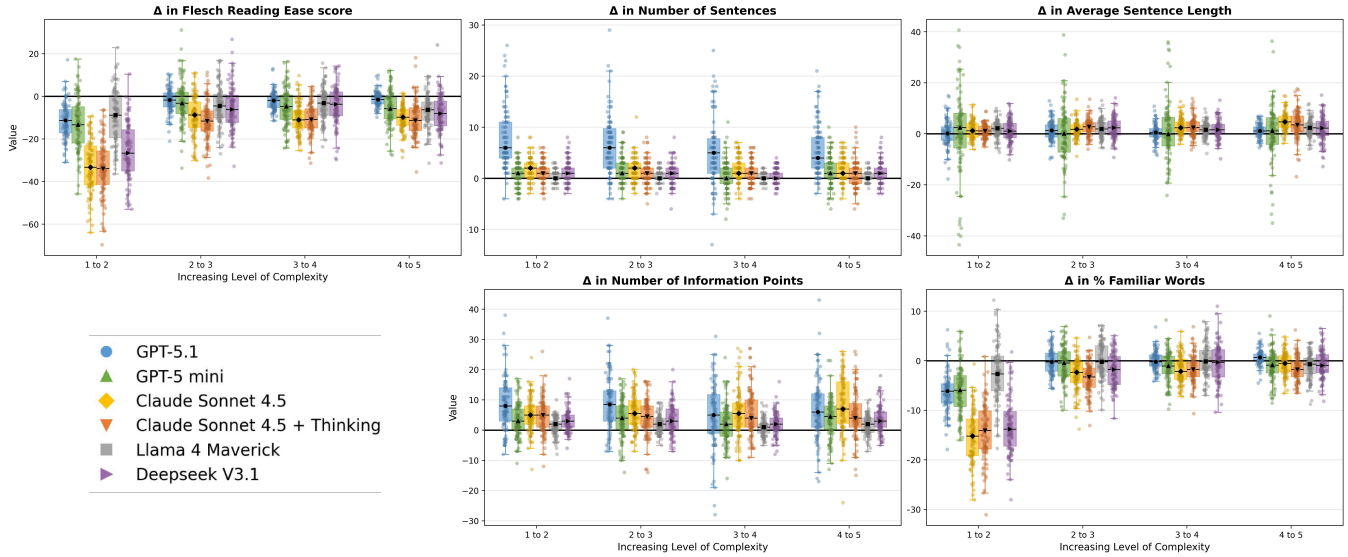

Figure 2: Model performance according to complexity metrics. The x-axes span the 4 transitions that occur (Level 1 to Level 2, Level 2 to Level 3, etc), and the y-axes represent the change in the corresponding metric that occurs for a transition. That is, each boxplot shows the distribution of changes in a metric between the consecutive versions that a particular model produces for $N = 98$ inputs. A scatter overlay is shown on each boxplot [26]; outliers were removed for visualization (Flesch Reading Ease Score: 1, Number of Sentences: 13, Average Sentence Length: 18, % Familiar Words: 1, Number of Information Points: 2).

for some inputs while decreasing the complexity for others; ideally, the models should increase the complexity for all inputs. To give a specific example, when transitioning from Level 3 to 4 in Δ Average Sentence Length, GPT-5 mini increases the average sentence length for 48% of the inputs and decreases it for 50% of the inputs.

Upon manual inspection of a few points, we qualitatively observe that the 5 versions of text can reflect the quantitative increases and decreases in the complexity metrics. An example is shown in Figure 3 for the % Familiar Words metric. Although the intended complexity is meant to increase through the 5 versions, the actual complexity of the versions according to the % Familiar Words metric alternates between increasing and decreasing. Qualitatively, the text seems to follow these fluctuations, one example in the last transition being: "*These approaches significantly improve image resolution by mitigating dispersion artifacts.*" → "*These computational strategies effectively correct dispersion-induced distortions, thereby enhancing image fidelity.*".

Secondly, models do not increase complexity by constant amounts. In all metrics, models can change the complexity by different amounts between consecutive versions. This trend has already been shown through the alternating increases and decreases described in the previous takeaway. However, even when the model monotonically increases the complexity, the amount by which the complexity increases between levels can vary. For example, in Number of Sentences, GPT-5.1 produces the changes (+13, +14, +20, +5) and (+5, +10, +1, +4) for two inputs when transitioning between the 5 levels. Sometimes, there is no change; for example, when transitioning from Level 3 to 4 in Δ Number of Information Points, Llama 4 Maverick produce no change for 8 of the 98 inputs.

Computational methods include time-frequency analysis and iterative optimization [0], a generic system dispersion compensation method [1], a **stepped detection** algorithm in the fractional Fourier domain [2], and a method using fractional Fourier transform [4].

↓ **+9.33%** (more familiar)

Computationally, techniques such as time-frequency analysis with iterative optimization [0], a generic dispersion compensation method [1], and algorithms operating in the fractional Fourier domain [2, 4] have been proposed.

↓ **-9.94%** (less familiar)

Computational techniques include iterative optimization **combined** with time-frequency analysis [0], generic system dispersion compensation [1], and **advanced signal processing** algorithms such as those operating in the fractional Fourier domain [2, 4]. **These methods enhance image clarity by correcting dispersion-induced artifacts.**

↓ **+6.65%**

Computationally, methods such as time-frequency analysis coupled with iterative optimization [0], generic dispersion compensation techniques [1], and **sophisticated** algorithms **like the stepped detection algorithm** in the fractional Fourier domain [2] and fractional Fourier transform-based methods [4] have been developed. **These approaches significantly improve image resolution by mitigating dispersion artifacts**.

↓ **-3.86%**

Computational methods **encompass a range of techniques**, including time-frequency analysis with iterative optimization [0], generic system dispersion compensation [1], and **advanced** algorithms operating within the fractional Fourier domain [2, 4]. **These computational strategies effectively correct dispersion-induced distortions, thereby enhancing image fidelity.**

Figure 3: 5 versions of a snippet of an input, increasing in complexity; Δ % Familiar Words is shown between versions. The expected trend is that % Familiar Words would decrease as the complexity increases. In this example, % Familiar Words increases by 9.33, decreases by 9.94, and so on. We bold words and phrases that differ between consecutive versions and qualitatively seem indicative of complexity.

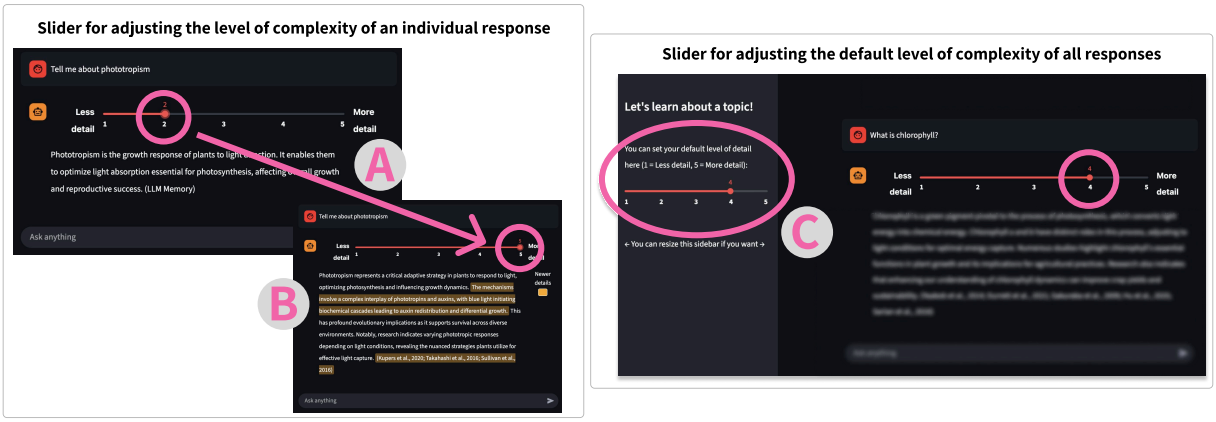

Figure 4: Features of the chatbot interface. We built this interface as a prototype that incorporates slider mechanisms for adjusting the complexity of individual responses (A) and the default complexity of all responses (C). When the user moves the per-response slider (A) to a different notch, sentences that are significantly different from those in the previously-displayed version are highlighted (B); significant differences are determined by comparing sentence-level BERTScores [41] to a threshold.

4 Next Steps: Human-Centered Evaluation

While the model evaluation displays the empirical relationships between the levels of complexity that LLMs produce, it does not address how we *should* define complexity levels (e.g., Should models should increase complexity by constant amounts? Do users notice if models alternate between increasing and decreasing complexity?). Thus, we plan to conduct within-subjects, task-based interviews, conditioned based on the trends we found in the model evaluation.

We built a chatbot prototype[2] that incorporates slider mechanisms for adjusting the complexity of model responses to user queries (Figure 4). We plan to recruit approximately 24 junior- and senior-level STEM researchers in and outside of academia. Using the interface, participants will complete a few information-seeking tasks (e.g., "Find 5 papers related to geotropism.", "What is the main idea of each paper you chose?") under 3 conditions from the model evaluation that dictate the scale for the sliders:

(1) Response increasing in complexity by a constant amount.
(2) Response increasing in complexity by variable amounts.
(3) Response increasing and decreasing in complexity.

That is, when the participant enters a query, the model will be prompted to respond with 5 levels of complexity that fulfill one of these conditions to "populate" the slider. An open methodological question is deciding which metrics should be used to enforce the conditions. Since we use multiple metrics to gauge complexity, our current approach is to provide all the metric values as feedback to the model and have it regenerate its response until the response fulfills the condition. However, this may result in longer response times during the user study; thus, we believe it may be beneficial to select one or two metrics to use as feedback to the model.

Lastly, we will evaluate a mix of quantitative and qualitative data. After each condition in the study, we will collect subjective ratings about participants' perceptions of the text versions and interactivity. Additionally, participants will be asked to think-aloud during the study and to write the information they found for the tasks into a document. Their interactions with the interface (e.g., queries, slider adjustments) will be tracked as well. We will conclude each study with an exit interview in which we will ask about the participant's experience and any interesting interactions or comments.

[2]Built using https://streamlit.io/

References

[1] Sabita Acharya, Barbara Di Eugenio, Andrew Boyd, Richard Cameron, Karen Dunn Lopez, Pamela Martyn-Nemeth, Carolyn Dickens, and Amer Ardati. 2018. Towards Generating Personalized Hospitalization Summaries. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, Silvio Ricardo Cordeiro, Shereen Oraby, Umashanthi Pavalanathan, and Kyeongmin Rim (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, USA, 74–82. doi:10.18653/v1/N18-4011

[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. doi:10.1145/3025453.3025631

[3] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, D'arcy Mike, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Dan Weld, Graham Neubig, Doug Downey, Wen-tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. 2024. OpenScholar: Synthesizing Scientific Literature with Retrieval-Augmented Language Models. *arXiv* (2024). https://arxiv.org/abs/2411.14199

[4] 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 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 14, 26 pages. doi:10.1145/3613904.3642289

[5] Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating Scientific Definitions with Controllable Complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 8298–8317. doi:10.18653/v1/2022.acl-long.569

[6] 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 (Sept. 2023), 38 pages. doi:10.1145/3589955

[7] Nadine Beks van Raaij, Daan Kolkman, and Ksenia Podoynitsyna. 2024. Clearer Governmental Communication: Text Simplification with ChatGPT Evaluated by Quantitative and Qualitative Research. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, Giorgio Maria Di Nunzio, Federica Vezzani, Liana Ermakova, Hosein

Azarbonyad, and Jaap Kamps (Eds.). ELRA and ICCL, Torino, Italia, 152–178. https://aclanthology.org/2024.determit-1.15/

[8] Jeanne Sternlicht Chall and Edgar Dale. 1995. Readability revisited : the new Dale-Chall readability formula. https://api.semanticscholar.org/CorpusID:61078711

[9] Michael Färber, Parisa Aghdam, Kyuri Im, Mario Tawfelis, and Hardik Ghoshal. 2025. SimplifyMyText: An LLM-Based System for Inclusive Plain Language Text Simplification. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part IV* (Lucca, Italy). Springer-Verlag, Berlin, Heidelberg, 418–424.

[10] Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology* 32, 3 (1948), 221–233. doi:10.1037/h0057532

[11] Raymond Fok, Joseph Chee Chang, Tal August, Amy X. Zhang, and Daniel S. Weld. 2024. Qlarify: Recursively Expandable Abstracts for Dynamic Information Retrieval over Scientific Papers. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 145, 21 pages. doi:10.1145/3654777.3676397

[12] Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent Skimming Support for Scientific Papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 476–490. doi:10.1145/3581641.3584034

[13] Science Journal for Kids. [n. d.]. Science Journal for Kids and Teens. https://www.sciencejournalforkids.org/

[14] Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. APPLS: Evaluating Evaluation Metrics for Plain Language Summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9194–9211. doi:10.18653/v1/2024.emnlp-main.519

[15] Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated Lay Language Summarization of Biomedical Scientific Reviews. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 1 (May 2021), 160–168. doi:10.1609/aaai.v35i1.16089

[16] Jerry Zhi-Yang He, Sashrika Pandey, Mariah L Schrum, and Anca Dragan. 2025. Context Steering: Controllable Personalization at Inference Time. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=xQCXInDq0m

[17] 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. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 413, 18 pages. doi:10.1145/3411764.3445648

[18] Elias Hedlin, Ludwig Estling, Jacqueline Wong, Carrie Demmans Epp, and Olga Viberg. 2025. Got It! Prompting Readability Using ChatGPT to Enhance Academic Texts for Diverse Learning Needs. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*. Association for Computing Machinery, New York, NY, USA, 115–125. doi:10.1145/3706468.3706483

[19] Terje Hillesund. 2010. Digital reading spaces: How expert readers handle books, the Web and electronic paper. *First Monday* 15, 4 (April 2010). doi:10.5210/fm.v15i4.2762

[20] Fred Hohman, Matthew Conlen, Jeffrey Heer, and Duen Horng Chau. 2020. Communicating with Interactive Articles. *Distill* (2020). doi:10.23915/distill.00028

[21] Brihi Joshi, Keyu He, Sahana Ramnath, Sadra Sabouri, Kaitlyn Zhou, Souti Chattopadhyay, Swabha Swayamdipta, and Xiang Ren. 2025. ELI-Why: Evaluating the Pedagogical Utility of Language Model Explanations. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 25466–25499. doi:10.18653/v1/2025.findings-acl.1306

[22] Taewook Kim, Dhruv Agarwal, Jordan Ackerman, and Manaswi Saha. 2025. Steering AI-driven Personalization of Scientific Text for General Audiences. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW479 (Oct. 2025), 28 pages. doi:10.1145/3757660

[23] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence* 6, 4 (April 2024), 383–392.

[24] Junyi Li and Ani Nenkova. 2015. Fast and Accurate Prediction of Sentence Specificity. *Proceedings of the AAAI Conference on Artificial Intelligence* 29, 1 (Feb. 2015). doi:10.1609/aaai.v29i1.9517

[25] Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X Zhang. 2025. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. In *Second Conference on Language Modeling*. https://openreview.net/forum?id=p0BwJk3R1p

[26] Justin Matejka and George Fitzmaurice. 2017. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. 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, 1290–1294. doi:10.1145/3025453.3025912

[27] Bryan Min, Allen Chen, Yining Cao, and Haijun Xia. 2025. Malleable Overview-Detail Interfaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 688, 25 pages. doi:10.1145/3706598.3714164

[28] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12076–12100. doi:10.18653/v1/2023.emnlp-main.741

[29] Meredith Ringel Morris. 2023. Scientists' Perspectives on the Potential for Generative AI in their Fields. arXiv:2304.01420 [cs.CY] https://arxiv.org/abs/2304.01420

[30] Alexandra Mudd, Tiffany Conroy, Siri Lygum Voldbjerg, Anita Goldschmied, Rebecca Feo, and Lambert Schuwirth. 2025. Developing and Evaluating the Use of ChatGPT as a Screening Tool for Nurses Conducting Structured Literature Reviews: Proof of Concept Study Results. *Journal of Clinical Nursing* (2025), 1–13. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jocn.17818 doi:10.1111/jocn.17818

[31] David Nicholas, Peter Williams, Ian Rowlands, and Hamid R. Jamali. 2010. Researchers' e-journal use and information seeking behaviour. *J. Inf. Sci.* 36, 4 (Aug. 2010), 494–516. doi:10.1177/0165551510371883

[32] OpenAI. 2025. ChatGPT with Deep Research. https://chat.openai.com/

[33] Perplexity AI. 2025. Sonar Deep Research. https://docs.perplexity.ai/getting-started/models/models/sonar-deep-research

[34] Paul Rust, Julian Frings, Sven Meister, and Leonard Fehring. 2025. Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations. *Communications Medicine* 5, 1 (29 May 2025), 208. doi:10.1038/s43856-025-00927-2

[35] Paul Rust, Julian Frings, Sven Meister, and Leonard Fehring. 2025. Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations. *Communications Medicine* 5, 1 (May 2025), 208. doi:10.1038/s43856-025-00927-2

[36] Amanpreet Singh, Joseph Chee Chang, Dany Haddad, Aakanksha Naik, Jena D. Hwang, Rodney Kinney, Daniel S Weld, Doug Downey, and Sergey Feldman. 2025. Ai2 Scholar QA: Organized Literature Synthesis with Attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Pushkar Mishra, Smaranda Muresan, and Tao Yu (Eds.). Association for Computational Linguistics, Vienna, Austria, 513–523. doi:10.18653/v1/2025.acl-demo.49

[37] S. Shyam Sundar, Qian Xu, and Saraswathi Bellur. 2010. Designing interactivity in media interfaces: a communications perspective. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2247–2256. doi:10.1145/1753326.1753666

[38] Sharon Whitfield and Melissa A. Hofmann. 2023. Elicit: AI literature review research assistant. *Public Services Quarterly* 19, 3 (2023), 201–207. arXiv:https://doi.org/10.1080/15228959.2023.2224125 doi:10.1080/15228959.2023.2224125

[39] 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: 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12–15, 2023, Proceedings, Part I* (Foshan, China). Springer-Verlag, Berlin, Heidelberg, 695–707. doi:10.1007/978-3-031-44693-1_54

[40] Ali Zaidi and Karrie Karahalios. 2025. From Sociotechnical Gaps to Solutions: Designing AI Tools with Parents to Address Special Education Advocacy Barriers in IEP Processes. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 2619–2636. doi:10.1145/3715336.3735778

[41] 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

[42] Dagný Halla Ágústsdóttir, Jacob Rosenberg, and Jason Joe Baker. 2025. ChatGPT-4o Compared With Human Researchers in Writing Plain-Language Summaries for Cochrane Reviews: A Blinded, Randomized Non-Inferiority Controlled Trial. *Cochrane Evidence Synthesis and Methods* 3, 4 (Jul 2025). doi:10.1002/cesm.70037