Explain Like I'm 5 or Whatever I Choose: Evaluating the Interactive Potential of Language Model Responses

Anonymous ACL submission

Abstract

Evaluations of large language models (LLMs) in scientific information seeking tasks have become increasingly use-centric, such as conducting live or multi-turn evaluations with real users. These evaluations still assume a single, static chat interface, but as models are integrated into new interfaces, evaluations must shift to incorporate interface-specific criteria. We propose a new evaluation framework based on a formative study with 16 participants that tests models' ability to generate multiple responses to one query that differ along an interpretable axis of language (language complexity), inspired by direct manipulation interfaces from human-centered design literature. We evaluate GPT-5.1, GPT-5 mini, Claude Sonnet 4.5 + Thinking, and DeepSeek-V3.1 by generating 5 responses at different levels of language complexity for 98 scientific queries. While models vary complexity across responses, most changes remain inconsistent, with the best performing model (Claude Sonnet 4.5) only shifting reliable complexity measures in the correct direction 46% of the time. Our findings hold with increased sample size and alternative complexity levels.

1 Introduction

Current evaluations of large language models (LLMs) in information seeking tasks have become increasingly use-centric in response to the rapid improvement of models and their integration into deployed systems. For example, more evaluations focus on conducting live or multi-turn evaluations with users (Bragg et al., 2026b). However, these evaluations generally assume a single, static user interface (i.e., a chat interface). As models are integrated into new interfaces (e.g., reading or writing interfaces, Joshi and Vogel, 2026; Lee et al., 2024), evaluations must shift to incorporate interface-specific criteria.

We focus on one such task and use context: scientific information seeking with mixed-expertise users. Scientists increasingly use LLMs for reading (Fok et al., 2023; Lo et al., 2023) and synthesizing literature (Asai et al., 2026; OpenAI, 2025). Readers of scientific language can vary in their preferred responses depending on their background (e.g., a junior or senior researcher) and particular context (e.g., reading within a familiar or unfamiliar discipline), such as preferring simpler or more complex summaries and explanations (Guo et al., 2021; August et al., 2022; Joshi et al., 2025).

Past evaluations have focused on models' ability to generate responses that align with different envisioned audiences (Joshi et al., 2025), or personalize a response to given user (Guo et al., 2024b; Murthy et al., 2022). However, a single response can be misaligned with a user, fail to incorporate correct context (Fok et al., 2024), or be inherently insufficient even if correct. For example, the design paradigm of *details on demand* (Min et al., 2025) suggests that users may—after seeing an initial summary—want more details, even if they did not want a detailed first response. While users might prompt a model for more details, often direct manipulation of text (e.g., a slider for text complexity) can be a more effective mechanism for end-user control (Zhang et al., 2026).

Rather than asking if models can generate a single best scientific summary, in this paper we ask if models can generate *multiple* summaries that enable effective user selection and control. We focus on language complexity (e.g., jargon, information, Guo et al., 2024a; Joshi et al., 2025) in model responses, and validate our approach with a formative user study ($N = 16$) where participants explored unfamiliar STEM topics using a prototype chat interface that enabled direct manipulation of response language complexity (Fig. 1). Using an existing scientific QA benchmark (Asai et al., 2026), we test 5 recent models (GPT-5.1,

GPT-5 mini, Claude Sonnet 4.5 + Thinking, and DeepSeek-V3.1) by generating 5 responses at different levels of response complexity[1] to 98 scientific queries. We define model performance based on the relationship *between* versions of a response.

We find that while models are able to vary complexity across responses, most changes remain inconsistent. For example, the best-performing model for jargon change (Claude Sonnet 4.5) only changed jargon in the correct direction 46% of the time across the 5 responses for an individual query (33% of the time for our measure of information, Sec. 4.3). Additionally, though models consistently increased complexity measures for lower complexity responses, in higher levels, changes in measures neared chance level for most models. We also show that our findings hold with increased sample size ($N = 459$) and alternative audience labels. In summary, our paper makes the following contributions:

1. A new evaluation framework that tests models' ability to generate multiple, distinct versions of a response across a dimension of interest. We instantiate the framework for scientific information seeking.

2. A suite of three complexity measures motivated from prior work and grounded to a user study with 16 participants reading scientific literature in unfamiliar disciplines. To evaluate models, we define a criterion for the relationship of complexity measures *between* versions of a response.

3. Evaluation results from 5 models on scientific queries showing that models often fail to reliably adjust complexity measures in the correct direction across versions. Our findings hold across models and when we shift anchors for version generation (i.e., make anchors more distant from one another).

2 Related Work

2.1 LLMs for Information-Seeking

With the rising capabilities of LLMs being able to quickly produce different versions of text (Kirk et al., 2024; August et al., 2024; Joshi et al., 2025; Wu et al., 2023a), many LLM-powered interfaces have been developed to help people search for papers (Mudd et al., 2025), skim papers (Fok et al., 2023), aggregate information across multiple documents (Singh et al., 2025; Whitfield and Hofmann, 2023), and understand content (August et al., 2023; Rust et al., 2025a; Fok et al., 2023; Rust et al., 2025b; Fok et al., 2024). A popular option has been conversation-based, question-answering systems, such as Elicit (Whitfield and Hofmann, 2023), ASTA (Singh et al., 2025), and OpenAI's Deep Research (OpenAI, 2025).

2.2 Adapting LLM Responses to Users

There have been two primary ways to adapt LLM responses to users: interactivity (i.e., the *user decides* what to see) (Min et al., 2025; Sundar et al., 2010; Färber et al., 2025; Head et al., 2021) and personalization (i.e., the *system decides* what the user sees) (Kim et al., 2025b; Adar et al., 2017; August et al., 2022; Acharya et al., 2018). Past evaluations in complexity adaption have had an implicit *personalization* context, meaning the goal was for the model to generate a single correct response, evaluating whether or not LLMs could simplify text by generating a response at a 5th grade reading level when prompted with "explain like I am a 5th grader" for example (Beks van Raaij et al., 2024; Hedlin et al., 2025; Joshi et al., 2025; Färber et al., 2025). We instead focus on an *interactive* context by evaluating how well models generate a range of responses, allowing users to choose between levels of complexity. We explore this alternative framing because users may need to select their own version depending on their needs in the moment (Fok et al., 2024) rather than on a static attribute (e.g., a professor wanting a simple definition of a term in their field). While users can prompt models for this information, specifying information needs iteratively can be time consuming and distracting (Zamfirescu-Pereira et al., 2023).

3 Formative User Study

Because our motivation for analyzing language complexity is linked to an interactive context (i.e., users directly manipulating language), we start by conducting a formative study to structure our subsequent model evaluation (Sec. 4). Specifically, we aim to (i) validate the utility of interactive complexity, (ii) corroborate characteristics of language that participants associate with complexity, and (iii) identify criteria for model responses that enable successful interactive complexity. In this

[1] We anchor complexity levels in our prompt by using different envisioned audiences, similar to past work (Joshi et al., 2025; August et al., 2022), see Sec 3.1.

section, we describe our user study design (Sec. 3.1) and findings (Sec. 3.2).

3.1 Study Procedure

We conducted a within-subjects study with 16 participants primarily from research and STEM backgrounds. To test the utility of direct manipulation of response complexity, the study was counterbalanced between an interactive chat condition where participants used interactive complexity (described below) and a conventional chat condition. To emulate the experience of exploring new topics with the interfaces, we asked each participant to provide two topics that they were interested in but had little to no knowledge about before the study. In line with the motivation of information seeking in knowledge-intensive domains, we required that the topics were in STEM. During the study, participants had 15 minutes to interact with each interface and complete two simple information-seeking tasks; details on instructions and tasks are provided in Appendix A.9. While interacting with each interface, participants were asked to actively verbalize their thought process, reasoning, and impressions. After each condition, we asked a few questions about participants' experience with and perceptions of the interface and its chat responses; the interview guide is provided in Appendix A.8. This study was approved by our institution's IRB.

Conditions The interactive complexity condition provided users with a slider mechanism to adjust the language complexity of chat responses, choosing from 5 levels labeled 1 through 5, 1 being the least complex (Fig. 1). The conventional interface looked and functioned similarly, without the sliders. All responses were generated using GPT-5 mini, chosen as a recent model with low latency appropriate for an interactive context.

Response Generation We used topics provided by participants to pregenerate scientific reports using a recent RAG-based pipeline (ScholarQA, Singh et al., 2025) with Claude Sonnet 4.5. Using these reports as a single ground-truth document, we prompted GPT-5 mini to generate the slider responses with audiences defined as College student, Junior Ph.D. student, Senior Ph.D. student, Postdoctoral researcher, and Senior researcher. We based these levels off of previous work that has stratified model responses based on level of education (Joshi et al., 2025; Science Journal for Kids; August et al., 2024), adjusting the levels to be appropriate for the envisioned prompts (i.e., scientific literature queries). We use a single prompt to generate all 5 levels. Appendix A.1 includes the prompt and describes alternative prompts we tested.

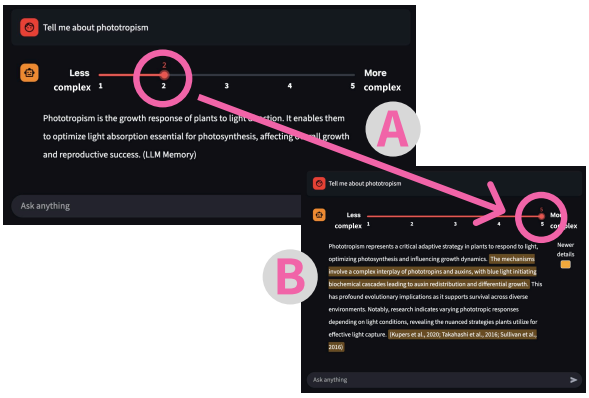


Figure 1: **Interactive language complexity interface** Users can manipulate textual complexity by moving the response slider (A) to a different notch. Sentences that are significantly different from those in the previously-displayed version are highlighted (B); significant differences are determined by comparing sentence-level BERTScores (Zhang et al., 2020) to a threshold. The preset default is 3, but there is also the option to change the default to apply to all sliders.

Participants We recruited 16 participants, primarily within academic institutions. Due to our snowball sampling process, most participants held an academic affiliation and had a STEM background. More participant details are provided in Appendix A.6. Studies were conducted over Zoom video calls and lasted 1 hour after which all participants received a $25 gift card, which is above the minimum wage policy in the area.

Evaluation To analyze the qualitative data from the study transcripts, we employ open coding (Saldaña, 2021) where one author created an initial codebook from 4 studies randomly sampled evenly between conditions, which all authors then iterated on until a final codebook was agreed upon. Using the final codebook, the same initial author coded the remaining studies. Since all authors were on consultation for study coding, we did not calculate inter-rater reliability (McDonald et al., 2019). The codes describe participants' sentiments (e.g., "*Appreciates agency of interactive chat*") and interactions (e.g., "*Prompted with desired complexity*").

3.2 Results

Interactive complexity is valuable over a conventional chat interface The majority of partic-

ipants (13/16) appreciated the flexibility that interactive complexity provided. Participants found the responses from the conventional chat interface inconsistent in terms of complexity, sometimes providing too much or too little information or jargon (13/16). In fact, 7 participants (4 of whom had not yet seen the interactive condition) ended up describing their preferred level of complexity in follow-up prompts, indicating a strong desire for control over perceived complexity of responses. As P16 describes: "*with the [conventional] chatbot, it felt like there was a misalignment between how I was interpreting my level of understanding and how [the model] interpreted it. So... I would manually adjust the prompt... The complexity slider was nice because it was just a very quick and easy way*".

Jargon, information, and length influenced perceived complexity As participants decreased complexity, they prioritized three primary, desired trends: decreases in jargon (16/16), amount of information (13/16), and length (12/16). This finding reinforces expectations that prior work has assumed (August et al., 2024; Joshi et al., 2025; Guo et al., 2024a) and forefronts the measures that we focus on in our model evaluation (Sec. 4).

Small changes in complexity are hard to perceive While 5 levels of complexity seemed generally appropriate (10/16), 10 participants noted the levels needed to be more distinct from each other. In particular, 6 participants noted that "*one and two and four and five [didn't] feel all that much different*" (P9). This observation suggests that some responses did not strictly increase in complexity, conflicting with what users expect and need.

4 Evaluating Interactive Complexity

Motivated by our formative findings, we test the potential for models to provide responses that enable interactive complexity control. Specifically, we evaluate model performance based on the relationship between multiple responses, rather than on a single response. Below we describe the data (Sec. 4.1) and models (Sec. 4.2) used in our evaluation, as well as our measures of complexity (Sec. 4.3) and our criterion to facilitate effective user control of model responses (Sec. 4.4).

4.1 Scientific Query Data

We use the queries from ScholarQA-Multi, a subset of ScholarQABench (Asai et al., 2026) that contains 98 expert-written scientific queries and answers across multiple STEM fields (Tab. 1). We restricted to this subset to use the expert-written answer reports for grounding model responses to a single ground truth report (similar to our user study), aligning with past work on the utility of human-written context (Tan et al., 2024; Zhang et al., 2023) and potential issues with reusing models for both initial ground truth generation and subsequent response generation (Xu et al., 2024; Panickssery et al., 2024). Since this restriction leads to a relatively small sample, we also evaluate a larger sample of 459 queries with reports generated from ScholarQA (Singh et al., 2025) as the RAG pipeline using Claude Sonnet 4.5 (Sec. 5.4).

4.2 Models

We prompt models the same way as for the formative user study (Sec. 3.1) —to generate 5 versions of the response, stratified for 5 envisioned audiences going from least to most complex: a College student, Junior Ph.D. student, Senior Ph.D. student, Postdoctoral researcher, and Senior researcher. We explore alternative anchors in Sec. 5.5. We evaluate 5 recent models across different sizes, model families, and reasoning abilities: GPT-5.1, GPT-5 mini, Claude Sonnet 4.5, Claude Sonnet 4.5 + Thinking, and DeepSeek-V3.1. Details about model configurations are in Appendix A.3.

Discipline	% of Data	Sample Query
Biology	20.4	*What are the biochemical analytical tools to assess the integrity and stability of LNPs?*
Biophysics	9.2	*Find some papers that discuss the methods to suppress multiple light scattering effect.*
Computer Science	33.7	*Could you please provide some references to work on multi-document summarization?*
Photonics	30.6	*What progress has been made in trapping and controlling multiple nanoparticles?*
Physics	6.1	*What are the ways to perform optomechanical cooling?*

Table 1: **Distribution of data across disciplines** ScholarQA-Multi consists of 98 queries distributed across Biology, Biophysics, Computer Science, Photonics, and Physics; a sample question from each domain is shown.

4.3 Complexity Measures

Past work has quantified complexity through several linguistic measures. These include readability formulas (e.g., Flesch-Kincaid score, Flesch, 1948), lexical features, and perplexity measures (Guo et al., 2024a; August et al., 2024; Joshi et al., 2025). We take inspiration from these works and our formative study results (Sec. 3.2) to curate a suite of three measures, each representing different dimensions of perceived language complexity:

- JARGON (Guo et al., 2024a; Joshi et al., 2025): This measure denotes the proportion of text that consists of less familiar words. We calculate the percentage of words in the text that is not on the Dale-Chall Word List, a list of 3,000 familiar English words (Chall and Dale, 1995).
- INFORMATION (Trienes et al., 2024; Guo et al., 2024a): More complex language often includes more information (e.g., technical details). We query GPT-4.1 to identify independent facts, using the pipeline for generating "atomic facts" from Min et al. (2023); we validate model performance by manually inspecting a subset of 25 examples.
- LENGTH (Joshi et al., 2025; Guo et al., 2024a): More complex language is often longer, though simpler language can also be longer when the same information elaborated upon (Wu et al., 2023b). We quantify length by total number of response characters.

The Flesch-Kincaid Reading Ease Score is a commonly-used scale for rating complexity (Flesch, 1948; Joshi et al., 2025; Ágústsdóttir et al., 2025; Färber et al., 2025). However, past work has shown that Flesch-Kincaid scores do not provide a reliable measure of complexity (Cachola et al., 2025; Tanprasert and Kauchak, 2021; Imperial and Tayyar Madabushi, 2023), and we found that our findings remain the same as the scores exhibit similar trends to JARGON and INFORMATION. Thus, we focus on the measures that we confirmed in our user study and provide the Flesch-Kincaid data in Appendix A.4.

4.4 Criterion for effective user control

Based on prior work (Guo et al., 2024a; Joshi et al., 2025; August et al., 2024) and our formative user study, we define model performance based on a model's ability to increase the measures as the intended complexity levels increase. We operationalize this by evaluating the *direction of changes* in the measures between the 5 levels of text that models generate for each query. **Models that are better at generating distinct levels of complexity will produce positive changes in all measures.** Negative changes indicate that the model decreased complexity when it should have increased.

5 Results

Fig. 2 plots the change in each measure between consecutive levels of intended complexity that each model generates. To quantify model performance, we report the percent of changes that move in the correct direction (i.e., increasing JARGON, INFORMATION, or LENGTH) in Tab. 3. Because Fig. 2 and Tab. 3 display results separated by each transition between levels, we also report the overall performance across all 5 response levels in Tab. 2 (i.e., the percent of inputs for which the generated 5 levels move in the correct direction across all transitions). We confirm in Appendix A.5 that the distribution of the responses we report here reflect the response distribution from our user study.

5.1 Models inconsistently increase complexity measures

All models vary between increasing and decreasing complexity measures when generating multiple levels. In Fig. 2, this can be seen where the distributions cover positive and negative changes for the same model and transition between levels, particularly in JARGON and INFORMATION. For example, when transitioning from Level 4 to 5, Claude Sonnet 4.5 increases JARGON for 55.10% of the inputs (Tab. 3). In other words, for the same transition, models can increase the complexity for some inputs while decreasing the complexity for others. Fig. 4 shows an example where JARGON and INFORMATION both decrease when the complexity is supposed to increase. The proportion of stagnant changes are at most 1.02% for JARGON, 7.14% for INFORMATION, and 0% for LENGTH; thus we focus the proportion of changes that decrease as a stronger indication of complexity going in the wrong direction.

5.2 Models increase length with complexity

Unlike JARGON and INFORMATION, LENGTH generally increases with increasing complexity, as is the intended trend. This can be seen in Tab. 2 as

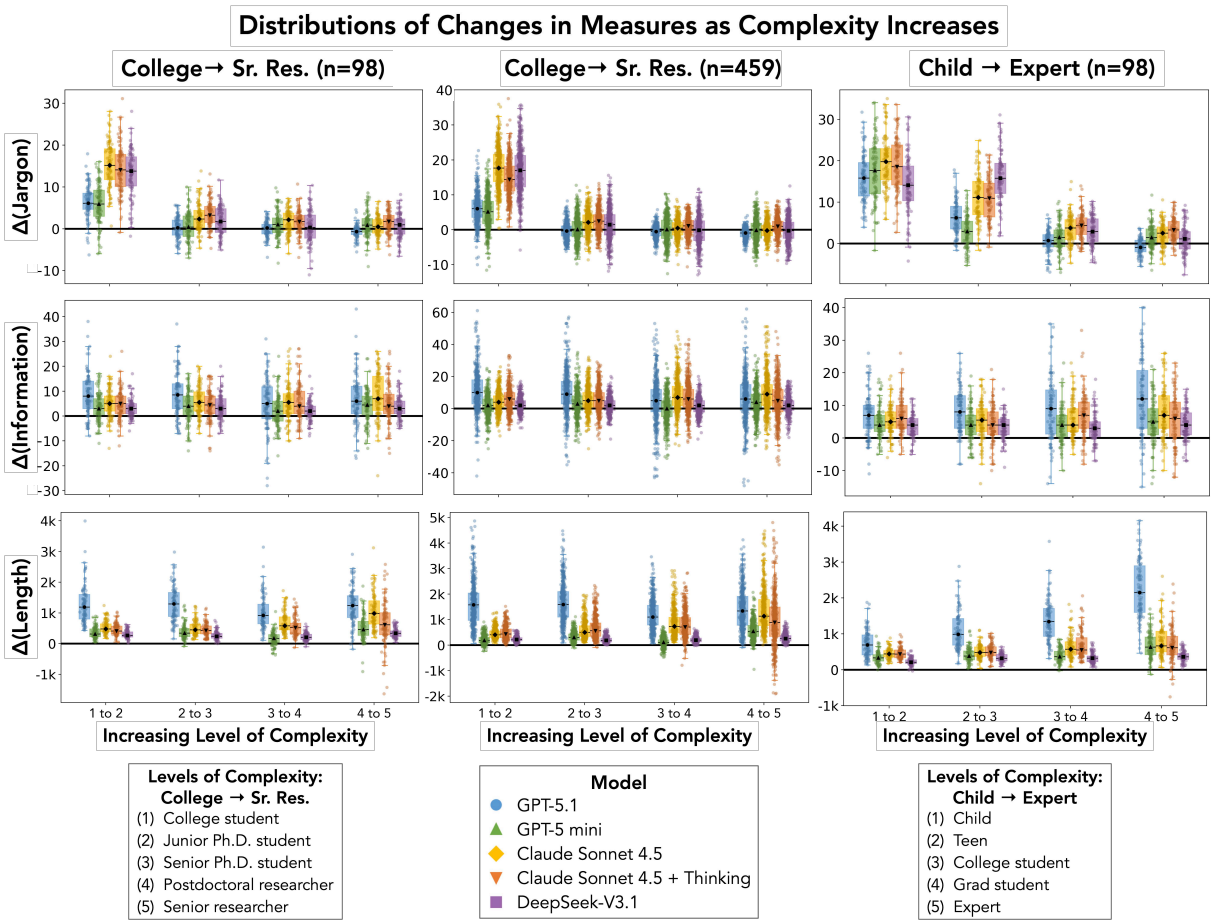


Figure 2: **Model performance shown as changes in complexity measures** Between consecutive levels of complexity, models produce changes in JARGON, INFORMATION, and LENGTH that vary between increasing and decreasing, particularly for JARGON and INFORMATION. Each point in the scatter overlay represents one of the 98 inputs. The three subsets (e.g., "College → Sr. Res. (n=98)") correspond to the evaluation that used the listed audience range and sample size; "Senior researcher" is abbreviated to "Sr. Res.". That is, the sample size increases from "College → Sr. Res. (n=98)" to "College → Sr. Res. (n=459)", while the audience range increases from "College → Sr. Res. (n=98)" to " Child → Expert (n=98)". Extreme outliers removed for visualization.

all models increase length across the 5 levels for the majority of their inputs (e.g., DeepSeek-V3.1 increases length across all levels for 98.98% of the inputs). However, an increase in length may not always indicate an increase in complexity, especially when JARGON and INFORMATION decrease. When directly examining responses, we notice that cases where length increases on its own (i.e., without increasing the other complexity measures) can be indicative of elaborative simplification (Wu et al., 2023b); an example is shown in Fig. 3.

5.3 Model struggle to differentiate responses in later audience levels

In addition to the overall performance discussed in the previous two findings, we consider how the performance varies by transition. We find that models increase complexity measures more often

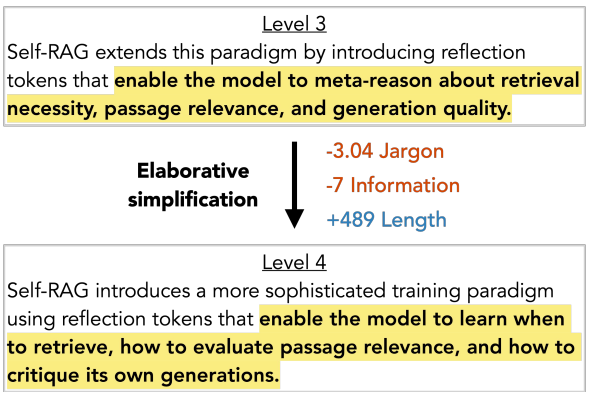


Figure 3: **Example of elaborative simplification** These are snippets of responses generated by Claude Sonnet 4.5 that are supposed to increase in complexity. Between the two, LENGTH increases, while JARGON and INFORMATION decrease. We notice that the additional text in the Level 4 snippet explains in simple language what "meta-reason" in the Level 3 snippet entails.

College → Sr. Res. (n=98)

Model	Jargon	Info.	Length	All
GPT-5.1	5.10	31.63	97.96	3.06
GPT-5 mini	13.27	14.29	76.53	0.00
Claude Sonnet 4.5	32.65	**42.86**	**100.0**	14.29
+ Thinking	**45.92**	32.65	86.73	**17.35**
DeepSeek-V3.1	18.37	24.49	98.98	6.12

College → Sr. Res. (n=459)

GPT-5.1	1.96	26.36	99.56	0.87
GPT-5 mini	3.92	5.88	69.93	0.22
Claude Sonnet 4.5	15.03	**32.90**	**100.0**	**5.45**
+ Thinking	**31.59**	28.76	83.66	4.79
DeepSeek-V3.1	7.19	7.41	98.47	0.44

Child → Expert (n=98)

GPT-5.1	18.37	**51.02**	**100.0**	11.22
GPT-5 mini	26.53	35.71	95.92	11.22
Claude Sonnet 4.5	60.20	45.92	**100.0**	**33.67**
+ Thinking	**79.59**	40.82	93.88	31.63
DeepSeek-V3.1	46.94	36.73	97.96	16.33

Table 2: **Model performance shown as percent of inputs where measures increase across all 5 levels.** These percentages represent how often models generate sets of levels that adhere to the desired increase in the complexity measures, so higher is better. The "All" column shows the percent of inputs for which the model increases all three measures.

	Model	Jargon	Info.	Length
1 to 2	GPT-5.1	94.90	**81.63**	**100.0**
	GPT-5 mini	87.76	72.45	**100.0**
	Claude Sonnet 4.5	**100.0**	80.61	**100.0**
	+ Thinking	98.98	77.55	**100.0**
	DeepSeek-V3.1	98.98	77.55	**100.0**
2 to 3	GPT-5.1	54.08	**83.67**	**100.0**
	GPT-5 mini	57.14	74.49	97.96
	Claude Sonnet 4.5	83.67	80.61	**100.0**
	+ Thinking	**88.78**	79.59	**100.0**
	DeepSeek-V3.1	69.39	74.49	**100.0**
3 to 4	GPT-5.1	55.10	71.43	**100.0**
	GPT-5 mini	65.31	57.14	80.61
	Claude Sonnet 4.5	71.42	**78.57**	**100.0**
	+ Thinking	**77.55**	75.51	98.98
	DeepSeek-V3.1	51.02	69.39	98.98
4 to 5	GPT-5.1	27.55	**77.55**	97.96
	GPT-5 mini	60.20	69.39	95.92
	Claude Sonnet 4.5	55.10	83.67	**100.0**
	+ Thinking	**76.53**	76.53	86.73
	DeepSeek-V3.1	60.20	75.51	**100.0**

Table 3: **Model performance per transition for College → Sr. Res. (n=98)** Each model's performance is shown as the percent of inputs where the measure goes in the correct direction at each transition. A higher percentage means that the model performed better at that transition, by more often increasing complexity.

when transitioning from Level 1 (College student) to Level 2 (Junior Ph.D.) than for the later transitions. This can be seen in Tab. 3 where all models have the highest proportion of inputs going in the correct direction of complexity in the "1 to 2" row compared to the later rows. For example, in the JARGON column, GPT-5 mini correctly increases complexity for 87.76% of the inputs when going from Level 1 to 2, which is higher than 57.14%, 65.31%, and 60.20% for the later transitions.

5.4 Increasing the sample size does not change these findings

The findings discussed thus far came from providing 98 queries and their expert-written reports as input. To test the effect of sample size, we evaluate a larger set of queries but with model-generated reports as input. We randomly sample 500 scientific queries from ScholarQABench (excluding the 98 original queries) and generate a report for each using the ScholarQA pipeline (Singh et al., 2025) with a Claude Sonnet 4.5 backend. We use these reports as input to each model and use the same framework as our original evaluation. We report results on 459 queries due to model refusal on some of the queries. Our results show the same mix of increasing and decreasing complexity measures (Fig. 2 and Tab. 2) along with the tendency to perform better when shifting from Level 1 to 2 than on the later transitions (per transition performance is provided in Appendix A.2).

5.5 Expanding audience levels can improve performance but exhibits the same trends

We chose the audiences "College student" through "Senior researcher" because the wording of the questions in ScholarQABench implied that the inquirer had at least a college education. We investigate if this choice of audience labels affected the similarity between levels by running an identical evaluation using a different set of audience labels. Specifically, we use labels from a popular video series called "5 Levels" by WIRED[2] that focuses on communicating specialized concepts to 5 different audiences: Child, Teen, College student, Grad student, and Expert. Results for this evaluation are included in Fig. 2 and Tab. 2. While complexity measures increase more often (i.e., the percentages of model responses with complexity measures in the correct direction are higher, Tab. 2), the same trend

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

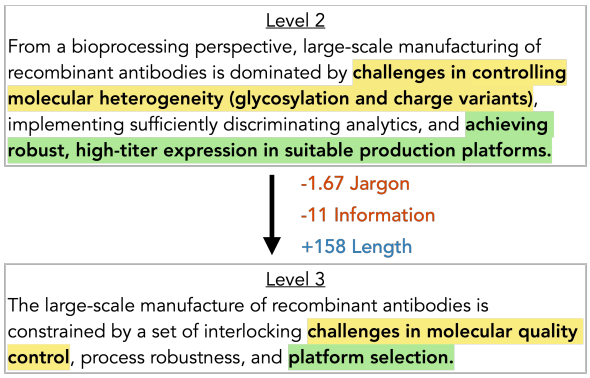


Figure 4: **Example of text incorrectly decreasing in complexity** Shown are two snippets of text generated by GPT-5.1 that are meant to increase in complexity from Level 2 to Level 3. However, we qualitatively observe that the complexity decreases between analogous phrases (e.g., "platform selection" in the Level 3 snippet reads simpler than "achieving robust, high-titer expression in suitable production platforms" from Level 2). Decreases in JARGON and INFORMATION reflect this observation; note that the measures represent the texts that the snippets are from, not the snippets in isolation, explaining why LENGTH increases.

	Model	Jargon	Info.	Length
1 to 2	GPT-5.1	**100.0**	81.63	**100.0**
	GPT-5 mini	98.98	81.63	98.98
	Claude Sonnet 4.5	**100.0**	87.76	**100.0**
	+ Thinking	**100.0**	**91.84**	**100.0**
	DeepSeek-V3.1	97.96	87.76	97.96
2 to 3	GPT-5.1	95.92	**92.86**	**100.0**
	GPT-5 mini	75.51	80.61	**100.0**
	Claude Sonnet 4.5	97.96	80.61	**100.0**
	+ Thinking	97.96	81.63	**100.0**
	DeepSeek-V3.1	**100.0**	81.63	**100.0**
3 to 4	GPT-5.1	63.27	78.57	**100.0**
	GPT-5 mini	65.31	73.47	98.98
	Claude Sonnet 4.5	79.59	83.67	**100.0**
	+ Thinking	**88.78**	**84.69**	**100.0**
	DeepSeek-V3.1	78.57	73.47	**100.0**
4 to 5	GPT-5.1	36.73	**85.71**	**100.0**
	GPT-5 mini	69.39	81.63	97.96
	Claude Sonnet 4.5	78.57	83.67	**100.0**
	+ Thinking	**92.86**	71.43	93.88
	DeepSeek-V3.1	64.29	82.65	**100.0**

Table 4: **Model performance per transition for Child → Expert (n=98)** Each model's performance is shown as the percent of inputs where the measure goes in the correct direction at each transition. A higher percentage means that the model performed better at that transition, by more often increasing complexity according to these measures.

of measures going in the wrong direction holds. For example, for INFORMATION, the performance of GPT-5.1 increased from 31.63% to 51.02% with the change in audience labels (Tab. 2); however, 51.02% is still close to chance level. Additionally, we observe the same trend where models differentiate between Levels 1 and 2 better than the later levels (Tab. 4). This can be seen for JARGON when the performance of GPT-5 mini decreases from 98.98% in the first transition to 69.39% in the last.

6 Discussion & Conclusion

Enabling users to directly adjust the language of model responses allows LLM-powered systems to better accommodate user needs. While prompting remains the default for adjusting model responses, new malleable interfaces promise more direct user control beyond articulating information needs through language (Min et al., 2025; Zhang et al., 2026). However, while interfaces are empowering users to interact with models beyond prompting, evaluations of model responses remain fixed to the traditional chat interface. In this paper, we propose the idea of evaluating models' potential for powering interfaces beyond chat. We do this by testing multiple responses relative to each other rather than single responses. This approach has parallels with other recent trends in model evaluations, such as evaluating multi-turn conversations (Laban et al., 2026), integrated or live use settings (Mehta et al., 2026; Bragg et al., 2026a), and performance measures based on different intended audiences (Joshi et al., 2025). Through a formative user study, we establish that these interactive use cases can be desirable, specifically for controlling language complexity, and identify measures and a criterion for distinguishing between levels of complexity.

To investigate how current models adhere to or stray from users' expectations, we evaluate model-generated levels of complexity for a dataset of scientific questions. We show that evaluating models across levels of complexity reveals model weaknesses that would be difficult to identify in single response evaluations. While models tend to increase length with complexity levels, they frequently decrease jargon and information, suggesting that models often neglect other attributes that are important for perceived complexity and end-user control. This finding holds even when increasing the sample size and extending complexity response anchors to encompass a wider range of intended audiences.

Limitations

We evaluated models using linguistic measures that were strongly supported by prior work and our formative study. However, there are other potential measures for complexity that could go beyond linguistic characteristics towards the content of the text, such as elaborative simplification and analogies. Thus, one avenue for future work would be to quantify and evaluate the impact of these additional measures.

Additionally, the questions in ScholarQABench are expert-written, scientific questions (Tab. 1). As a result, the content of the question may not always match the audience, particularly when we tested the WIRED audiences (e.g., a child is unlikely to ask about "ways to perform optomechanical cooling"). We did test a range of audiences that would be more likely to pose such questions (College student to Senior researcher); however, another option for future work could be to try queries that vary by audience.

Lastly, we establish that models do not consistently increase complexity, causing the differences between generated levels of complexity to fluctuate. However, even if models did follow the correct direction of complexity, we do not know exactly how much of a difference between levels there needs to be for a user to notice. Thus, an interesting follow-up work would be to quantify that difference by performing a human-centered evaluation on multiple versions of text that vary systematically in one or more complexity measures.

Ethical Considerations

This work uses LLMs to generate and evaluate responses. In addition to their impact on the environment (Ren et al., 2024; Desislavov et al., 2023), LLMs can hallucinate and exhibit biases that affect the information they generate (Venkit et al., 2024; Sharma et al., 2024; Zhou and Di Eugenio, 2025). At the same time, we believe that the contribution of this work towards making information from knowledge-intensive domains accessible to people of varying expertise is valuable.

References

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*, pages 74–82, New Orleans, Louisiana, USA. Association for Computational Linguistics.

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*, CHI '17, page 3188–3200, New York, NY, USA. Association for Computing Machinery.

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, and 6 others. 2026. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, page 857–863.

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*, CHI '24, New York, NY, USA. Association for Computing Machinery.

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)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

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).

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*, pages 152–178, Torino, Italia. ELRA and ICCL.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, and 20 others. 2026a. Astabench: Rigorous benchmarking of ai agents with a scientific research suite. *Preprint*, arXiv:2510.21652.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi Mishra, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha

Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, and 20 others. 2026b. AstaBench: Rigorous benchmarking of AI agents with a scientific research suite. In *The Fourteenth International Conference on Learning Representations*.

Isabel Cachola, Daniel Khashabi, and Mark Dredze. 2025. Evaluating the evaluators: Are readability metrics good measures of readability? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24011–24027, Suzhou, China. Association for Computational Linguistics.

Jeanne Sternlicht Chall and Edgar Dale. 1995. Readability revisited : the new dale-chall readability formula.

Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. 2023. Trends in ai inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems*, 38:100857.

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*, page 418–424, Berlin, Heidelberg. Springer-Verlag.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

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*, UIST '24, New York, NY, USA. Association for Computing Machinery.

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*, IUI '23, page 476–490, New York, NY, USA. Association for Computing Machinery.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024a. APPLS: Evaluating evaluation metrics for plain language summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9194–9211, Miami, Florida, USA. Association for Computational Linguistics.

Yue Guo, Joseph Chee Chang, Maria Antoniak, Erin Bransom, Trevor Cohen, Lucy Wang, and Tal August. 2024b. Personalized jargon identification for enhanced interdisciplinary communication. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4535–4550, Mexico City, Mexico. Association for Computational Linguistics.

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):160–168.

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*, CHI '21, New York, NY, USA. Association for Computing Machinery.

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, page 115–125, New York, NY, USA. Association for Computing Machinery.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.

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*, pages 25466–25499, Vienna, Austria. Association for Computational Linguistics.

Nikhita Joshi and Daniel Vogel. 2026. Designing and evaluating ai margin notes in document reader software. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA. Association for Computing Machinery.

Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025a. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Taewook Kim, Dhruv Agarwal, Jordan Ackerman, and Manaswi Saha. 2025b. Steering ai-driven personalization of scientific text for general audiences. *Proc. ACM Hum.-Comput. Interact.*, 9(7).

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):383–392.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*.

Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, and 17 others. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

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 *Proceedings of the 2nd Conference on Language Modeling (COLM)*.

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, and 36 others. 2023. The semantic reader project: Augmenting scholarly documents through ai-powered interactive reading interfaces. In *Communications of the ACM (CACM)*.

Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Sushant Mehta, Logan Ritchie, Suhaas Garre, Ian Niebres, Nick Heiner, and Edwin Chen. 2026. Enterprisebench corecraft: Training generalizable agents on high-fidelity rl environments. *Preprint*, arXiv:2602.16179.

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, New York, NY, USA. Association for Computing Machinery.

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*, pages 12076–12100, Singapore. Association for Computational Linguistics.

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*, pages 1–13.

Sonia K. Murthy, Kyle Lo, Daniel King, Chandra Bhagavatula, Bailey Kuehl, Sophie Johnson, Jon Borchardt, Daniel S. Weld, Tom Hope, and Doug Downey. 2022. Accord: A multi-document approach to generating diverse descriptions of scientific concepts. In *EMNLP: System Demonstrations*, pages 200–213. Association for Computational Linguistics.

OpenAI. 2025. ChatGPT with Deep Research.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802. Curran Associates, Inc.

Shaolei Ren, Bill Tomlinson, Rebecca W. Black, and Andrew W. Torrance. 2024. Reconciling the contrasting narratives on the environmental impact of large language models. *Scientific Reports*, 14(1):26310.

Paul Rust, Julian Frings, Sven Meister, and Leonard Fehring. 2025a. Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations. *Communications Medicine*, 5(1):208.

Paul Rust, Julian Frings, Sven Meister, and Leonard Fehring. 2025b. Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations. *Communications Medicine*, 5(1):208.

Johnny Saldaña. 2021. *The coding manual for qualitative researchers*, fourth edition edition. Sage Publications.

Science Journal for Kids. Science journal for kids and teens.

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

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)*, pages 513–523, Vienna, Austria. Association for Computational Linguistics.

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*, CHI '10, page 2247–2256, New York, NY, USA. Association for Computing Machinery.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227, Bangkok, Thailand. Association for Computational Linguistics.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 1–14, Online. Association for Computational Linguistics.

Jan Trienes, Sebastian Joseph, Jörg Schlötterer, Christin Seifert, Kyle Lo, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2024. InfoLossQA: Characterizing and recovering information loss in text simplification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4263–4294, Bangkok, Thailand. Association for Computational Linguistics.

Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. 2024. An audit on the perspectives and challenges of hallucinations in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6528–6548, Miami, Florida, USA. Association for Computational Linguistics.

Sharon Whitfield and Melissa A. Hofmann. 2023. Elicit: Ai literature review research assistant. *Public Services Quarterly*, 19(3):201–207.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023a. 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*, page 695–707, Berlin, Heidelberg. Springer-Verlag.

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023b. Elaborative simplification as implicit questions under discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537, Singapore. Association for Computational Linguistics.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Chao Zhang, Yiren Liu, Lunyiu Nie, Jeffrey M. Rzeszotarski, Yun Huang, and Tal August. 2026. From words to widgets for controllable llm generation. *Preprint*, arXiv:2604.10925.

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*.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Merging generated and retrieved knowledge for open-domain QA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4710–4728, Singapore. Association for Computational Linguistics.

Yue Zhou and Barbara Di Eugenio. 2025. Veracity bias and beyond: Uncovering LLMs' hidden beliefs in problem-solving reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21298–21310, Vienna, Austria. Association for Computational Linguistics.

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).

A Appendix

A.1 Prompt for Interactive Complexity

We use 5 levels because it allows for some nuance between the versions of text, and a popular video series called "5 Levels" by WIRED[3] focuses on communicating specialized concepts to 5 different audiences which aligns well with the motivation for this work. To populate the slider with 5 responses for the user study and run the model evaluation, we experimented with different characteristics of prompts, with highlights indicating what worked better: specifying audience vs. generic levels, single vs. multi-prompt, and defining endpoints of complexity vs. defining all 5 complexity levels. Our final prompt for the model evaluation is shown in Fig. 5; we allowed the model to include a "References" section in its responses during the user study. We also enforced a JSON schema.

We chose the audience labels to range from College student to Senior Researcher because the wording of the questions in ScholarQABench implied that the inquirer had at least a college education. However, as mentioned in Sec. 5.5, we tested using the audiences from the WIRED 5 levels series (Child, Teen, College student, Grad student, Expert) to check the effect of the audience labels, using the same prompt as a template.

A.2 Additional Data for Increased Sample Size

Tab. 5 shows how often models moved complexity in the correct direction per transition when evaluating the larger sample of 459 ScholarQABench queries (Sec. 5.4).

A.3 Model Configuration Details

Tab. 6 shows the model configurations for the model evaluation. We prompted all but 1 input only once; the 1 input that we had to regenerate was initially refused by Sonnet 4.5, possibly due to the topic. We ran the evaluation during November 2025 through January 2026.

A.4 Flesch-Kincaid Reading Ease Score Data

Fig. 6 shows the changes in Flesch-Kincaid scores across levels labeled with our initial audiences and the WIRED audiences. Flesch-Kincaid scores follow the opposite trend to the measures in the main paper; a *decrease* in the score indicates higher complexity. Similar to JARGON and INFORMATION (Fig. 2), we observe inconsistent increases in complexity (i.e., decreasing Flesch-Kincaid scores).

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

	Model	Jargon	Info.	Length
1 to 2	GPT-5.1	93.03	**84.75**	**100.0**
	GPT-5 mini	86.06	59.48	92.37
	Claude Sonnet 4.5	**100.0**	76.03	**100.0**
	+ Thinking	**100.0**	81.05	99.78
	DeepSeek-V3.1	99.56	69.50	99.56
2 to 3	GPT-5.1	44.01	**79.30**	**100.0**
	GPT-5 mini	51.20	66.01	99.35
	Claude Sonnet 4.5	74.29	76.69	**100.0**
	+ Thinking	**79.74**	76.47	99.78
	DeepSeek-V3.1	62.96	61.22	**100.0**
3 to 4	GPT-5.1	35.51	67.76	**100.0**
	GPT-5 mini	50.76	48.58	77.34
	Claude Sonnet 4.5	56.43	**80.17**	**100.0**
	+ Thinking	**63.83**	76.03	98.04
	DeepSeek-V3.1	48.37	59.04	99.35
4 to 5	GPT-5.1	25.05	70.59	99.56
	GPT-5 mini	47.93	69.28	98.04
	Claude Sonnet 4.5	45.32	**79.96**	**100.0**
	+ Thinking	**67.32**	66.45	85.19
	DeepSeek-V3.1	45.97	64.92	99.56

Table 5: **Model performance per transition for College → Sr. Res. (n=459)** Each model's performance is shown as the percent of inputs where the measure goes in the correct direction at each transition. A higher percentage means that the model performed better at that transition, by more often increasing complexity according to these measures.

A.5 Complexity of Responses in User Study

In this section, we report the changes in complexity for the aggregated 45 responses across the 16 participants from the interactive condition of the user study. We use GPT-5 mini to supply the responses as a low-latency option. The distributions generally match those of GPT-5 mini in the model evaluation (Fig. 2), confirming that the user study was a fair instantiation of the model evaluation.

A.6 Participant Details

Tab. 7 lists participants' background and LLM usage. Most participants held an academic affiliation and had a STEM background. To ascertain general LLM usage, we asked "How often do you use LLMs and LLM-infused applications?", and the choices were:

- Never
- Rarely, about 1–2 times a month
- Sometimes, about 3–4 times a month
- Often, about twice a week
- Always, about once or more a day

Some wording came from Kim et al. (2025a).

College student to Senior researcher Prompt

```
You are given a user query and a report responding to that query as input.

Using information only from the report and query, rewrite the chatbot response into 5 versions where
each version responds with a level of complexity appropriate for a College student, Junior Ph.D.
student, Senior Ph.D. student, Postdoctoral researcher, or Senior researcher respectively. That is,
Version 1 should be written with a level of complexity that a college student would understand,
Version 2 should be written with a level of complexity that a junior Ph.D. student would understand,
and so on. Assume that the user understands the words in their query.

Preserve mentions of papers and citations by preserving abbreviated, in-text citations. To be clear,
do not write out a References section, just use in-text citations like this: (Vaswani, 2017).

Do not add any additional text like greetings or ornamental words.
```

Figure 5: **Prompt for Interactive Complexity**

Model	Hyperparameters
`gpt-5.1-2025-11-13`	temperature = 0
`gpt-5-mini-2025-08-07`	none specified
`Claude Sonnet 4.5`	temperature = 0; max_tokens = 9000
`Claude Sonnet 4.5 + Thinking`	max_tokens = 11192 (includes 3000 for thinking)
`deepseek.v3-v1:0`	temperature = 0; maxTokens = 9000

Table 6: **Model configurations**

For LLM usage in research, we asked "Do you use LLMs and LLM-infused applications for any parts of the research process? Select all that apply.", and the choices were:

- Yes, for information seeking (e.g., discovering papers, generating summaries, discovering topics)
- Yes, for editing writing (e.g., fixing grammar or rephrasing, looking up synonyms, formatting papers)
- Yes, for direct writing (e.g., rewriting to another style, shortening, summarizing)
- Yes, for data cleaning & analysis (e.g., cleaning and reformatting data, statistical reporting, qualitative analysis)
- Yes, for ideation & framing (e.g., brainstorming research questions, coming up with ways to frame a paper, getting inspiration for methods)
- Yes, for data generation (e.g., generating synthetic data, producing examples and labels)
- Yes, for other purposes (Select this box and Other and specify below in Other)
- No
- Other:

These answer choices mostly came from Liao et al. (2025).

A.7 User Study Interface

We extended a chat interface from the Streamlit framework[4] to build our conventional and interactive interfaces. More details are provided in Sec. 3.1.

A.8 User Study Interview Guide

We include the interview guide we used to structure questions during the user study. The questions aim to probe at participants' perceptions of and interactions with both interfaces.

The following questions were asked for both conditions:

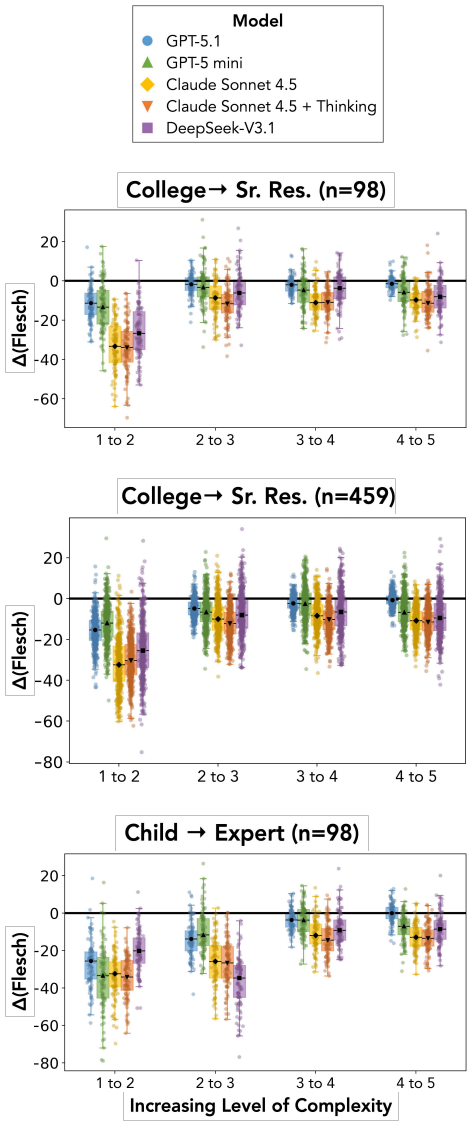
- How did you feel about the level of complexity in the chat responses? Were there times that you felt that you had too much or too little information? Was the amount of complexity appropriate or overwhelming?
- How did you use the system generally? What worked well and what didn't?
- What about the text in the responses would you change if anything?
- How did you feel about not reading the papers? (if the participant expressed something about this)

The following questions were asked only for the

[4] https://docs.streamlit.io/develop/tutorials/chat-and-llm-apps/build-conversational-apps

Background	Field of Study	Research Experience	LLM Usage	LLM Research Usage
Engineer	Computer Science	7 years	Always	Information seeking, Ideation & framing, Other: Coding, Indexing
Postdoctoral researcher	Comparative Literature	8 years	Rarely	Information seeking, Ideation & framing
1st year Ph.D. student	Computer Science	1 year	Rarely	None
Master's student	Computer Science	2 years	Always	Information seeking, Editing writing, Direct writing, Data cleaning & analysis, Ideation & framing
2nd year Ph.D. student	Computer Science	4 years	Always	Information seeking, Editing writing, Direct writing, Data cleaning & analysis, Ideation & framing, Data generation, Other: Getting feedback on paper drafts
Undergraduate student	Statistics, Actuarial Science	None	Often	Information seeking, Editing writing, Direct writing, Data cleaning & analysis, Ideation & framing
5th year Ph.D. student	Applied Mathematics	4 years	Always	Information seeking, Editing writing, Direct writing
Master's student	Computer Science	2 years	Always	Information seeking, Editing writing
Master's student	Computer Science	1 year	Always	Information seeking, Editing writing, Ideation & framing
4th year Ph.D. student	Computer Science	5 years	Often	Information seeking, Other: Coding
4th year Ph.D. student	Computer Science	4 years	Always	Information seeking, Editing writing, Direct writing, Data cleaning & analysis, Ideation & framing
2nd year Ph.D. student	Computational Biology	4 years	Often	Information seeking, Data cleaning & analysis
2nd year Ph.D. student	Computer Science	2 years	Always	Editing writing, Other: Coding
College Graduate	Statistics, Computer Science	None	Always	Information seeking, Editing writing, Data cleaning & analysis
Master's student	Electrical & Computer Engineering	2 years	Always	Information seeking, Editing writing, Ideation & framing
5th year Ph.D. student	Computer Science	6 years	Always	Editing writing, Ideation & framing

Table 7: **Participants' background and LLM usage**



Levels of Complexity: College → Sr. Res.	Levels of Complexity: Child → Expert
(1) College student	(1) Child
(2) Junior Ph.D. student	(2) Teen
(3) Senior Ph.D. student	(3) College student
(4) Postdoctoral researcher	(4) Grad student
(5) Senior researcher	(5) Expert

Figure 6: **Flesch-Kincaid Reading Ease score data** These plots show the distributions of changes in the Flesch-Kincaid scores between consecutive levels for the two sets of audience labels that we prompted with. Since a higher Flesch-Kincaid score means that the text is more readable, *decreasing* the scores as complexity increases is desirable.

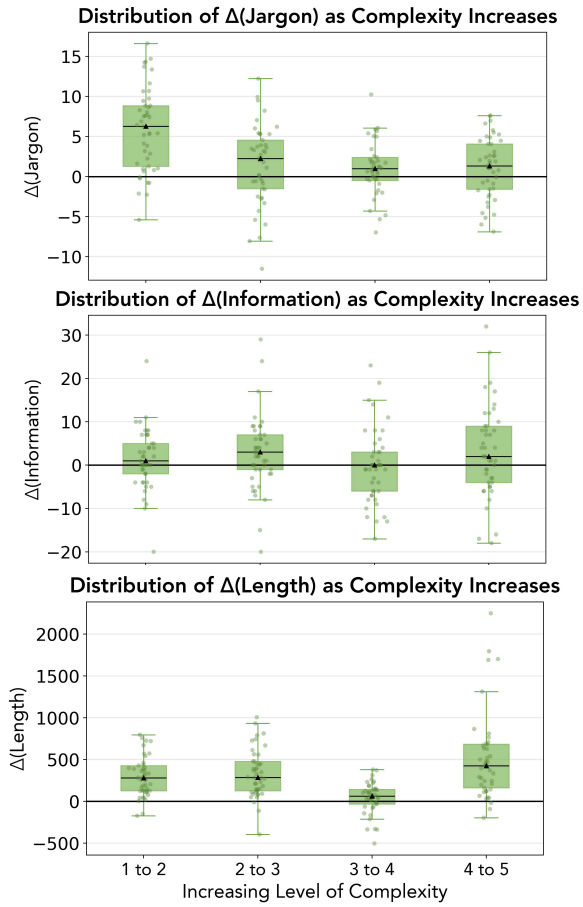


Figure 7: **Model performance during user study based on complexity measures** Between consecutive levels of complexity, the interactive condition had comparable changes in JARGON, INFORMATION, and LENGTH to the model evaluation (Fig. 2), varying between increasing and decreasing. We use GPT-5 mini to supply the responses as a low-latency option. Each point in the scatter overlay represents one of the 45 participant queries. Ideally, all measures should increase between levels (i.e., all points should be strictly above the zero line). Extreme outliers removed for visualization.

List 3 interesting facts that you learned:

1.

2.

3.

Find 1 paper that is relevant to the topic and list as much of the information below as possible.

Title:

Last name of first author:

Main Idea (1 sentence or phrase):

Why you chose this paper (1 sentence or phrase):

Figure 8: **Task Questions**

interactive condition:

- How did you feel about having a choice of 5 responses with varying complexity as opposed to one response with fixed complexity? [after participants had experienced both conditions]
- When did you find the slider helpful or not helpful?
- When did you ask follow-up questions versus use the slider versus use the response as is?
- Is there anything about the progression of the text between the 5 levels that you would want to control?
- Did the variations in complexity match what you expected from the 5 levels? If not, what would you have wanted?
- Was there anything hard about going through the different levels?
- Did you read all 5 levels, why or why not?
- Could you tell the difference between the 5 levels [or the levels you did read] and how?
- Does 5 levels feel appropriate?
- Do you want the slider for every response?

A.9 Study Materials

Participants were guided through a 1-hour Zoom session where they interacted with two interfaces, one conventional chat interface and one with interactive complexity as described in Sec. 3.1. For the conventional chat interface, participants were told "*This is a simple chatbot where you ask a question and it provides a response*". Since the interactive complexity version had slider features (Fig. 1), participants watched a 1-minute video tutorial. After learning how to use the interfaces, participants were then given task instructions as shown in Fig. 9. The task questions are displayed in Fig. 8; participants were also provided with a note-taking box in the same document. After the study, participants were provided with the disclaimer shown in Fig. 10.

Task Instructions Given to Participants

Before this session, you provided a topic: [topic provided by participant].
This document contains a few tasks about this topic for you to complete using the chatbot. It also contains an area for you to take notes if you would like to.

You will have 15 minutes to complete the tasks using only the chatbot; I will provide you with time warnings.

Please focus on using the chatbot rather than reading through links or papers. You can open links to verify content from the system, but do not read the entire paper.

While doing the tasks, you will be "thinking aloud". Basically, we want you to tell us everything that goes through your mind from the start to the end as you complete the tasks. When you are thinking aloud:
- There are no right or wrong answers. You are not being tested.
- Be honest. If you feel confused, frustrated, or surprised, please say so.
- Keep talking. Demonstrate your entire process, what you're doing, why you're doing it, what you're thinking, and what you're expecting
- Focus on the task. Complete the tasks to the best of your ability, while simultaneously describing your process.

For example, if you click something, you could say "I am clicking this because" followed by your reason. If you're confused, you could say "I'm not sure what to do here because" again followed by a reason.

Figure 9: **Task Instructions Given to Participants**

Post-Study Disclaimer

To create a realistic setting, we showed AI answers that are directly from responses from an actual AI system. As known, AI systems can make up information. Please note that the AI answers you saw in this study may have been inaccurate, incomplete, or inconsistent, even when they sounded convincing.

Figure 10: **Post-Study Disclaimer**