

# Explain Like I’m 5 or Whatever I Choose: Evaluating the Potential for Interactive Complexity in Language Model Responses

Indu Panigrahi and Tal August

Siebel School of Computing and Data Science

University of Illinois Urbana-Champaign

{indup2, taugust}@illinois.edu

## Abstract

Given the increased use of large language models (LLMs) for reading and seeking information and the varying needs and backgrounds of users, direct manipulation of text can be a valuable mechanism for end-user control. However, enabling direct manipulation requires that models generate text consistent with user-requested changes. We focus on one such setting, manipulating the complexity of scientific information, and ask if models can generate multiple summaries along a continuum of text complexity that enable effective user selection and control. Through a within-subjects user study, we identify reliable complexity measures and a criterion for how models should conform to these measures. We evaluate five LLMs of different families, sizes, and reasoning abilities on 557 scientific queries. We find that while models can vary complexity across responses, most changes remain inconsistent, with models shifting the complexity measures in the correct direction at most 33.67% of the time.

## 1 Introduction

Scientists and students are increasingly using LLMs for reading (Fok et al., 2023; Lo et al., 2023) and synthesizing literature (Asai et al., 2026; OpenAI, 2025). Readers vary in their preferred responses based on their background (e.g., a junior vs. senior researcher) and context (e.g., reading within a familiar topic vs. new discipline), 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 a given user (Guo et al., 2024b; Murthy et al., 2022). However, a single response can be misaligned with a user or fail to incorporate correct context (Gu et al., 2024). Further, responses may be inherently insufficient

even if correct (Fok et al., 2024). 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 request more details, repetitive prompting is time-intensive (Zamfirescu-Pereira et al., 2023). 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).

From an evaluation standpoint, this requires a shift towards evaluating the quality of *multiple* responses and how they relate to each other, rather than of single responses (Joshi et al., 2025; Murthy et al., 2022). Thus, 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. Triangulating with prior work, we identify measures of language complexity through a within-subjects user study where participants ( $N = 16$ ) explored unfamiliar STEM<sup>1</sup> 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 models of different sizes, families, and reasoning capabilities (GPT-5.1, GPT-5 mini, Claude Sonnet 4.5  $\pm$  Thinking, and DeepSeek-V3.1) by generating 5 versions, at different levels of response complexity<sup>2</sup>, of answers to 557 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, one of the better-performing models, Claude Sonnet 4.5 + Thinking, increases complexity 17.35% of the time across the

<sup>1</sup>Science, Technology, Engineering, and Mathematics

<sup>2</sup>We label 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.

five versions for an individual query (and at most 76.53% of the time for a single transition between versions). Additionally, models increase complexity measures less consistently between higher complexity responses than between lower complexity responses (e.g., 80.61% of the time between the first two versions reduces to 43.88% between the last two versions for Claude Sonnet 4.5). In summary, our paper makes the following contributions:

1. A new evaluation approach that tests models' ability to generate multiple versions of a response across a dimension of interest. We instantiate the framework for scientific information seeking and language complexity.
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, further justified with additional human validation.
3. Evaluation results from 5 models 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 the complexity levels.

## 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

Inspired by small-scale formative studies in Human-Computer Interaction research (Kim et al., 2023; Sauro and Lewis, 2012), and their growing use in NLP to identify a contextualized task of interest to then evaluate at scale (Guo et al., 2024b; Kim et al., 2023; Bhatt et al., 2025), we conduct a formative study with 16 participants to structure our subsequent model evaluation (Sec. 4). Our formative study: (1) validates the utility of interactive complexity in the context of scientific information seeking, (2) corroborates characteristics of language associated with user-perceived complexity, and (3) identifies requirements for model responses that enable successful interactive complexity.

### 3.1 Study Design and Analysis

**Procedure** We recruited 16 participants to interact with two interfaces. An interactive complexity chat provided users with a slider mechanism to adjust the language complexity of responses, choosing from 5 levels, 1 being the least complex (Fig. 1). The conventional interface looked and functioned similar to a basic chat interface (i.e., without the slider). All studies were counterbalanced between

the two interfaces.

To emulate the experience of exploring new, knowledge-intensive topics with the interfaces, we asked each participant to provide two STEM topics before the study that they had interest in but little to no knowledge about. For each interface, participants were instructed to identify facts and a paper within their chosen topics.

While interacting with each interface, participants actively verbalized their reasoning and impressions. After each condition, we interviewed participants about their experience with the interface. More details and materials are provided in Appendix A.5.

**Response Generation** We used the participant-provided topics to generate scientific reports using a RAG-based pipeline (ScholarQA, Singh et al., 2025) with Claude Sonnet 4.5. Using these reports as inputs, we prompted GPT-5 mini<sup>3</sup> to generate 5 versions of the reports for audiences defined as College student, Junior Ph.D. student, Senior Ph.D. student, Postdoctoral researcher, and Senior researcher<sup>4</sup> to populate the slider. Appendix A.1 includes the prompt and describes alternative prompts we tested.

**Analysis** To analyze the participant responses, we employ open coding (Saldaña, 2021) (details in Appendix A.5.6). 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 participants (13/16) appreciated the flexibility that interactive complexity provided beyond prompting. 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, seven participants (four of whom had not yet seen the interactive condition) ended up describing their preferred level

<sup>3</sup>Chosen for low latency appropriate for an interactive context.

<sup>4</sup>We based these levels off of previous work that stratifies model responses using level of education (Joshi et al., 2025; August et al., 2024), adjusting the levels to be appropriate for the envisioned prompts (i.e., scientific literature queries).

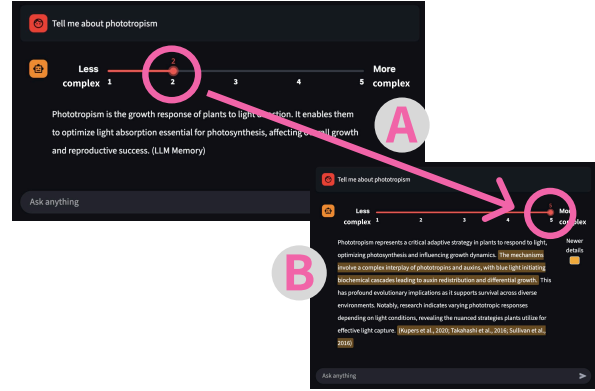


**Figure 1: Interactive language complexity interface** Users can manipulate textual complexity by moving the response slider (A) to a different notch. Sentences that are significantly different from those in the previously-displayed version are highlighted (B); significant differences are determined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

of complexity in follow-up prompts, indicating a strong desire for control over perceived complexity of responses.

**Jargon, information, and length influenced perceived complexity** As participants decreased complexity, they prioritized three primary, desired trends: decreases in scientific jargon (16/16), amount of information (13/16), and length (12/16). This finding reinforces what prior work has proposed for measuring complexity (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 five levels of complexity seemed generally appropriate (10/16), ten participants noted the levels needed to be more distinct from each other. In particular, six participants shared the sentiment that P9 expressed of “one and two [the least complex versions] and four and five [didn’t] feel all that much different”. 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. 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 over model responses (Sec. 4.4).

#### 4.1 Scientific Query Data

We use 557 queries from ScholarQABench (Asai et al., 2026) which contains expert-written scientific queries across multiple STEM fields (Appendix A.2.1). 98 of these queries come with expert-written answer reports. We restrict one evaluation to this subset to use the expert-written reports for grounding model responses to a single ground truth report (similar to the user study); this choice aligns 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 run a second evaluation on a larger sample of queries from the rest of ScholarQABench with reports generated from ScholarQA (Singh et al., 2025), similar to the user study. Specifically, we randomly sample 500 scientific queries from ScholarQABench (excluding the aforementioned 98) 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 inputs to each model. We report results on 459 queries due to model refusal on some of the queries.

#### 4.2 Models

We prompt models the same way as for the user study (Sec. 3.1)—to generate 5 versions of the input, stratified for five envisioned audiences going from least to most complex (College student → Senior researcher). Since we label each version with an intended audience, we run a third evaluation that uses alternative audience labels (Child → Expert) (Sec. 5.4). We evaluate 5 models across different sizes, model families, and reasoning abilities: GPT-5.1, GPT-5 mini, Claude Sonnet 4.5 (with and without Thinking), and DeepSeek-V3.1. Details about model configurations are in Appendix A.3.

#### 4.3 Complexity Measures

Past work has quantified complexity through several linguistic measures. These include readability formulas (Flesch, 1948), lexical features, and perplexity measures (Guo et al., 2024a; August et al., 2024; Joshi et al., 2025). We

triangulate between these works and our formative study results (Sec. 3.2) to curate a suite of three measures,<sup>5</sup> each representing different dimensions of perceived language complexity:

- JARGON (Guo et al., 2024a; Joshi et al., 2025): This measure denotes the number of words that are considered specialized or technical. We calculate the percentage of words in the response that are not on the Dale-Chall Word List, a list of 3,000 familiar English words (Chall and Dale, 1995).
- INFORMATION (Tienes 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 the number of independent facts, using the pipeline for generating “atomic facts” from Min et al. (2023); we validate the performance by manually inspecting a subset of 25 examples (Appendix A.6.1).
- LENGTH (Joshi et al., 2025; Guo et al., 2024a): More complex language, similar to containing more information, is often longer. However, simpler language can also be longer when the same information is explained or elaborated<sup>6</sup> on (Wu et al., 2023b). We quantify length as the total number of response characters.

#### 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 as a model’s ability to increase complexity measures as the intended complexity levels increase. We operationalize this by evaluating the *direction of changes* in the complexity measures across the 5 versions of a response. **Models that are better at generating distinct levels of complexity will produce positive changes in all measures.**

<sup>5</sup>The Flesch-Kincaid Reading Ease Score (Flesch, 1948) is a commonly-used scale for rating complexity; however, past work has shown that Flesch-Kincaid scores do not provide a reliable measure of complexity (Cachola et al., 2025). We focus on the measures confirmed in our user study and provide Flesch-Kincaid data (reflects similar trends as the other measures) in Appendix A.4.1.

<sup>6</sup>We consider information density (the ratio between INFORMATION and LENGTH) as way to measure elaborative simplification. However, we already include both measures separately, and model performance was lower than and reflected the same trends as with our selected measures (Appendix A.4.2)

Our criterion rests on the assumption that our measures of complexity capture a user’s perception of complexity. To validate this assumption, we perform a comparison between the complexity measures and human ratings of perceived complexity. We randomly sampled 100 pairs of generated text from consecutive levels in our model evaluation (evenly sampled across the four transitions and five models) and conducted manual binary comparisons of which text was perceived as more complex. Comparisons were done by two authors in the research team in a blind setting<sup>7</sup>. Importantly, ratings were not focused on measuring JARGON, INFORMATION, or LENGTH, but rather a user’s overall perceived complexity. We compare agreement, measured as Cohen’s  $\kappa$ <sup>8</sup>, between these human ratings and the complexity measures (Tab. 1). We find that the agreement between the ratings and the combination of all measures is substantial ( $\kappa = 0.84$ ), and much higher than any single measure alone (e.g., JARGON alone is  $\kappa = 0.56$ ). This further emphasizes the importance of all measures increasing together as opposed to individually.

| Measure(s)    | Agreement   | # Agreed / # Inputs |
|---------------|-------------|---------------------|
| Jargon        | 0.56        | 78 / 100            |
| + Information | 0.77        | 53 / 60             |
| + Length      | <u>0.83</u> | 64 / 70             |
| Information   | 0.36        | 68 / 100            |
| + Length      | 0.69        | 59 / 70             |
| Length        | 0.60        | 80 / 100            |
| All           | <b>0.84</b> | 46 / 50             |

Table 1: **Agreement on measures for criterion** Each row corresponds to either a single measure (e.g., JARGON) or a combination of measures (e.g., JARGON + INFORMATION). “Agreement” indicates the agreement between the measure(s) and the human rater as Cohen’s  $\kappa$ . “# Agreed / # Inputs” indicates how many inputs the ratings and measure(s) agree on out of the number of inputs that the measure(s) agree on.

## 5 Results

We present results from the three evaluations, varying the input source (expert-written and model-generated) and the audience labels (College student  $\rightarrow$  Senior researcher and Child  $\rightarrow$  Expert).

<sup>7</sup>The texts within each sampled pair were shuffled, and two authors rated the first 25 pairs. Since the authors achieved high agreement (88%, Cohen’s  $\kappa = 0.75$ ), one author rated the remaining 75 pairs.

<sup>8</sup>Chosen for two raters classifying samples according to a binary category.

Since the two evaluations that vary the input source share the same trends, we discuss them together throughout Sec. 5.1-5.3. Sec. 5.4 discusses the evaluation of alternate audience labels.

Fig. 2 plots the change in each measure between consecutive levels of complexity (descriptive statistics provided in Appendix A.2.2). Tab. 2 reports the percent of changes that move in the correct direction (i.e., increasing JARGON, INFORMATION, or LENGTH). Because Fig. 2 and Tab. 2 separate results by the four transitions between levels, we also report the overall performance across all transitions in Tab. 3 (i.e., the percent of inputs for which each model moves in the correct direction at all transitions between the five levels). To test if the distribution of changes moving in the wrong direction is significantly different from the distribution of changes moving in the right direction at each transition, we run Welch’s  $t$ -tests, with a Bonferroni correction, comparing those distributions at each of the four transitions (Appendix A.2.3). Lastly, we confirm in Appendix A.5.3 that the distribution of the responses we report here reflects the response distribution from our user study. We also check factual consistency and consistency between runs (Appendix A.6).

### 5.1 Models inconsistently increase complexity measures

Models often increase the complexity for some inputs while decreasing the complexity for others for the same transition. In Fig. 2, this can be seen where distributions cover positive and negative changes for the same model and transition, particularly in JARGON and INFORMATION. For example, when transitioning from Level 3 to 4, Claude Sonnet 4.5 + Thinking increases all three measures for only 58.16% of the expert-written inputs (48.58% for model-generated inputs) (Tab. 2). Fig. 4 shows an example where JARGON and INFORMATION both decrease when the complexity is supposed to increase. LENGTH often moves in the correct direction; however, as shown by our criterion, all measures should move in the correct direction together (discussed further in Sec. 5.2).

When considering all transitions (Tab. 3), Claude Sonnet 4.5 + Thinking is the best-performing but only consistently increases complexity measures 17.35% of the time (4.79% for generated inputs). Our significance tests further show that for JARGON and INFORMATION, the mean of changes moving in the wrong direction is

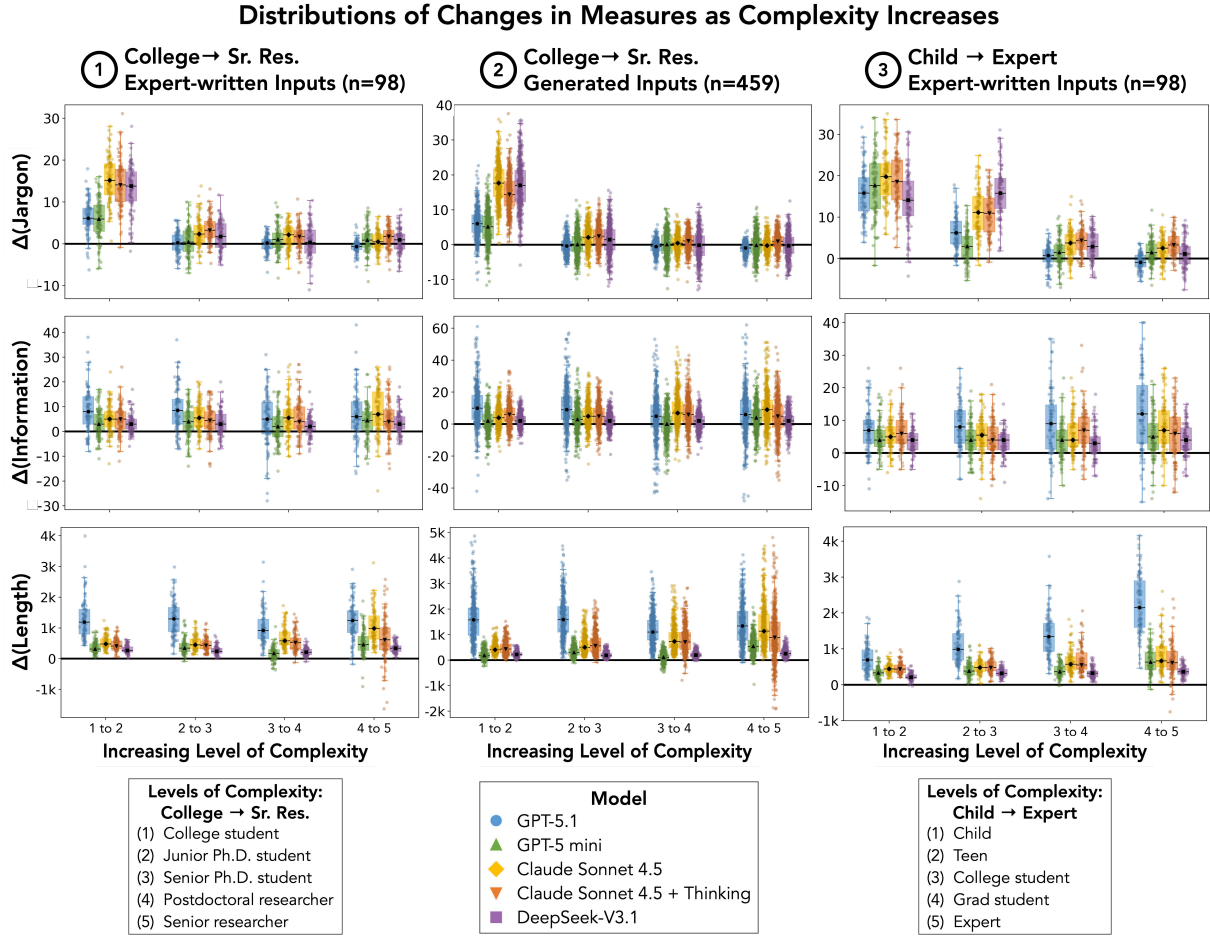


Figure 2: **Model performance shown as changes in complexity measures for the three evaluations** Each point in the scatter overlays represents an input. Between consecutive levels of complexity, models produce changes in JARGON, INFORMATION, and LENGTH that vary between increasing and decreasing, particularly for JARGON and INFORMATION, regardless of input source and audience labels. Extreme outliers removed for visualization.

often significantly different from those moving in the right direction ( $p < 0.01$  for expert-written and generated inputs), implying responses that fail to increase in complexity instead significantly decrease in complexity (as opposed to having no significant difference) (Appendix A.2.3).

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, models generally increase LENGTH with increasing complexity. This can be seen in Tab. 3 as all models increase LENGTH across the 5 levels for the majority of their inputs (e.g., DeepSeek-V3.1 increases LENGTH across all levels 98.98% of the time, 98.47% for generated inputs), reflecting the

per-transition performance in Tab. 2. However, an increase in LENGTH alone may not always indicate an increase in complexity. When directly examining responses, we notice that cases where LENGTH increases on its own (without increasing JARGON or INFORMATION) can be indicative of elaborative simplification (Wu et al., 2023b); an example is shown in Fig. 3.

## 5.3 Models 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 when transitioning from Level 1 to Level 2 than for the later transitions. This can be seen in Tab. 2 where most models have the highest percentage of inputs moving in the correct direction of complexity in the “1 to 2” row compared to the later rows. For

| ① College → Sr. Res.<br>Expert-written (n=98) |               |              |              |              | ② College → Sr. Res.<br>Generated (n=459) |              |              |              | ③ Child → Expert<br>Expert-written (n=98) |              |              |              |              |
|-----------------------------------------------|---------------|--------------|--------------|--------------|-------------------------------------------|--------------|--------------|--------------|-------------------------------------------|--------------|--------------|--------------|--------------|
| Model                                         | Jar.          | Inf.         | Len.         | All          | Jar.                                      | Inf.         | Len.         | All          | Jar.                                      | Inf.         | Len.         | All          |              |
| 1 to 2                                        | GPT-5.1       | 94.90        | <b>81.63</b> | <b>100.0</b> | 77.55                                     | 93.03        | <b>84.75</b> | <b>100.0</b> | 79.09                                     | <b>100.0</b> | 81.63        | <b>100.0</b> | 81.63        |
|                                               | GPT-5 mini    | 87.76        | 72.45        | <b>100.0</b> | 62.25                                     | 86.06        | 59.48        | 92.37        | 48.58                                     | 98.98        | 81.63        | 98.98        | 80.61        |
|                                               | Sonnet 4.5    | <b>100.0</b> | 80.61        | <b>100.0</b> | <b>80.61</b>                              | <b>100.0</b> | 76.03        | <b>100.0</b> | 76.04                                     | <b>100.0</b> | 87.76        | <b>100.0</b> | 87.76        |
|                                               | + Thinking    | 98.98        | 77.55        | <b>100.0</b> | 76.53                                     | <b>100.0</b> | 81.05        | 99.78        | <b>81.05</b>                              | <b>100.0</b> | <b>91.84</b> | <b>100.0</b> | <b>91.84</b> |
|                                               | DeepSeek-V3.1 | 98.98        | 77.55        | <b>100.0</b> | 76.53                                     | 99.56        | 69.50        | 99.56        | 69.06                                     | 97.96        | 87.76        | 97.96        | 83.67        |
| 2 to 3                                        | GPT-5.1       | 54.08        | <b>83.67</b> | <b>100.0</b> | 43.88                                     | 44.01        | <b>79.30</b> | <b>100.0</b> | 35.29                                     | 95.92        | <b>92.86</b> | <b>100.0</b> | <b>88.78</b> |
|                                               | GPT-5 mini    | 57.14        | 74.49        | 97.96        | 42.86                                     | 51.20        | 66.01        | 99.35        | 33.33                                     | 75.51        | 80.61        | <b>100.0</b> | 60.20        |
|                                               | Sonnet 4.5    | 83.67        | 80.61        | <b>100.0</b> | 67.35                                     | 74.29        | 76.69        | <b>100.0</b> | 56.43                                     | 97.96        | 80.61        | <b>100.0</b> | 79.59        |
|                                               | + Thinking    | <b>88.78</b> | 79.59        | <b>100.0</b> | <b>71.43</b>                              | <b>79.74</b> | 76.47        | 99.78        | <b>62.75</b>                              | 97.96        | 81.63        | <b>100.0</b> | 79.59        |
|                                               | DeepSeek-V3.1 | 69.39        | 74.49        | <b>100.0</b> | 54.08                                     | 62.96        | 61.22        | <b>100.0</b> | 38.78                                     | <b>100.0</b> | 81.63        | <b>100.0</b> | 81.63        |
| 3 to 4                                        | GPT-5.1       | 55.10        | 71.43        | <b>100.0</b> | 38.78                                     | 35.51        | 67.76        | <b>100.0</b> | 25.49                                     | 63.27        | 78.57        | <b>100.0</b> | 50.00        |
|                                               | GPT-5 mini    | 65.31        | 57.14        | 80.61        | 27.55                                     | 50.76        | 48.58        | 77.34        | 20.26                                     | 65.31        | 73.47        | 98.98        | 51.02        |
|                                               | Sonnet 4.5    | 71.42        | <b>78.57</b> | <b>100.0</b> | 55.10                                     | 56.43        | <b>80.17</b> | <b>100.0</b> | 45.10                                     | 79.59        | 83.67        | <b>100.0</b> | 66.33        |
|                                               | + Thinking    | <b>77.55</b> | 75.51        | 98.98        | <b>58.16</b>                              | <b>63.83</b> | 76.03        | 98.04        | <b>48.58</b>                              | <b>88.78</b> | <b>84.69</b> | <b>100.0</b> | <b>75.51</b> |
|                                               | DeepSeek-V3.1 | 51.02        | 69.39        | 98.98        | 35.71                                     | 48.37        | 59.04        | 99.35        | 28.32                                     | 78.57        | 73.47        | <b>100.0</b> | 57.14        |
| 4 to 5                                        | GPT-5.1       | 27.55        | <b>77.55</b> | 97.96        | 18.37                                     | 25.05        | 70.59        | 99.56        | 17.43                                     | 36.73        | <b>85.71</b> | <b>100.0</b> | 32.65        |
|                                               | GPT-5 mini    | 60.20        | 69.39        | 95.92        | 41.84                                     | 47.93        | 69.28        | 98.04        | 32.24                                     | 69.39        | 81.63        | 97.96        | 56.12        |
|                                               | Sonnet 4.5    | 55.10        | 83.67        | <b>100.0</b> | 43.88                                     | 45.32        | <b>79.96</b> | <b>100.0</b> | 34.86                                     | 78.57        | 83.67        | <b>100.0</b> | <b>65.31</b> |
|                                               | + Thinking    | <b>76.53</b> | 76.53        | 86.73        | <b>55.10</b>                              | <b>67.32</b> | 66.45        | 85.19        | <b>38.56</b>                              | <b>92.86</b> | 71.43        | 93.88        | <b>65.31</b> |
|                                               | DeepSeek-V3.1 | 60.20        | 75.51        | <b>100.0</b> | 47.96                                     | 45.97        | 64.92        | 99.56        | 29.85                                     | 64.29        | 82.65        | <b>100.0</b> | 52.04        |

Table 2: **Model performance per transition** Each model’s performance is shown as the percent of inputs for which the measure moves in the correct direction at each transition; “All” means where all three measures move in the correct direction. A higher percentage means that the model performed better by more often increasing complexity.

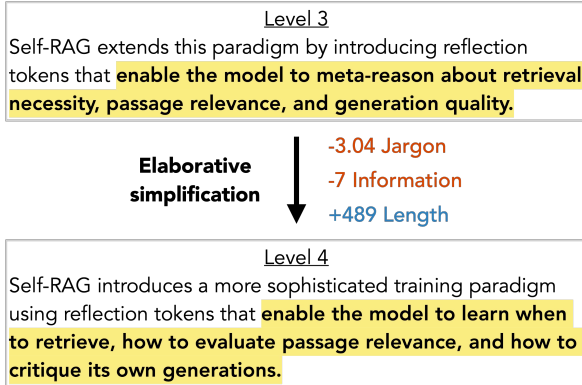


Figure 3: **Example of elaborative simplification** Snippets of responses generated by Claude Sonnet 4.5 that are supposed to increase in complexity. Between the two, LENGTH increases, while JARGON and INFORMATION decrease. The additional text in the Level 4 snippet explains in simple language the term “meta-reason” in the Level 3 snippet.

example, GPT-5 mini correctly increases all measures for 62.25% of the inputs when going from Level 1 to 2, which is higher than 42.86%, 27.55%, and 41.84% for the later transitions.

#### 5.4 Expanding audience labels can improve performance but exhibits the same trends

We chose the audiences “College student” through “Senior researcher” because the wording of the

queries in ScholarQABench implies that the inquirer has at least a college education. We investigate if this choice of audience labels affects the similarity between levels by running an identical evaluation to the first using a different set of audience labels. Specifically, we use labels from a popular video series called “5 Levels” by WIRED<sup>9</sup> that focuses on communicating specialized concepts to 5 different audiences: Child, Teen, College student, Grad student, and Expert.

While complexity measures increase more often (i.e., the percentages of responses moving in the correct direction are generally higher than those in the first evaluation, Tab. 3), the same trend of measures moving in the wrong direction holds. For example, the percent of inputs that Claude Sonnet 4.5 increases in complexity across all transitions improved from 14.29% to 33.67% with the change in audience labels (Tab. 3); however, 33.67% remains relatively low.

In terms of per-transition performance (Tab. 2), the “Child → Expert” labels increase complexity measures more often than the “College → Senior researcher” labels in the first evaluation, suggesting that the model interprets the former as more distinct from each other. However, we

<sup>9</sup><https://www.wired.com/video/series/5-levels/>

① College → Sr. Res., Expert-written (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         |
| Sonnet 4.5    | 32.65        | <b>42.86</b> | <b>100.0</b> | 14.29        |
| + Thinking    | <b>45.92</b> | 32.65        | 86.73        | <b>17.35</b> |
| DeepSeek-V3.1 | 18.37        | 24.49        | 98.98        | 6.12         |

② College → Sr. Res., Generated (n=459)

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

③ Child → Expert, Expert-written (n=98)

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

Table 3: **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.

observe the same trend where models differentiate between Levels 1 and 2 better than the later levels. This can be seen when the performance of GPT-5 mini decreases from 80.61% in the first transition to 56.12% in the last (Tab. 2). Similar to the first evaluation, our significance tests further imply that for JARGON and INFORMATION, responses that fail to increase in complexity often significantly decrease in complexity ( $p < 0.01$  for expert-written and generated inputs) (Appendix A.2.3).

## 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). In this paper, we propose an evaluation approach for these settings that tests multiple responses relative to each other rather than single responses in isolation. Through a formative user study and additional human validation, we establish that these interactive use cases can be

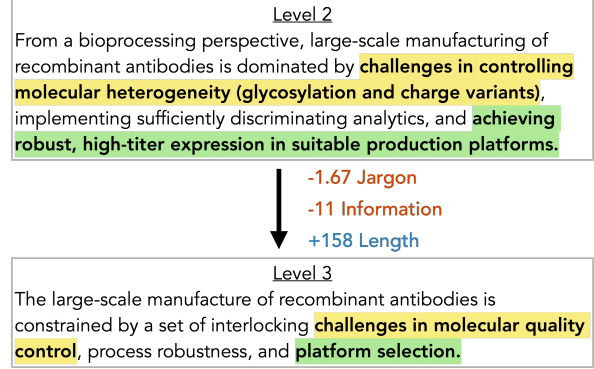


Figure 4: **Example of text incorrectly decreasing in complexity** Shown are two snippets of text generated by GPT-5.1 that are meant to increase in complexity from Level 2 to Level 3. However, we qualitatively observe that the complexity decreases between analogous phrases (e.g., “platform selection”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

desirable, specifically for controlling language complexity, and identify measures and a criterion for distinguishing between levels of complexity. To investigate how current models fulfill or stray from users’ expectations, we evaluate model-generated levels of complexity for answers to a dataset of scientific queries, varying the answer source and range of intended audiences. We show that evaluating models across levels of complexity reveals model weaknesses that would be difficult to identify in single-response evaluations. Specifically, models do not consistently increase complexity and that their performance can worsen for higher complexity levels. Further, while models tend to increase length with complexity levels, they frequently decrease jargon and information, suggesting that models often neglect attributes that are important for perceived complexity and end-user control.

Our evaluation 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., 2026), and performance measures based on different intended audiences (Joshi et al., 2025). More generally, 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.

## 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, potentially with additional open-source LLMs as well to increase accessibility.

We structure our user study around the target setting of people learning unfamiliar scientific topics. While sample sizes for formative studies are often small (Guo et al., 2024b; Kim et al., 2023; Sauro and Lewis, 2012), we noticed the same salient comments and interactions among our participants, suggesting data saturation for readers in unfamiliar STEM fields. We recognize this is a focused subset of the broader audience of scientific papers; an interesting expansion would be to analyze text from other disciplines.

Lastly, there are also other options for interactivity to explore. For our evaluation, the model generates a full response for each of the five levels, and we test different prompting techniques for this (Appendix A.1). However, one could imagine, for example, an interface where users select a specific snippet of text to be rewritten into a different complexity, which leads to other possibilities for direct manipulation.

## 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. Additionally, we focus on English text; however, this does not necessarily account for contexts with ESL learners who may have different experiences that impact scientific information seeking.

## 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 Podoyntsyna. 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.
- Shaily Bhatt, Tal August, and Maria Antoniak. 2025. [Research borderlands: Analysing writing across research cultures](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26238–26266, Vienna, Austria. Association for Computational Linguistics.

- Jonathan Bragg, Mike D’Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Hadad, 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. 2026. [Astabench: Rigorous benchmarking of ai agents with a scientific research suite](#). *Preprint*, arXiv:2510.21652.
- 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. [Simplifmytext: 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.
- Ziwei Gu, Ian Arawjo, Kenneth Li, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. [An ai-resilient text rendering technique for reading and skimming documents](#). In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, 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.
- Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. ["help me help the ai": Understanding how explainability can support human-ai interaction](#). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 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*.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.
- 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 csw 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, Xinxu 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.
- Jeff Sauro and James R. Lewis. 2012. *Chapter 7 - What Sample Sizes Do We Need?: Part 2: Formative Studies*, page 143–184. Morgan Kaufmann, Boston.
- 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, Lunyu Nie, Jeffrey M. Rzeszutarski, 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<sup>10</sup> 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. We include example results for our choice between single vs. multi-prompt in Tab. 4. 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.

| Jargon            |        |        |        |        |
|-------------------|--------|--------|--------|--------|
| Model             | 1 to 2 | 2 to 3 | 3 to 4 | 4 to 5 |
| GPT-5.1           | 70.10  | 50.52  | 61.86  | 63.92  |
| GPT-5 mini        | 49.48  | 80.41  | 44.33  | 67.01  |
| Claude Sonnet 4.5 | 96.91  | 76.29  | 47.42  | 51.55  |
| + Thinking        | 96.91  | 85.57  | 49.48  | 44.33  |
| DeepSeek-V3.1     | 67.01  | 72.16  | 57.73  | 53.61  |

| Information       |        |        |        |        |
|-------------------|--------|--------|--------|--------|
| Model             | 1 to 2 | 2 to 3 | 3 to 4 | 4 to 5 |
| GPT-5.1           | 69.07  | 59.79  | 57.73  | 43.30  |
| GPT-5 mini        | 85.57  | 64.95  | 44.33  | 43.30  |
| Claude Sonnet 4.5 | 65.98  | 43.30  | 45.36  | 32.99  |
| + Thinking        | 67.01  | 28.87  | 64.95  | 25.77  |
| DeepSeek-V3.1     | 54.64  | 50.52  | 44.33  | 41.24  |

| Length            |        |        |        |        |
|-------------------|--------|--------|--------|--------|
| Model             | 1 to 2 | 2 to 3 | 3 to 4 | 4 to 5 |
| GPT-5.1           | 89.69  | 67.01  | 54.64  | 50.52  |
| GPT-5 mini        | 97.94  | 82.47  | 31.96  | 44.33  |
| Claude Sonnet 4.5 | 73.20  | 57.73  | 54.64  | 14.43  |
| + Thinking        | 85.57  | 24.74  | 57.73  | 18.56  |
| DeepSeek-V3.1     | 63.92  | 59.79  | 49.48  | 44.33  |

Table 4: **Model performance for College → Sr. Res. (Expert-written Inputs, n=98) for multi-prompt approach** To test a multi-prompt approach, we ran an evaluation (n=98) that prompts for each of the levels (College student through Senior researcher) in separate calls (all else the same) and found that it performs worse than the single prompt approach (Tab. 2). Results are shown as the percent of inputs moving in the correct direction at each transition.

We chose the audience labels to range from College student to Senior Researcher because the word-

ing of the questions in ScholarQABench implied that the inquirer had at least a college education. However, as mentioned in Sec. 5.4, 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 Analysis and Information

#### A.2.1 Queries from ScholarQABench

For the first evaluation, we use the queries from ScholarQA-Multi, a subset of ScholarQABench (Asai et al., 2026)<sup>11</sup> that contains 98 expert-written scientific queries and answers across multiple STEM fields. The distribution of queries across fields, along with sample queries, is shown in Table 5.

#### A.2.2 Descriptive Statistics of Boxplots

This section contains descriptive statistics, specifically the means and standard deviations, for the distributions shown in Fig. 2. Tables 6, 7, and 8 are for the “College → Sr. Res., Expert-written Inputs (n=98)”, “College → Sr. Res., Generated Inputs (n=459)”, and “Child → Expert, Expert-written Inputs (n=98)” groups respectively.

#### A.2.3 Significance Testing of Results

As mentioned in Sec. 5, to test if the distributions of changes going in the wrong direction are significantly different from the distributions of changes going in the right direction (Fig. 2), we run Welch’s *t*-tests, applying the Bonferroni correction, comparing those distributions at each of the 4 transitions. That is, at each of the 4 transitions (Level 1 to 2, Level 2 to 3, etc.), we compare whether or not the magnitudes of the changes in the wrong direction are significantly different from the magnitudes of the changes in the right direction. We report the 95% confidence intervals with their significance for each of the evaluations in Tables 9, 10, and 11 respectively. For all models, we find that when there are more than a few inputs for which the model moves the complexity measure towards the incorrect direction, the change often moves significantly in the wrong direction (i.e., when the model incorrectly decreases JARGON, it does so to a significant degree).

<sup>11</sup><https://github.com/AkariAsai/ScholarQABench>

<sup>10</sup><https://www.wired.com/video/series/5-levels/>

### 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

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

Table 5: **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.

### A.3 Model Configuration Details

Tab. 12 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 Additional Measures

#### A.4.1 Flesch-Kincaid Reading Ease Score Data

The Flesch-Kincaid Reading Ease Score is a commonly-used scale for rating complexity (Flesch, 1948; Joshi et al., 2025; Ágústssdóttir 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.

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).

#### A.4.2 Information Density

INFORMATION is one of the three measures used in our evaluations. An alternative to providing the raw number of information points as we do for INFORMATION is to calculate the ratio between information and length. We test such a measure, using the number of sentences to represent length following the idea that a sentence can represent a single information point. The results are detailed in Tab. 13 as the percent of inputs moving in the correct direction at each transition, similar to Tab. 2. We find that model performance according to this measure is worse than according to the other measures, having fewer inputs moving in the correct direction per transition. Additionally, while we divide by the number of sentences, there is no standardized way to calculate information density.

|        | Model             | Jargon           | Info.            | Length               |
|--------|-------------------|------------------|------------------|----------------------|
| 1 to 2 | GPT-5.1           | 6.36 $\pm$ 3.89  | 8.31 $\pm$ 8.56  | 1314.48 $\pm$ 715.12 |
|        | GPT-5 mini        | 6.01 $\pm$ 4.79  | 3.39 $\pm$ 5.55  | 355.53 $\pm$ 181.59  |
|        | Claude Sonnet 4.5 | 15.31 $\pm$ 5.33 | 5.13 $\pm$ 5.63  | 532.77 $\pm$ 206.22  |
|        | + Thinking        | 14.21 $\pm$ 5.65 | 4.73 $\pm$ 5.64  | 455.89 $\pm$ 185.6   |
|        | DeepSeek-V3.1     | 13.2 $\pm$ 5.5   | 3.06 $\pm$ 3.62  | 311.71 $\pm$ 149.61  |
| 2 to 3 | GPT-5.1           | 0.21 $\pm$ 2.55  | 9.02 $\pm$ 8.97  | 1280.07 $\pm$ 552.53 |
|        | GPT-5 mini        | 0.63 $\pm$ 3.45  | 3.98 $\pm$ 6.29  | 404.54 $\pm$ 269.97  |
|        | Claude Sonnet 4.5 | 2.79 $\pm$ 3.27  | 5.88 $\pm$ 5.9   | 515.05 $\pm$ 267.77  |
|        | + Thinking        | 3.53 $\pm$ 3.0   | 4.32 $\pm$ 5.92  | 459.96 $\pm$ 237.09  |
|        | DeepSeek-V3.1     | 1.91 $\pm$ 3.4   | 3.57 $\pm$ 4.85  | 272.45 $\pm$ 158.17  |
| 3 to 4 | GPT-5.1           | 0.00 $\pm$ 1.91  | 4.78 $\pm$ 10.46 | 1015.32 $\pm$ 522.71 |
|        | GPT-5 mini        | 1.16 $\pm$ 3.07  | 2.04 $\pm$ 6.57  | 177.31 $\pm$ 243.21  |
|        | Claude Sonnet 4.5 | 1.69 $\pm$ 2.77  | 5.88 $\pm$ 7.41  | 681.84 $\pm$ 340.44  |
|        | + Thinking        | 1.72 $\pm$ 2.7   | 5.69 $\pm$ 7.59  | 547.86 $\pm$ 299.72  |
|        | DeepSeek-V3.1     | 0.44 $\pm$ 3.82  | 1.48 $\pm$ 4.83  | 221.73 $\pm$ 121.54  |
| 4 to 5 | GPT-5.1           | -0.73 $\pm$ 1.62 | 6.77 $\pm$ 9.88  | 1228.54 $\pm$ 602.67 |
|        | GPT-5 mini        | 0.94 $\pm$ 3.03  | 3.79 $\pm$ 6.56  | 509.57 $\pm$ 405.4   |
|        | Claude Sonnet 4.5 | 0.55 $\pm$ 2.27  | 8.49 $\pm$ 8.96  | 1098.28 $\pm$ 703.87 |
|        | + Thinking        | 1.61 $\pm$ 2.49  | 5.14 $\pm$ 7.66  | 661.05 $\pm$ 722.9   |
|        | DeepSeek-V3.1     | 0.68 $\pm$ 2.86  | 3.3 $\pm$ 4.5    | 351.92 $\pm$ 164.12  |

Table 6: **Model performance per transition for College  $\rightarrow$  Sr. Res. (Expert-written Inputs, n=98)** Mean and standard deviation are shown.

## A.5 User Study Details and Materials

### A.5.1 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. 8. The task questions are displayed in Fig. 7; participants were also provided with a note-taking box in the same document. After the study, participants received a \$25 gift card, which is above the minimum wage policy in the area, and were provided with the disclaimer shown in Fig. 9. This study was approved by our institution’s IRB.

### A.5.2 User Study Interface

We extended a chat interface from the Streamlit framework<sup>12</sup> to build our conventional and interactive interfaces. More details are provided in Sec. 3.1.

<sup>12</sup><https://docs.streamlit.io/develop/tutorials/chat-and-llm-apps/build-conversational-apps>

### A.5.3 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.5.4 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

|        | Model             | Jargon           | Info.             | Length               |
|--------|-------------------|------------------|-------------------|----------------------|
| 1 to 2 | GPT-5.1           | 6.02 $\pm$ 4.21  | 11.27 $\pm$ 12.86 | 1629.2 $\pm$ 853.56  |
|        | GPT-5 mini        | 5.1 $\pm$ 4.92   | 1.62 $\pm$ 6.66   | 208.1 $\pm$ 171.21   |
|        | Claude Sonnet 4.5 | 17.74 $\pm$ 5.67 | 4.58 $\pm$ 6.59   | 457.04 $\pm$ 224.88  |
|        | + Thinking        | 14.56 $\pm$ 5.18 | 5.97 $\pm$ 7.53   | 491.03 $\pm$ 291.95  |
|        | DeepSeek-V3.1     | 16.99 $\pm$ 6.88 | 2.72 $\pm$ 4.5    | 238.97 $\pm$ 115.29  |
| 2 to 3 | GPT-5.1           | -0.27 $\pm$ 2.45 | 10.15 $\pm$ 14.89 | 1664.0 $\pm$ 781.91  |
|        | GPT-5 mini        | 0.24 $\pm$ 3.58  | 3.45 $\pm$ 8.34   | 326.81 $\pm$ 186.53  |
|        | Claude Sonnet 4.5 | 2.1 $\pm$ 3.13   | 5.8 $\pm$ 7.99    | 625.99 $\pm$ 446.66  |
|        | + Thinking        | 2.52 $\pm$ 3.13  | 5.42 $\pm$ 7.71   | 691.07 $\pm$ 488.99  |
|        | DeepSeek-V3.1     | 1.45 $\pm$ 4.59  | 2.03 $\pm$ 4.85   | 211.61 $\pm$ 126.37  |
| 3 to 4 | GPT-5.1           | -0.64 $\pm$ 1.8  | 5.27 $\pm$ 15.2   | 1195.75 $\pm$ 605.11 |
|        | GPT-5 mini        | 0.09 $\pm$ 3.64  | -0.14 $\pm$ 8.83  | 120.39 $\pm$ 189.7   |
|        | Claude Sonnet 4.5 | 0.34 $\pm$ 2.54  | 8.54 $\pm$ 10.3   | 900.51 $\pm$ 648.02  |
|        | + Thinking        | 0.98 $\pm$ 2.6   | 6.77 $\pm$ 10.06  | 847.88 $\pm$ 664.22  |
|        | DeepSeek-V3.1     | -0.28 $\pm$ 4.2  | 1.45 $\pm$ 5.26   | 214.11 $\pm$ 123.19  |
| 4 to 5 | GPT-5.1           | -1.08 $\pm$ 1.56 | 5.75 $\pm$ 14.88  | 1407.18 $\pm$ 797.09 |
|        | GPT-5 mini        | -0.16 $\pm$ 3.45 | 3.79 $\pm$ 8.55   | 567.35 $\pm$ 349.88  |
|        | Claude Sonnet 4.5 | -0.09 $\pm$ 2.31 | 9.92 $\pm$ 13.48  | 1399.97 $\pm$ 976.72 |
|        | + Thinking        | 0.98 $\pm$ 2.44  | 4.7 $\pm$ 12.34   | 990.22 $\pm$ 1165.34 |
|        | DeepSeek-V3.1     | -0.22 $\pm$ 3.48 | 2.52 $\pm$ 6.15   | 285.56 $\pm$ 170.85  |

Table 7: **Model performance per transition for College  $\rightarrow$  Sr. Res. (Generated Inputs, n=459)** Mean and standard deviation are shown.

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.5.5 Participant Details

We recruited participants primarily through email listservs and Slack channels at multiple institutions. Tab. 14 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. \(2025a\)](#).

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)

|        | Model             | Jargon           | Info.             | Length               |
|--------|-------------------|------------------|-------------------|----------------------|
| 1 to 2 | GPT-5.1           | 15.94 $\pm$ 5.75 | 7.09 $\pm$ 6.67   | 719.06 $\pm$ 367.96  |
|        | GPT-5 mini        | 17.44 $\pm$ 7.19 | 4.81 $\pm$ 4.84   | 337.99 $\pm$ 161.3   |
|        | Claude Sonnet 4.5 | 19.8 $\pm$ 6.31  | 5.57 $\pm$ 4.56   | 451.6 $\pm$ 134.68   |
|        | + Thinking        | 19.01 $\pm$ 6.45 | 6.4 $\pm$ 5.24    | 463.13 $\pm$ 154.94  |
|        | DeepSeek-V3.1     | 14.26 $\pm$ 6.82 | 3.64 $\pm$ 3.38   | 223.52 $\pm$ 99.24   |
| 2 to 3 | GPT-5.1           | 6.43 $\pm$ 3.99  | 8.48 $\pm$ 6.39   | 1081.16 $\pm$ 509.61 |
|        | GPT-5 mini        | 2.85 $\pm$ 3.81  | 4.05 $\pm$ 4.96   | 410.45 $\pm$ 194.22  |
|        | Claude Sonnet 4.5 | 11.09 $\pm$ 5.52 | 5.15 $\pm$ 5.71   | 511.41 $\pm$ 210.74  |
|        | + Thinking        | 10.73 $\pm$ 5.09 | 4.37 $\pm$ 5.65   | 529.91 $\pm$ 236.12  |
|        | DeepSeek-V3.1     | 16.23 $\pm$ 6.13 | 3.36 $\pm$ 3.49   | 340.62 $\pm$ 139.68  |
| 3 to 4 | GPT-5.1           | 0.43 $\pm$ 2.72  | 8.83 $\pm$ 9.52   | 1371.26 $\pm$ 572.88 |
|        | GPT-5 mini        | 1.5 $\pm$ 3.43   | 4.1 $\pm$ 5.76    | 400.8 $\pm$ 185.05   |
|        | Claude Sonnet 4.5 | 3.47 $\pm$ 3.4   | 5.65 $\pm$ 5.57   | 625.2 $\pm$ 316.5    |
|        | + Thinking        | 4.28 $\pm$ 3.15  | 6.84 $\pm$ 6.7    | 678.54 $\pm$ 372.7   |
|        | DeepSeek-V3.1     | 3.06 $\pm$ 3.52  | 2.94 $\pm$ 4.36   | 333.94 $\pm$ 146.99  |
| 4 to 5 | GPT-5.1           | -0.79 $\pm$ 1.91 | 12.46 $\pm$ 11.68 | 2226.66 $\pm$ 816.5  |
|        | GPT-5 mini        | 1.63 $\pm$ 2.91  | 5.19 $\pm$ 6.64   | 669.51 $\pm$ 359.69  |
|        | Claude Sonnet 4.5 | 2.2 $\pm$ 2.8    | 7.86 $\pm$ 7.53   | 829.68 $\pm$ 515.69  |
|        | + Thinking        | 3.44 $\pm$ 2.7   | 5.51 $\pm$ 7.87   | 686.22 $\pm$ 527.81  |
|        | DeepSeek-V3.1     | 1.13 $\pm$ 3.4   | 4.28 $\pm$ 4.58   | 378.29 $\pm$ 144.78  |

Table 8: **Model performance per transition for Child  $\rightarrow$  Expert (Expert-written Inputs, n=98)** Mean and standard deviation are shown.

- No
- Other:

These answer choices mostly came from [Liao et al. \(2025\)](#).

### A.5.6 Qualitative Coding Procedure

As mentioned in Sec. 3.1, to analyze the qualitative data from the study transcripts, we employ open coding ([Saldaña, 2021](#)). We follow work where multiple authors create a codebook from sampled interviews and one author codes the remaining interviews with the codebook ([Kim et al., 2023](#); [McDonald et al., 2019](#)). Specifically, 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 (Tab. 15). Using the final codebook, the same initial author coded the remaining studies.

## A.6 Validity checks

### A.6.1 Manual validation of Information measure

We used the prompt provided by [Min et al. \(2023\)](#) for identifying what they call “atomic facts” (defined as “pieces of information”). Since we used a more recent model (GPT-4.1 instead of Instruct-GPT), we manually validated a randomly-sampled subset of 25 responses according to the following procedure and achieved 80% agreement between the Information measure and the human rater:

1. One author manually identified all atomic facts following the procedure and definition from [Min et al. \(2023\)](#).
2. The number of manually-identified facts was compared to the number of atomic facts that GPT-4.1 identified.
3. If the GPT-4.1 number was  $\pm 2$  of the author’s number, it was considered consistent.

### A.6.2 Factuality

We run an evaluation of factual correctness using SummaC ([Laban et al., 2022](#)) to compare the model responses to the human-written reports that were used. SummaC is a popular NLI-based measure of how much a summary follows from a source; higher scores indicate that more of the information in the original report is retained and correct. We find that the responses have moderate scores, consistent across levels within the same model (Tab. 16). This suggests, unsurprisingly, that information is lost in text simplification and to a similar degree across levels.

### A.6.3 Consistency between runs

To check if the performance across multiple model calls is consistent, we run the College  $\rightarrow$  Sr. Res. (n=98) evaluation 5 times using Claude Sonnet 4.5 (one of our better-performing models). We find that the performance is consistent across the 5 runs (Tab. 17).

| <b>Jargon</b>     |                |               |               |               |
|-------------------|----------------|---------------|---------------|---------------|
| Model             | 1 to 2         | 2 to 3        | 3 to 4        | 4 to 5        |
| GPT-5.1           | [7.67,12.36]** | [3.58,4.77]** | [2.57,3.53]** | [1.77,2.84]** |
| GPT-5 mini        | [8.2,11.16]**  | [4.75,6.4]**  | [3.99,5.57]** | [3.69,5.35]** |
| Claude Sonnet 4.5 | n/a            | [4.16,6.03]** | [4.25,5.55]** | [3.14,4.21]** |
| + Thinking        | n/a            | [4.58,6.54]** | [3.86,5.58]** | [3.66,5.06]** |
| DeepSeek-V3.1     | n/a            | [5.12,6.62]** | [5.05,6.96]** | [4.01,5.41]** |

| <b>Information</b> |                 |                 |                 |                 |
|--------------------|-----------------|-----------------|-----------------|-----------------|
| GPT-5.1            | [12.21,16.03]** | [13.19,17.18]** | [14.05,20.62]** | [13.2,18.78]**  |
| GPT-5 mini         | [7.55,10.39]**  | [9.07,12.79]**  | [8.66,11.9]**   | [9.63,12.91]**  |
| Claude Sonnet 4.5  | [7.77,11.23]**  | [8.41,11.27]**  | [10.01,13.92]** | [12.72,19.38]** |
| + Thinking         | [7.63,10.91]**  | [8.47,12.81]**  | [9.67,13.36]**  | [9.87,14.47]**  |
| DeepSeek-V3.1      | [4.62,6.44]**   | [6.63,9.08]**   | [5.17,9.09]**   | [5.75,7.95]**   |

| <b>Length</b>     |     |                  |                  |                     |
|-------------------|-----|------------------|------------------|---------------------|
| GPT-5.1           | n/a | n/a              | n/a              | n/a                 |
| GPT-5 mini        | n/a | [325.31,608.84]* | [311.7,451.77]** | [227.28,1647.9]     |
| Claude Sonnet 4.5 | n/a | n/a              | n/a              | n/a                 |
| + Thinking        | n/a | n/a              | n/a              | [1078.36,1734.76]** |
| DeepSeek-V3.1     | n/a | n/a              | n/a              | n/a                 |

Table 9: **Significance of difference between incorrect and correct changes in measures for College → Sr. Res. (Expert-written Inputs, n=98) evaluation** 95% confidence intervals of the difference between the means are shown. \*\* indicates  $p < 0.01$ , \* indicates  $p < 0.05$ , n/a means that no or few inputs went in the wrong direction.

| <b>Jargon</b>     |                |               |               |               |
|-------------------|----------------|---------------|---------------|---------------|
| Model             | 1 to 2         | 2 to 3        | 3 to 4        | 4 to 5        |
| GPT-5.1           | [7.8,9.68]**   | [3.4,4.02]**  | [2.54,2.97]** | [2.24,2.6]**  |
| GPT-5 mini        | [8.22,9.67]**  | [5.26,6.06]** | [5.46,6.25]** | [5.03,5.82]** |
| Claude Sonnet 4.5 | n/a            | [4.89,5.66]** | [3.79,4.36]** | [3.34,3.89]** |
| + Thinking        | n/a            | [4.95,5.72]** | [3.89,4.5]**  | [3.55,4.17]** |
| DeepSeek-V3.1     | [-13.56,53.79] | [6.79,7.84]** | [6.26,7.19]** | [5.15,5.93]** |

| <b>Information</b> |                 |                 |                 |                 |
|--------------------|-----------------|-----------------|-----------------|-----------------|
| GPT-5.1            | [18.18,22.45]** | [20.71,25.76]** | [20.74,24.97]** | [20.95,25.42]** |
| GPT-5 mini         | [9.35,11.0]**   | [11.38,13.53]** | [11.88,14.07]** | [12.55,14.7]**  |
| Claude Sonnet 4.5  | [9.51,11.14]**  | [11.2,13.27]**  | [14.74,17.48]** | [17.95,21.33]** |
| + Thinking         | [11.19,13.04]** | [11.3,13.3]**   | [13.99,16.55]** | [17.45,20.66]** |
| DeepSeek-V3.1      | [6.43,7.51]**   | [6.9,8.05]**    | [7.83,9.03]**   | [8.28,9.82]**   |

| <b>Length</b>     |                   |                  |                   |                     |
|-------------------|-------------------|------------------|-------------------|---------------------|
| GPT-5.1           | n/a               | n/a              | n/a               | [1416.27,1572.11]** |
| GPT-5 mini        | [261.89,310.97]** | [295.02,489.73]* | [272.72,325.22]** | [513.46,977.82]**   |
| Claude Sonnet 4.5 | n/a               | n/a              | n/a               | n/a                 |
| + Thinking        | n/a               | n/a              | [944.9,1332.41]** | [1648.57,1946.75]** |
| DeepSeek-V3.1     | [229.77,273.36]** | n/a              | [213.65,248.93]** | [219.26,400.55]     |

Table 10: **Significance of difference between incorrect and correct changes in measures for College → Sr. Res. (Generated Inputs, n=459) evaluation** 95% confidence intervals of the difference between the means are shown. \*\* indicates  $p < 0.01$ , \* indicates  $p < 0.05$ , n/a means that no or few inputs went in the wrong direction.

| <b>Jargon</b>      |                 |                 |                 |                    |
|--------------------|-----------------|-----------------|-----------------|--------------------|
| Model              | 1 to 2          | 2 to 3          | 3 to 4          | 4 to 5             |
| GPT-5.1            | n/a             | [6.55,8.54]**   | [3.59,5.17]**   | [2.7,3.59]**       |
| GPT-5 mini         | n/a             | [5.69,7.49]**   | [4.58,6.33]**   | [3.8,5.3]**        |
| Claude Sonnet 4.5  | n/a             | [9.59,15.01]*   | [5.08,6.7]**    | [4.26,5.73]**      |
| + Thinking         | n/a             | [10.21,12.76]** | [5.19,6.55]**   | [4.11,6.03]**      |
| DeepSeek-V3.1      | [3.89,30.38]    | n/a             | [5.15,6.96]**   | [4.66,6.33]**      |
| <b>Information</b> |                 |                 |                 |                    |
| GPT-5.1            | [9.59,13.23]**  | [8.63,15.48]**  | [12.76,17.71]** | [16.53,23.04]**    |
| GPT-5 mini         | [6.27,8.77]**   | [7.31,10.06]**  | [6.98,9.94]**   | [9.01,13.65]**     |
| Claude Sonnet 4.5  | [7.18,9.88]**   | [8.26,12.1]**   | [6.86,9.49]**   | [10.06,14.1]**     |
| + Thinking         | [7.19,10.83]**  | [8.13,11.82]**  | [9.34,13.1]**   | [11.58,15.35]**    |
| DeepSeek-V3.1      | [5.03,7.64]**   | [4.96,7.48]**   | [5.46,7.78]**   | [6.34,9.27]**      |
| <b>Length</b>      |                 |                 |                 |                    |
| GPT-5.1            | n/a             | n/a             | n/a             | n/a                |
| GPT-5 mini         | n/a             | n/a             | n/a             | [647.79,913.09]**  |
| Claude Sonnet 4.5  | n/a             | n/a             | n/a             | n/a                |
| + Thinking         | n/a             | n/a             | n/a             | [746.12,1315.55]** |
| DeepSeek-V3.1      | [177.56,322.69] | n/a             | n/a             | n/a                |

Table 11: **Significance of difference between incorrect and correct changes in measures for Child → Expert (Expert-written Inputs, n=98) evaluation** 95% confidence intervals of the difference between the means are shown. \*\* indicates  $p < 0.01$ , \* indicates  $p < 0.05$ , n/a means that no or few inputs went in the wrong direction.

| 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 12: **Model configurations**

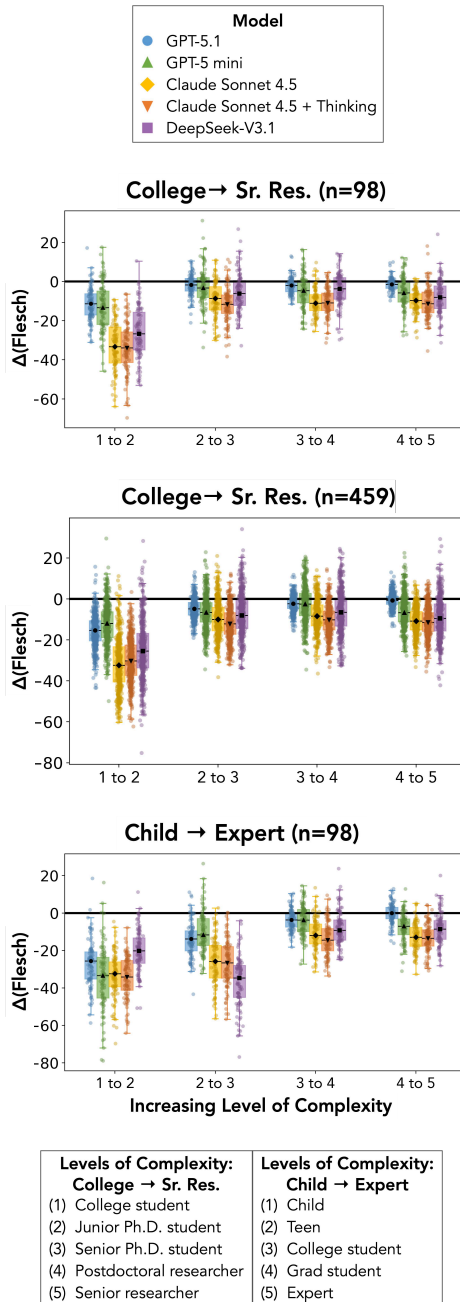


Figure 6: **Flesch-Kincaid Reading Ease score data**  
 These plots show the distributions of changes in the Flesch-Kincaid scores between consecutive levels for the two sets of audience labels that we prompted with. Expert-written inputs were used for  $n = 98$ , and model-generated inputs were used for  $n = 459$ . Since a higher Flesch-Kincaid score means that the text is more readable, *decreasing* the scores as complexity increases is desirable.

| Model             | 1 to 2 | 2 to 3 | 3 to 4 | 4 to 5 |
|-------------------|--------|--------|--------|--------|
| GPT-5.1           | 26.53  | 39.80  | 38.78  | 40.82  |
| GPT-5 mini        | 47.96  | 42.86  | 47.96  | 44.90  |
| Claude Sonnet 4.5 | 66.33  | 61.22  | 56.12  | 67.35  |
| + Thinking        | 68.37  | 63.27  | 66.33  | 63.27  |
| DeepSeek-V3.1     | 57.14  | 63.27  | 56.12  | 56.12  |

Table 13: **Model performance per transition for College → Sr. Res. (Expert-written Inputs, n=98) according to information density** Each model’s performance is shown as the percent of inputs where information density moves in the correct direction at each transition. A higher percentage means that the model performed better at that transition, by more often increasing complexity.

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 7: **Task Questions**

| 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 14: **Participants' background and LLM usage**

| Theme                                                                                                          | Sample Code                                         | Sample Quote                                                                                                                                                                                                                     |
|----------------------------------------------------------------------------------------------------------------|-----------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Complexity is situational                                                                                      | Complexity is situational                           | P3: “I liked the option to see different complexities for a topic because again sometimes I might want more complex, sometimes I might want less complex depending on the task, depending on what I’m actually trying to learn.” |
| Inconsistent satisfaction with the conventional chatbot                                                        | Conventional: inconsistent satisfaction with answer | P13: “sometimes you’re getting the right complexity with some sentences and sometimes you’re not. Yeah...it’s not super consistent in its complexity.”                                                                           |
| Users appreciate the convenience and agency that interactive complexity adds                                   | Appreciates agency of interactive chat              | P9: “I like the amount of choice. I feel like if I don’t use it, I don’t use it, but I like having it there if I want to use it.”                                                                                                |
| Users look for decreasing jargon, information, and length as they decrease the level of complexity             | Want less information when decreasing complexity    | P9: “I want it to be shorter. I think shorter and honestly like less information.”                                                                                                                                               |
| While 5 levels of complexity seemed generally appropriate, the levels need to be more distinct from each other | Levels are too similar                              | P15: “I didn’t notice like a meaningful enough jump between each level”                                                                                                                                                          |

Table 15: **Codebook for analyzing our formative study** Shown is a list of themes with sample codes and quotes. Some codes match the themes as they formed themes on their own.

| Model             | Level 1     | Level 2     | Level 3     | Level 4     | Level 5     |
|-------------------|-------------|-------------|-------------|-------------|-------------|
| GPT-5.1           | 0.46 ± 0.13 | 0.45 ± 0.12 | 0.46 ± 0.13 | 0.46 ± 0.15 | 0.45 ± 0.14 |
| GPT-5 mini        | 0.44 ± 0.13 | 0.41 ± 0.11 | 0.4 ± 0.11  | 0.39 ± 0.11 | 0.39 ± 0.12 |
| Claude Sonnet 4.5 | 0.45 ± 0.12 | 0.45 ± 0.13 | 0.44 ± 0.14 | 0.41 ± 0.13 | 0.39 ± 0.13 |
| + Thinking        | 0.46 ± 0.11 | 0.47 ± 0.13 | 0.42 ± 0.12 | 0.41 ± 0.12 | 0.4 ± 0.13  |
| DeepSeek-V3.1     | 0.46 ± 0.12 | 0.44 ± 0.12 | 0.41 ± 0.14 | 0.4 ± 0.11  | 0.38 ± 0.11 |

Table 16: **Average factuality scores across models and levels of complexity.** Means and standard deviations of distributions of SummaC scores are shown (using “College → Sr. Res., Expert-written Inputs (n=98)” evaluation).

| Measure     | 1 to 2       | 2 to 3       | 3 to 4       | 4 to 5       |
|-------------|--------------|--------------|--------------|--------------|
| Jargon      | 100 ± 0      | 82.68 ± 2.56 | 66.80 ± 3.88 | 63.09 ± 5.88 |
| Information | 80.41 ± 2.44 | 81.24 ± 2.87 | 81.03 ± 2.66 | 82.68 ± 1.52 |
| Length      | 100 ± 0      | 100 ± 0      | 100 ± 0      | 100 ± 0      |

Table 17: **Consistency in performance between separate model calls.** Means and standard deviations of the distribution of per-transition model performances, similar to Tab. 2, across 5 calls to Claude Sonnet 4.5 are shown (using “College → Sr. Res., Expert-written Inputs (n=98)” evaluation).

### 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 8: 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 9: Post-Study Disclaimer

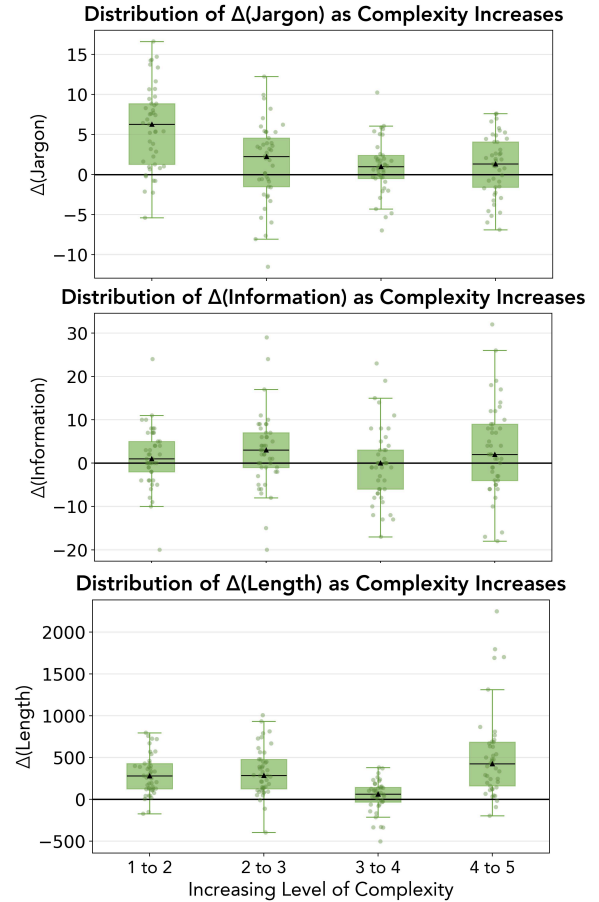


Figure 10: Model performance during user study based on complexity measures Between consecutive levels of complexity, the interactive condition had comparable changes in JARGON, INFORMATION, and LENGTH to the model evaluation (Fig. 2), varying between increasing and decreasing. We use GPT-5 mini to supply 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.