

# ELI-WHY: Evaluating the Pedagogical Utility of Language Model Explanations

Brihi Joshi<sup>♥</sup> Keyu He<sup>♥</sup> Sahana Ramnath<sup>♥</sup> Sadra Sabouri<sup>♥</sup> Kaitlyn Zhou<sup>♦</sup>  
Souti Chattopadhyay<sup>♥</sup> Swabha Swayamdipta<sup>♥</sup> Xiang Ren<sup>♥</sup>

<sup>♥</sup>University of Southern California <sup>♦</sup>Stanford University  
{brihijos, frankhe}@usc.edu

[inklab.usc.edu/eli-why](https://inklab.usc.edu/eli-why)

## Abstract

Language models today are widely used in education, yet their ability to tailor responses for learners with varied informational needs and knowledge backgrounds remains under-explored. To this end, we introduce ELI-WHY, a benchmark of 13.4K “Why” questions to evaluate the pedagogical capabilities of language models. We then conduct two extensive human studies to assess the utility of language model-generated explanatory answers (explanations) on our benchmark, tailored to three distinct educational grades: elementary, high-school and graduate school. In our first study, human raters assume the role of an “educator” to assess model explanations’ fit to different educational grades. We find that GPT-4-generated explanations match their intended educational background only 50% of the time, compared to 79% for lay human-curated explanations. In our second study, human raters assume the role of a learner to assess if an explanation fits their own informational needs. Across all educational backgrounds, users deemed GPT-4-generated explanations 20% less suited on average to their informational needs, when compared to explanations curated by lay people. Additionally, automated evaluation metrics reveal that explanations generated for different informational needs remain indistinguishable in their grade level, for different language model families, limiting their pedagogical effectiveness.

## 1 Introduction

Language models are increasingly used in education to seek information (Suri et al., 2024), tutoring (Chevalier et al., 2024), and automated assessment (Tili et al., 2023; Stahl et al., 2024). A critical aspect of their pedagogical utility is their potential to *tailor* responses to learners with varying informational needs (Adolphe et al., 2023; Puech et al., 2024; Davies et al., 2021; Chevalier et al., 2024;

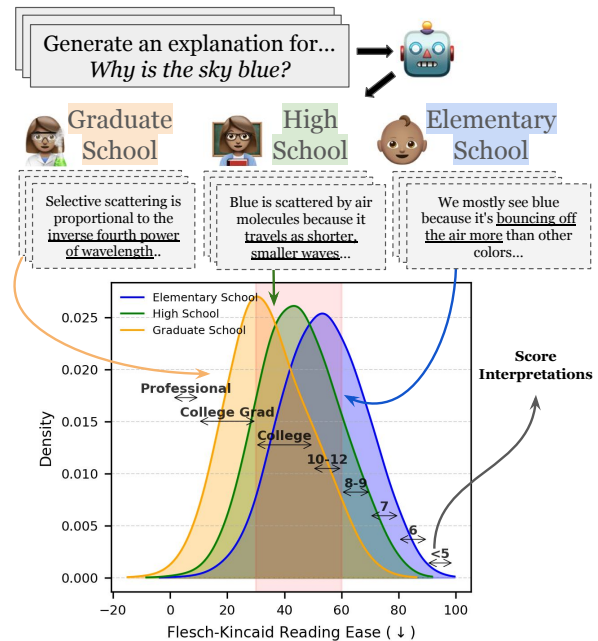


Figure 1: **Distribution of Flesch-Kincaid Reading Ease scores for tailored GPT-4-explanations in ELI-WHY:** Explanations are generated for elementary, high school, and graduate-level backgrounds for “Why” questions. Interpretations of Flesch-Kincaid Reading Ease scores depicting grade-level complexity (lower = more complex) often overlap (within the high school-college range, region highlighted in pink).

Jurenka et al., 2024; Sun and Zhou, 2024; Ross and Andreas, 2024). This is particularly important in scientific communication, where complex concepts must be conveyed effectively to nonexperts (August et al., 2023), and in policy or legal communication, where text must balance technical accuracy with readability (Cheong et al., 2024). Despite the potential of language models to modify explanations in their complexity (August et al., 2024), formality (Luo et al., 2023), and domain specificity (Karabacak and Margetis, 2023; Wang et al., 2023) at inference time, it remains unclear whether they can effectively generate responses that are useful both to educators (Kim et al., 2024a)

\* Equal contribution.

and to learners alike (Lee et al., 2023).

One critical challenge in pedagogy is answering “Why” questions. These require explanatory answers to meet different learners where they are. For example, for the question “*Why is the sky blue?*”, a high school student might find the explanation “*Sunlight scatters when it hits air molecules*” more understandable, while a physics graduate might find a more technical answer “*Selective scattering is proportional to the inverse fourth power of wavelength*” more satisfactory. Although language models are capable of step-by-step reasoning across various tasks (Wei et al., 2022; Prystawski et al., 2023), by default they generate a *one-size-fits-all* explanation, that might not fit the informational needs of a user interacting with it (August et al., 2024). Can the prompt-following skills of language models (Wei et al., 2021; Zeng et al., 2023; Lee et al., 2024) help them tailor their explanations<sup>1</sup> to users with different informational needs?

We introduce ELI-WHY<sup>2</sup>, a dataset of 13.4K “Why” questions that span different disciplines such as science, medicine, and humanities, such as “*Why do countries have flags?*” or “*Why do leaves change color in the fall?*” to examine the pedagogical utility of language model explanations. While prior studies have explored the ability of language models to generate general-purpose explanations in a pedagogical setting (Joshi et al., 2023; Li et al., 2024), it is important that explanations adapt to the prior knowledge of learners (Schmucker et al., 2024; Ye et al., 2024; Lee et al., 2023).

Our experimental settings involve using the *highest educational degree attained* as a proxy for the informational needs of a user. Specifically, we prompt language models to generate three different explanations to ELI-WHY questions, fit for users with elementary school, high school, or graduate level education. We conduct automated evaluations and two human studies to assess the utility of language model generated grade-tailored explanations. Our first human study is conducted from the perspective of an educator to test the appropriateness of an explanation for users with different educational backgrounds on a subset of ELI-WHY. We find that GPT-4-generated explanations match their intended background only 50% of the time,

compared to 79% for explanations curated by lay humans (§4). We then use automated metrics to assess the grade-level readability of explanations; while explanations become more lengthy and contain ‘complex’ words as the educational level increases, their complexity in terms of grade-level readability often overlaps (shown in Figure 1 and Section 4.2). We extend this automated metric analysis to three more model families apart from GPT-4, and report similar findings. Our second human study tests the appropriateness of an explanation from the perspective of a learner’s own self-reported informational needs (Section 5). To capture the information needs of users, we asked participants to rate the explanations based on whether they provide new information and whether the explanations connect to their prior knowledge. Studies with participants from elementary, high school, and graduate backgrounds (Physics and Psychology) reveal that GPT-4-generated explanations are relatively 20% less informative than explanations curated by lay humans. This gap is particularly pronounced for users with graduate-level and high-school backgrounds.

Overall, our results highlight the limitations of current language model-driven pedagogy and suggest that explicitly prompting for audience adaptation alone might be insufficient. We believe that in addition to ELI-WHY being a valuable resource to evaluate language models’ pedagogical utility, our human-centered evaluation framework can help evaluate personalized agents catered to the informational needs of individual users.

## 2 The ELI-WHY Benchmark

Existing work in pedagogical evaluation of language models has either focused on objective benchmark-driven question-answering tasks (e.g. multiple-choice science-based question answering) (Lu et al., 2022; Mitra et al., 2024; Chang et al., 2025) or subjective use-case driven tasks (e.g. evaluating academic achievements induced by language model assistants) (Höper and Schulte, 2024; Sun and Zhou, 2024). Combining these two, we focus on the task of answering “Why” questions; they ensure a good balance between having a knowledge-seeking setting and having room for subjectivity in the manner in which knowledge is presented (Sulik et al., 2023). To this end, we introduce **ELI-WHY**, which consists of 13,392 “Why” questions curated across STEM and Non-STEM

<sup>1</sup>For brevity, we henceforth refer to free-text explanatory answers to “Why” questions as *explanations*.

<sup>2</sup>Name inspired by the subreddit “Explain Like I’m Five”, where users seek simpler answers to questions. <https://www.reddit.com/r/explainlikeimfive/>

domains. There are 6,217 STEM questions (across disciplines like Physics, Chemistry, Computer Science, Material Engineering etc.), and 7,175 non-STEM questions (across disciplines like Sociology, Law, Culture, History, Public Relations, etc.). Our dataset is created by (1) **over-generating** “Why” questions from GPT-4 via few-shot prompting followed by (2) **extensive filtering** by checking validity of the generated questions. We expand upon these steps below, and provide full details, including prompts, model settings and filtering process about ELI-WHY curation in Appendix B.

### Overgenerating “Why” questions from GPT-4.

We use a set of 50 seed “Why” questions from Sulik et al. (2023) (Table 2) and split them into different disciplines. We use a random subset of these as in-context examples to prompt GPT-4 (Table 3) to generate more questions in a given discipline (Liu et al., 2022). This led to a set of ~30k questions.

**Filtering generated questions.** We then manually deduplicated questions from the set. We additionally removed niche, domain-specific questions (e.g. questions like “*Why is the electron cloud model currently the most accepted atomic model?*”) with the help of crowdworkers. Details about the filtering process can be found in Appendix B.2. This resulted in the final 13,392 questions.

## 3 Generating Explanations for different Educational Backgrounds

Users with varying educational or conceptual backgrounds differ in expectations of answers to their questions (Kolb et al., 2007; Bertrand et al., 2023). Tailoring responses to users with different educational backgrounds is important to improve language models’ use in pedagogy (Adolphe et al., 2023; Puech et al., 2024). In this section, we describe the different educational levels we used for evaluating language model explanations and our methodology for generating grade-tailored explanations.

**Educational backgrounds.** We choose three educational levels with different informational needs<sup>3</sup> for our users<sup>4</sup>: **Elementary School**, **High School** and **Graduate School**, in the

context of education in the United States<sup>5</sup>. **Elementary School** group typically covers content up to U.S. Grade 4, and adults with this education level may have limited theoretical knowledge of individual disciplines. The **High School** group extends through U.S. Grade 12 to approximately the sophomore year of undergraduate studies, and adults at this level have a foundational grasp of academic subjects but may still struggle with discipline-specific terminology. **Graduate School** group typically follow a bachelor’s degree, offering advanced, specialized education and adults with this education have few knowledge gaps and possess expertise in specific areas without needing foundational instruction<sup>6</sup>.

### Generating grade-tailored explanations.

For any given “Why” question, our goal is to generate three responses corresponding to users whose highest educational degree is at the **Elementary School**, **High School** and **Graduate School** level. We generate explanations for each question by zero-shot prompting language models from four model families—GPT-4-0613<sup>7</sup> (henceforth shortened to GPT-4), Llama-3.2-3B-Instruct, Qwen 2.5 14B Instruct and DeepSeek R1 Distill Llama 8B. We instruct each language model to assume the role of an *expert* in order to provide suitable explanations for each of the three educational backgrounds (prompt detailed in Appendix C).<sup>8</sup> Additionally, our prompts contain instructions like “*do not add any additional text like greetings or ornamental words*” to ensure that language models tailors the response in terms of knowledge and not just stylistic cues. For example, GPT-4 would often add context, such as “*playing in the park*” or “*other kids*” while generating explanations for **Elementary School** background. We try to limit such generations using specific instructions in the prompt, so that it puts less emphasis on stylistic verbiage, when compared to knowledge content (Table 6). Throughout the rest of this paper, we use *intended educational background* of an explanation to refer

<sup>5</sup><https://usahello.org/education/children/grade-levels/>

<sup>6</sup>Description of these backgrounds are informed by [https://en.wikipedia.org/wiki/Educational\\_attainment\\_in\\_the\\_United\\_States](https://en.wikipedia.org/wiki/Educational_attainment_in_the_United_States) and Falk and Needham (2013).

<sup>7</sup>Last accessed August 2024.

<sup>8</sup>Language models like GPT-4 contains knowledge from varied disciplines, as per <https://openai.com/index/gpt-4-research/>.

<sup>3</sup>As educational levels can often overlap with informational needs, we choose three levels that are ideally least overlapping.

<sup>4</sup>Throughout this study, we refer to adults aged 18+ as our user base.

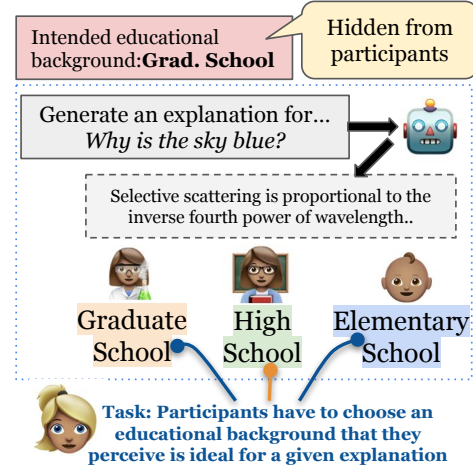
to the educational background used to generate the explanation. All model parameters used to generate explanations are detailed in Appendix C.2. While all four language models are used for automated evaluations, we only use explanations generated by GPT-4 for our human studies.

**Baseline explanations.** In addition to the above explanations, we prompt language models to produce the **Default** explanation for a given question, without providing any educational background (prompt detailed in Appendix C). We also collect baseline **Web-Retrieved** explanations using the Google API<sup>9</sup>; we use the *Featured Snippet* provided by Google.<sup>10</sup>

**Web Explanations Curated by Lay Humans.** Lastly, for a subset of 40 questions in ELI-WHY, authors of this work manually curated explanations ( **Manually Web-Retrieved** ) for each educational background, by searching appropriate websites. All explanations are curated independently by two authors, then discussed together to preserve the most plausible explanation. For e.g., we retrieve **Graduate School** level explanations for a question by searching through journals and research papers on the topic, and **Elementary School** level explanations by searching through the Explain Like I’m Five (ELI5) subreddit<sup>11</sup>. For **High School**, we retrieve explanations from blog posts and web pages intended for lay users. These are not meant to be expert-level explanations, but simulate a process of obtaining explanations for different grade levels in contrast to language model generations (Oh et al., 2008; Ward, 2021).

#### 4 Do language model explanations match their intended educational background?

In this section, we evaluate whether grade-tailored language model explanations match their intended educational backgrounds, using human evaluations. We then extend to a large-scale empirical analysis on all of ELI-WHY and model variants, where we employ different automated metrics and reconcile these findings with that of the user study.



**Figure 2: User study for evaluating perceived background match:** Participants assume the role of an educator and determine the educational background of the explanation presented to them, without knowing the explanation’s intended educational background. In this example, the explanation is generated for **Graduate School**, but participants perceive it to be ideal for **High School** users.

#### 4.1 Intended vs. perceived educational backgrounds of tailored explanations

We define the *intended educational background* as the grade-level for which an explanation was generated. We then define the *perceived educational background* as the grade-level that a *human user* associates with an explanation. To identify if language model explanations are successfully tailored for different grade levels, we conduct a user study in which participants assume the role of an educator; they read questions and a language model explanation to indicate their perceived educational background of the explanation. We then evaluate the percentage of explanations where the intended educational background matches the perceived educational background. We term this as *Perceived Background Match*. This formulation allows us to directly measure whether tailored explanations match the grade level they were generated for.

**User study design.** We conducted a user study with a subset of 400 “Why” questions from ELI-WHY, along with explanations generated by GPT-4 tailored for each of the three grade levels we consider. The participants were presented with a question-explanation pair and were asked to identify the perceived educational background of the explanation (Figure 2). Before making their judgments, the participants received detailed task in-

<sup>9</sup><https://serpapi.com/>

<sup>10</sup>These are pre-Gemini summary results, where the only use of a model had been to rank relevant snippets, according to <https://support.google.com/websearch/answer/9351707?hl=en>.

<sup>11</sup><https://www.reddit.com/r/explainlikeimfive/>



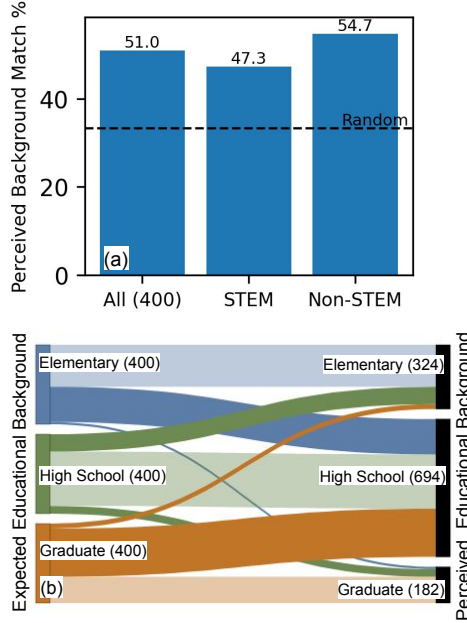


Figure 3: **Perceived background match results:** (a) Perceived background match % vs. domain split of questions. (b) Sankey diagram depicting change between intended background of explanations (left) to perceived background (right) after user study.

structions, including information on different educational backgrounds defined in Section 3. Additionally, pilot evaluations, conducted by the authors and a subset of participants, helped refine instruction clarity. Each participant annotated five question-explanation pairs, and each pair received three independent annotations, ensuring a diverse evaluation of perceived backgrounds; we considered a majority vote of the perceived educational background for all explanations. As a control, we also conducted a user study on **Manually Web-Retrieved** explanations for 40 questions to understand perceived explanation match trends for explanations curated by lay users. Further details on participant screening, demographics, and study setup are provided in Appendix D.

**Results.** Figure 3 presents results from the user study. Figure 3(a) shows that the perceived background match % of tailored explanations generated by GPT-4 is very low (close to 50%). This trend is observed across STEM and Non-STEM splits of the subset. Furthermore, the user study also reveals that tailoring *mismatch* is seen across the board for all educational backgrounds. Figure 3(b) shows the change between intended and perceived educational background, after the study. Most explana-

tions are perceived to be tailored for **High School**, which can be explained by GPT-4’s tendency to be conditioned towards a “lay-user” (August et al., 2024; Hsu et al., 2024). We also observe surprising mismatches—e.g., **Elementary School** explanations being perceived as **Graduate School**, and vice versa. We show examples of these cases, along with justification written by users in Appendix D.4. Additionally, the perceived role match of **Manually Web-Retrieved** explanations is much higher (79.16%). This reveals a concerning trend in GPT-4’s explanations: while GPT-4 can be easily prompted to generate explanations tailored for different educational backgrounds, it does not necessarily mean that users perceive these explanations fit for a given background, potentially hindering GPT-4’s utility in pedagogy (Kasneci et al., 2023).

## 4.2 What do automated metrics reveal about tailored language model explanations?

Section 4.1 demonstrated that GPT-4-generated rationales often mismatch their intended educational backgrounds. We extend the scale of our analysis to the full ELI-WHY benchmark and more language model families using automated metrics and show that careful interpretation of these metrics also highlight the above mismatch.

**Automated Metrics.** We use three categories of automated metrics, based on surface-form features, readability, and reasoning styles to evaluate whether these automated metrics distinguish between explanations tailored to different grades. Surface form metrics compute sentence count, average sentence length, estimated reading time (Demberg and Keller, 2008), and TE Score (August et al., 2024) (the TE score / Thing Explainer Out-of-Vocabulary score measures the proportion of ‘complex words’ in an explanation by taking the proportion of words outside a curated list of the 2,000 most common English words). We employ three popular readability metrics: Flesch-Kincaid Reading Ease (Flesch, 1948), Linsear Write Formula (O’hayre, 1966), and Dale-Chall Readability Score (Dale and Chall, 1948). Each of these metrics also map score ranges to an interpreted U.S. grade level (Kincaid et al., 1975). Score range mappings for each metric are detailed in Table 11. Finally, we analyze the type of reasoning in the explanations: whether they are mechanistic (describe *how* a phenomenon occurs, e.g. *pollen shedding occurs because of desiccation of anther tips*) vs.

Type	Metric	Language Model	Elementary School	High School	Graduate School	Default	Web-Retrieved
Surface-form	# Sentences	GPT-4	04.63 $\pm$ 01.34	07.08 $\pm$ 02.53	08.46 $\pm$ 02.62	05.07 $\pm$ 01.63	02.30 $\pm$ 00.90
		Llama-3.2-3B-Instruct	03.29 $\pm$ 01.63	06.70 $\pm$ 02.97	09.10 $\pm$ 03.33	04.24 $\pm$ 02.63	
		Qwen 2.5 14B Instruct	03.38 $\pm$ 00.86	03.77 $\pm$ 00.93	04.50 $\pm$ 01.17	02.97 $\pm$ 00.87	
		DeepSeek R1 Distill LLama 8B	04.45 $\pm$ 02.60	05.30 $\pm$ 02.88	06.50 $\pm$ 03.33	04.78 $\pm$ 02.93	
	Avg. # Words / Sentence	GPT-4	18.43 $\pm$ 03.47	19.17 $\pm$ 03.36	20.00 $\pm$ 03.38	19.35 $\pm$ 03.57	17.26 $\pm$ 06.90
		Llama-3.2-3B-Instruct	20.39 $\pm$ 05.10	21.30 $\pm$ 03.81	23.12 $\pm$ 03.69	23.74 $\pm$ 04.88	
		Qwen 2.5 14B Instruct	18.50 $\pm$ 03.75	19.92 $\pm$ 03.74	21.73 $\pm$ 03.96	21.54 $\pm$ 04.49	
		DeepSeek R1 Distill LLama 8B	20.22 $\pm$ 04.40	19.89 $\pm$ 03.97	20.03 $\pm$ 03.69	20.47 $\pm$ 04.42	
	Avg. Reading Time (s)*	GPT-4	06.36 $\pm$ 01.75	10.57 $\pm$ 03.65	13.93 $\pm$ 04.05	07.81 $\pm$ 02.41	02.93 $\pm$ 01.04
		Llama-3.2-3B-Instruct	04.61 $\pm$ 02.14	10.97 $\pm$ 05.00	17.05 $\pm$ 06.30	07.93 $\pm$ 04.71	
		Qwen 2.5 14B Instruct	04.60 $\pm$ 01.17	05.93 $\pm$ 01.41	08.08 $\pm$ 01.98	05.19 $\pm$ 01.40	
		DeepSeek R1 Distill LLama 8B	07.37 $\pm$ 04.38	08.77 $\pm$ 04.97	11.22 $\pm$ 06.10	08.15 $\pm$ 05.08	
	TE Score*	GPT-4	00.43 $\pm$ 00.09	00.49 $\pm$ 00.09	00.55 $\pm$ 00.09	00.50 $\pm$ 00.09	00.44 $\pm$ 00.12
		Llama-3.2-3B-Instruct	00.37 $\pm$ 00.11	00.47 $\pm$ 00.09	00.54 $\pm$ 00.10	00.53 $\pm$ 00.11	
		Qwen 2.5 14B Instruct	00.38 $\pm$ 00.10	00.47 $\pm$ 00.09	00.53 $\pm$ 00.10	00.51 $\pm$ 00.10	
		DeepSeek R1 Distill LLama 8B	00.58 $\pm$ 00.14	00.60 $\pm$ 00.14	00.66 $\pm$ 00.16	00.62 $\pm$ 00.14	
Readability	Flesch-Kincaid Reading Ease* ( $\downarrow$ )	GPT-4	53.82 $\pm$ 14.52	45.51 $\pm$ 14.26	34.70 $\pm$ 14.43	41.00 $\pm$ 15.73	53.35 $\pm$ 18.52
		Llama-3.2-3B-Instruct	60.91 $\pm$ 17.14	46.39 $\pm$ 15.11	34.56 $\pm$ 14.51	33.68 $\pm$ 16.87	
		Qwen 2.5 14B Instruct	57.59 $\pm$ 15.19	42.38 $\pm$ 16.05	30.67 $\pm$ 16.18	33.79 $\pm$ 17.44	
		DeepSeek R1 Distill LLama 8B	34.40 $\pm$ 16.79	34.59 $\pm$ 15.70	30.98 $\pm$ 15.23	30.94 $\pm$ 16.69	
	Linsear Write Formula ( $\uparrow$ )	GPT-4	11.67 $\pm$ 02.77	12.15 $\pm$ 02.70	13.21 $\pm$ 02.70	13.16 $\pm$ 02.82	11.10 $\pm$ 05.21
		Llama-3.2-3B-Instruct	12.29 $\pm$ 03.94	13.23 $\pm$ 03.09	14.88 $\pm$ 02.93	16.43 $\pm$ 03.93	
		Qwen 2.5 14B Instruct	11.46 $\pm$ 03.02	13.61 $\pm$ 02.95	15.40 $\pm$ 02.94	15.44 $\pm$ 03.49	
		DeepSeek R1 Distill LLama 8B	14.42 $\pm$ 03.70	14.03 $\pm$ 03.37	14.20 $\pm$ 03.17	14.86 $\pm$ 03.79	
	Dale-Chall Readability Score ( $\uparrow$ )	GPT-4	09.31 $\pm$ 01.16	09.84 $\pm$ 01.10	10.55 $\pm$ 01.10	10.35 $\pm$ 01.23	09.94 $\pm$ 01.56
		Llama-3.2-3B-Instruct	08.73 $\pm$ 01.47	09.49 $\pm$ 01.19	09.88 $\pm$ 01.13	10.80 $\pm$ 01.38	
		Qwen 2.5 14B Instruct	09.01 $\pm$ 01.32	10.32 $\pm$ 01.32	10.91 $\pm$ 01.27	11.11 $\pm$ 01.41	
		DeepSeek R1 Distill LLama 8B	11.09 $\pm$ 01.35	11.02 $\pm$ 01.25	11.14 $\pm$ 01.20	11.36 $\pm$ 01.36	
Reasoning	% Mechanistic Reasoning	GPT-4	54.76%	57.28%	63.54%	58.51%	65.16%
		Llama-3.2-3B-Instruct	55.04%	57.81%	65.65%	63.03%	
		Qwen 2.5 14B Instruct	52.75%	57.42%	63.61%	60.53%	
		DeepSeek R1 Distill LLama 8B	56.72%	56.14%	58.48%	57.15%	

Table 1: **Comparison of surface-form, readability, and reasoning-type metrics across different education levels, along with retrieved explanations.** \* represents metrics that have high correlation with user evaluations of perceived educational backgrounds.  $\uparrow$  and  $\downarrow$  depict direction of scores representing more complex explanations for readability metrics; for all the other metrics, higher values indicate higher complexity.

functional (the *purpose* why a phenomenon occurs, e.g. *pollen shedding occurs to facilitate reproduction*) (Sulik et al., 2023). Further details on the calculation of these metrics are in Appendix E.

**Automated metrics reveal that tailored explanations suffer from *interpretation collapse*.** Table 1 presents average and standard deviation of automated metrics for grade-tailored explanations, along with two baseline explanations: **Default** and **Web-Retrieved**. Across all language models, we can observe that the surface form metrics, specifically number of sentences, differ significantly across different educational levels. Particularly, generated explanations get lengthier as the educational level increases. All models also end up using more ‘complex words’ with increasing educational levels, as shown by the increasing TE Score for all models. Additionally, all models end up using more mechanistic reasoning and less teleological reasoning as educational levels increase; prior work has often shown that young children often endorse more teleological explanations (Schachner et al., 2017), also demonstrated

here.

**Default** explanations mimic **High School** explanations in all metrics, indicating that explanations generated by GPT-4 without any grade-level tailoring are often intended for a **High School** user. On the other hand, **Web-Retrieved** explanations are more concise than other explanations, but their complexity varies widely, shown by the high standard deviation for all readability tests. In Appendix E.5, we also compare different grade-tailored explanations with **Default** and **Web-Retrieved** in terms of informational overlap between explanations.

We observe an interesting pattern demonstrated by the readability metrics. Consider the Flesch-Kincaid Reading Ease metric (where a lower score indicates higher grade-level readability of a given text). This is also one of three metrics (among Avg. Reading Time and TE Score) that correlate significantly with user perceived educational levels that we obtain in Section 4.1 (Appendix E.3). For all models except DeepSeek R1 Distill LLama 8B, we observe that the Flesch-Kincaid Reading

Ease metrics are relatively distinct for different educational backgrounds. However, it is interesting to see that these values are so close to each other that they often fall under the same interpreted U.S. grade level. For example, for GPT-4 explanations we show the Flesch-Kincaid Reading Ease distributions for grade-tailored explanations in Figure 1. When these scores are mapped to their **interpreted U.S. grade levels**, the distributions collapse into a narrow range, primarily between high school and college-level readability. We term this as *interpretation collapse*, which is also observed for all language models (Appendix E.4). In fact, for DeepSeek R1 Distill Llama 8B, readability score distributions are almost overlapping for all educational levels. This is supported by our observations in Section 4.1, where participants often perceive most explanations as tailored for **High School**. The fact that explanations meant for vastly different backgrounds fall into overlapping score ranges suggests that grade-tailored explanations are not meaningfully differentiating at an interpretive level, even if surface form qualities like length and complexity of words increases. We suggest that automated metrics (like Flesch-Kincaid Reading Ease) to some extent can also be used to measure whether language model explanations are truly tailored to their intended educational backgrounds, provided they are carefully inspected with their corresponding grade-level interpretations.

#### 4.3 Case Study: Why is there a mismatch between intended and perceived educational backgrounds of tailored explanations?

As seen in Figure 3(b), we observe surprising mismatches while tailoring explanations to different educational backgrounds—particularly where explanations tailored for higher educational levels like **High School** or **Graduate School** are instead perceived as **Elementary School**. We hypothesize that such mismatches arise because of certain questions being always associated with a particular educational background, hindering GPT-4’s (and possibly other language models’) ability to generalize for a different educational background.

As a case study, we look at the ELI5 subreddit, where users often seek simplified explanations for different questions, most of them being “Why” questions (Appendix F). We observe that questions that exhibit perceived simplification—

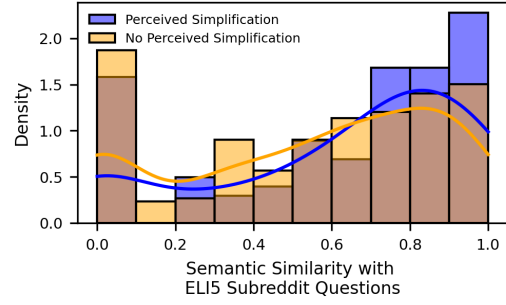


Figure 4: **Relationship between perceived simplification and semantic similarity to ELI5 questions:** Questions where explanations were perceived as significantly simpler than intended (e.g., intended **High School** or **Graduate School** but perceived as **Elementary School**) tend to have higher similarity to questions present the ELI5 subreddit.

GPT-4’s explanations tailored for **High School** and **Graduate School** that were perceived to be **Elementary School** by users—are significantly more similar to questions in the ELI5 subreddit than other questions ( $p < 0.05$ , Mann-Whitney U Test (Mann and Whitney, 1947)). This suggests that GPT-4 may overgeneralize and produce simpler explanations when a question closely resembles those always present in contexts pertaining to these educational backgrounds (Figure 4).

### 5 Do generated explanations help provide new information to users?

A fundamental notion of utility for language models in pedagogical cases is how much they assist users in learning *new* information (Joshi et al., 2023; Zhang et al., 2024; Schmucker et al., 2024; Lee et al., 2023). In this section, we discuss the utility of explanations in delivering new information to a *learner*, that aligns with the learner’s informational needs. Understanding this is crucial in determining whether language models like GPT-4 tailor explanations for different educational backgrounds merely stylistically or if they provide *new* and *relevant* information that contributes to learning and comprehension.

**Evaluating informativeness w.r.t user informational needs.** Consider a user with a high school-level background in physics, familiar with basic concepts about light such as scattering, wave-particle duality, and light interactions. Given a question, “*Why is the sky blue?*”, the user receives the following explanation: “*Because of a*

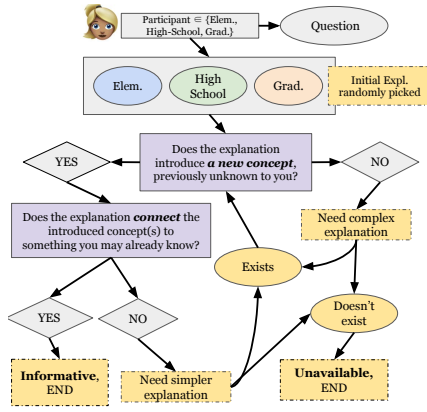


Figure 5: **User evaluation of explanation informativeness:** Participants are provided with a randomly selected explanation (from one of the educational backgrounds) for a given question. They then assume the role of a learner and determine if an explanation provides new information that connects with their information needs.

*solar zenith angle (SZA) of 90°, only 1/3 of the blue color of the sky at the zenith is caused by Rayleigh scattering.*” While this explanation introduces new terms like *solar zenith angle*, it fails to properly define them, making it difficult for the user to integrate the explanation into their existing knowledge. Conversely, an overly simplistic explanation such as “*When sunlight comes through the air bubble that surrounds the Earth, it sometimes hits little bits of air and gets scattered*” provides no meaningful new insights and is therefore uninformative.

We define that an explanation is *informative* in for a user if it satisfies two conditions: (1) it introduces *new concepts* that the user was previously unaware of, and (2) these new concepts *connect well* with the user’s existing background knowledge, making them easier to understand. We design the following user study to evaluate the informativeness of an explanation for a given user. We recruit users belonging to a specific educational background. The user is presented with a question and a randomly selected stimuli explanation, that could belong to **Elementary School**, **High School** or **Graduate School** backgrounds with equal probability. The user is then asked: “*Does the explanation introduce a new concept, previously unknown to you?*” If the user responds negatively, this implies that the explanation is too simple for them, so the system provides an explanation from the next higher educational background. If the user responds positively, they are asked a follow-up question: “*Does the explanation connect the introduced concept(s) to something you may already know?*” If

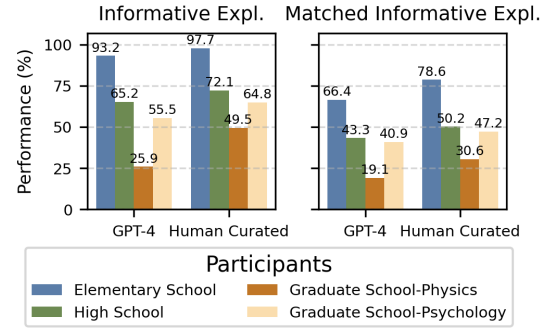


Figure 6: **Comparison of % Informative Explanations and % Matched Informative Explanations for across different educational backgrounds.** GPT-4 grade-tailored explanations are often informative for **Elementary School** participants; they struggle to align with the needs of **High School** and **Graduate School** participants, whereas **Manually Web-Retrieved** grade-tailored explanations perform consistently better across all participants.

they confirm that the concepts are well-integrated, the explanation is considered *informative*. However, if the new concepts do not align with their prior knowledge, the explanation introduces new information but lacks coherence, making it difficult for the user to integrate into their understanding; in this case, the system provides an explanation from the next lower educational background. Figure 5 summarizes this evaluation.

**Human Study and Metrics.** We recruited adult participants with the following highest education levels: elementary school, high school, and graduate degrees in two distinct disciplines—Physics for STEM and Psychology for Non-STEM. We select 40 questions from ELI-WHY, and derived from Section 4.1, we use GPT-4-generated explanations that were perceived to match their intended educational backgrounds. For each educational background and question, participants assume the role of a learner and determine if GPT-4 generates explanations that are *informative* for a question. Each question is answered by five participants, leading to 200 responses for each educational background. We compute two metrics: % Informative Explanations which is the % of questions where *any* one of the three GPT-4 grade-tailored explanations were found informative, and % Matched Informative Explanations which is the % of questions where explanations were informative and *matched* the participant’s educational background. Given that we aim to capture how useful



grade-tailored explanations are for an individual and that every individual may have different prior knowledge even within the same educational background, we do not do any majority voting while aggregating the above metrics for a question across participants. Lastly, we also replicate the same user study with the 40 ELI-WHY questions with **Manually Web-Retrieved** explanations.

**Results.** Figure 6 shows the % Informative Explanations and % Matched Informative Explanations results for participants with different educational background. We observe that participants with higher education backgrounds have lower % Informative Explanations and % Matched Informative Explanations. It is particularly stark for participants with a **Graduate School**-Physics background, where only 19% of questions have informative explanations that match the participant’s background. We find that on an aggregate basis, **Manually Web-Retrieved** explanations consistently outperform GPT-4 on both metrics across all educational backgrounds. While GPT-4 provides new information at a comparable rate for **Elementary School** and **High School** participants, its effectiveness declines significantly for **Graduate School**-background participants. On average, for all three educational backgrounds, **Manually Web-Retrieved** explanations are relatively 20% more informative than GPT-4 explanations. It is important to note that **Manually Web-Retrieved** explanations are curated by lay experts, not domain experts. These individuals rely on general knowledge, metadata about online resources to craft responses, yet they still provide more informative and better-aligned explanations than GPT-4. This suggests that GPT-4 struggles not just with domain expertise, but also with the broader research and adaptation strategies that even nonexperts employ when tailoring explanations. Recruiting actual subject matter experts could further widen this gap, highlighting limitations in delivering truly audience-appropriate information.

## 6 Conclusion

Our study introduces ELI-WHY, a benchmark for evaluating the pedagogical utility of language model explanations tailored to users belonging to different educational backgrounds. Through user studies, we find that language models like GPT-4 struggle to align explanations with intended ed-

ucational levels and fail to be informative, especially for advanced learners. Automated evaluations across multiple model families confirm that grade-tailored explanations often collapse into a similar and narrow complexity range, hinting at their limited effectiveness. Future work can explore methods for integrating measures of pedagogical utility for both educators and learners, as a signal to improve language models for users and cater to more personal learning goals, not bounded by educational backgrounds.

## Acknowledgments

We are grateful to members of the USC NLP Group, INK Lab, DILL Lab and Aditya Chetan, as well as anonymous reviewers and meta reviewers for their comments and feedback on the draft. This research is supported in part by the National Science Foundation (NSF) under grant IIS2403437, the Simons Foundation, and the Allen Institute for AI, the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. B. Joshi was supported by the Apple Scholars in AI/ML PhD Fellowship. This work was partially done when S. Swayamdipta was a visitor at the Simons Institute for the Theory of Computing.

## Limitations

**Prompting.** Our evaluations rely exclusively on zero-shot prompting to grade-tailor explanations for all language models, without exploring alternative prompting strategies such as retrieval augmentation or fine-tuning, which may improve such tailoring. Lay users of language models often provide prompt-level instructions without additional strategies, which led to our design decision. Additionally, user interactions are often multi-turn in nature, which we haven’t explored in this work.

**Human Evaluations.** Our evaluations are conducted in a controlled setting, where explanations are assessed in isolation rather than within real-world interactive learning contexts. In practice,

learners might seek clarification, ask follow-up questions, or engage in dialogue, which could impact how explanations are understood and used (Sulik et al., 2023; Zhao et al., 2024).

**Benchmark Design.** ELI-WHY questions are generated from GPT-4. While we conduct extensive validation checks, these questions may differ from actual questions that may be asked in pedagogical settings, different both in format or content. Our study categorizes learners into only three broad educational backgrounds (elementary, high school, graduate), whereas real learners exist on a continuum of knowledge levels, with varying prior expertise and learning needs. There might also be potential overlaps in learning needs amongst learners with different educational backgrounds. While our benchmark includes a diverse set of “Why” questions, our human evaluation studies are conducted on a subset of the dataset, as carefully conducted human experiments are very expensive.

## Ethics Statement

The benchmark we introduce, ELI-WHY, will be publicly released along with all model and human curated explanations. All user studies were conducted by participants from the U.S. We designed the task to compensate annotators above minimum wage (\$16.5/hour) and conducted extensive qualification rounds before task participation. Annotators who completed these qualification tasks received additional compensation to account for the time required to familiarize themselves with task instructions. We also maintained direct communication with participants to address queries and concerns. Additionally, we provided performance-based bonuses to annotators who flagged errors or consistently provided high-quality annotations. AI Assistants (Copilot and ChatGPT) are used as assistants in coding tasks.

## References

- Maxime Adolphe, Marion Pech, Masataka Sawayama, Denis Maurel, Alexandra Delmas, Pierre-Yves Oudeyer, and Hélène Sauzéon. 2023. [Exploring the potential of artificial intelligence in individualized cognitive training: a systematic review](#).
- 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, 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).
- Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R Eagan, and Winston Maxwell. 2023. On selective, mutable and dialogic xai: A review of what users say about different types of interactive explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.
- Maureen A Callanan and Lisa M Oakes. 1992. Preschoolers’ questions and parents’ explanations: Causal thinking in everyday activity. *Cognitive development*, 7(2):213–233.
- Serina Chang, Ashton Anderson, and Jake M. Hofman. 2025. [Chatbench: From static benchmarks to human-ai evaluation](#). *Preprint*, arXiv:2504.07114.
- Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023. [REV: Information-theoretic evaluation of free-text rationales](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2007–2030, Toronto, Canada. Association for Computational Linguistics.
- Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. [\(a\)i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice](#). *Preprint*, arXiv:2402.01864.
- Alexis Chevalier, Jiayi Geng, Alexander Wettig, Howard Chen, Sebastian Mizera, Toni Annala, Max Jameson Aragon, Arturo Rodríguez Fanlo, Simon Frieder, Simon Machado, et al. 2024. Language models as science tutors. *arXiv preprint arXiv:2402.11111*.
- Michelle M Chouinard, Paul L Harris, and Michael P Maratsos. 2007. Children’s questions: A mechanism for cognitive development. *Monographs of the society for research in child development*, pages i–129.
- Kathleen H Corriveau and Katelyn E Kurkul. 2014. “why does rain fall?”: Children prefer to learn from an informant who uses noncircular explanations. *Child development*, 85(5):1827–1835.
- Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.
- Huw C Davies, Rebecca Eynon, and Cory Salvesson. 2021. [The mobilisation of ai in education: A bourdieusean field analysis](#). *Sociology*, 55(3):539–560.

- Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.
- John H Falk and Mark D Needham. 2013. Factors contributing to adult knowledge of science and technology. *Journal of Research in Science Teaching*, 50(4):431–452.
- Nils Feldhus, Aliko Anagnostopoulou, Qianli Wang, Milad Alshomary, Henning Wachsmuth, Daniel Sonntag, and Sebastian Möller. 2024. Towards modeling and evaluating instructional explanations in teacher-student dialogues. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pages 225–230.
- Ronald Aylmer Fisher. 1970. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer.
- R Flesch. 1948. A new readability yardstick. *J. Appl. Psychol.*, 32(3):221–233.
- Rudolf Flesch. 1979. How to write plain english. *University of Canterbury*. Available at [http://www.mang.canterbury.ac.nz/writing\\_guide/writing/flesch.shtml](http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml). [Retrieved 5 February 2016].
- Lukas Höper and Carsten Schulte. 2024. New perspectives on the future of computing education: Teaching and learning explanatory models. In *Proceedings of the 24th Koli Calling International Conference on Computing Education Research*, pages 1–8.
- Yi-Sheng Hsu, Nils Feldhus, and Sherzod Hakimov. 2024. Free-text rationale generation under readability level control. *arXiv preprint arXiv:2407.01384*.
- Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. [Are machine rationales \(not\) useful to humans? measuring and improving human utility of free-text rationales](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7103–7128, Toronto, Canada. Association for Computational Linguistics.
- Irina Jurenka, Markus Kunesch, Kevin R. McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, Ankit Anand, Miruna Pislari, Stephanie Chan, Lisa Wang, Jennifer She, Parsa Mahmoudieh, Aliya Rysbek, Wei-Jen Ko, Andreea Huber, Brett Wiltshire, Gal Elidan, Roni Rubinfeld, Jasmin Rubinovitz, Amit Pitaru, Mac McAllister, Julia Wilkowski, David Choi, Roe Engelberg, Lidan Hackmon, Adva Levin, Rachel Griffin, Michael Sears, Filip Bar, Mia Mesar, Mana Jabbar, Arslan Chaudhry, James Cohan, Sridhar Thiagarajan, Nir Levine, Ben Brown, Dilan Gorur, Svetlana Grant, Rachel Hashimshoni, Laura Weidinger, Jieru Hu, Dawn Chen, Kuba Dolecki, Canfer Akbulut, Maxwell Bileschi, Laura Culp, Wen-Xin Dong, Nahema Marchal, Kelsie Van Deman, Hema Bajaj Misra, Michael Duah, Moran Ambar, Avi Caciularu, Sandra Lefdal, Chris Summerfield, James An, Pierre-Alexandre Kamienny, Abhinav Mohdi, Theofilos Strinopoulos, Annie Hale, Wayne Anderson, Luis C. Cobo, Niv Efron, Muktha Ananda, Shakir Mohamed, Maureen Heymans, Zoubin Ghahramani, Yossi Matias, Ben Gomes, and Lila Ibrahim. 2024. [Towards responsible development of generative ai for education: An evaluation-driven approach](#). *Preprint*, arXiv:2407.12687.
- Mert Karabacak and Konstantinos Margetis. 2023. Embracing large language models for medical applications: opportunities and challenges. *Cureus*, 15(5).
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- Deborah Kelemen. 1999. Why are rocks pointy? children’s preference for teleological explanations of the natural world. *Developmental psychology*, 35(6):1440.
- Minsun Kim, SeonGyeom Kim, Suyoun Lee, Yoosang Yoon, Junho Myung, Haneul Yoo, Hyunseung Lim, Jieun Han, Yoonsu Kim, So-Yeon Ahn, et al. 2024a. Designing prompt analytics dashboards to analyze student-chatgpt interactions in efl writing. *arXiv preprint arXiv:2405.19691*.
- Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024b. Evalm: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21.
- Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. [Derivation of new readability formulas \(automated readability index, fog count and flesch reading ease formula\) for navy enlisted personnel](#).
- David A Kolb et al. 2007. *The Kolb learning style inventory*. Hay Resources Direct Boston, MA.
- Katelyn E Kurkul and Kathleen H Corriveau. 2018. Question, explanation, follow-up: A mechanism for learning from others? *Child Development*, 89(1):280–294.
- Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. 2024. [Aligning to thousands of preferences via system message generalization](#). *Preprint*, arXiv:2405.17977.
- Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. Dapie: Interactive step-by-step explanatory dialogues to answer children’s why and how questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–22.



- Grace Li, Milad Alshomary, and Smaranda Muresan. 2024. "is chatgpt a better explainer than my professor?": Evaluating the explanation capabilities of llms in conversation compared to a human baseline. *arXiv preprint arXiv:2406.18512*.
- Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. [WANLI: Worker and AI collaboration for natural language inference dataset creation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tania Lombrozo and Nicholas Z Gwynne. 2014. Explanation and inference: Mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, 8:700.
- Silvia B Lovato, Anne Marie Piper, and Ellen A Wartella. 2019. Hey google, do unicorns exist? conversational agents as a path to answers to children’s questions. In *Proceedings of the 18th ACM international conference on interaction design and children*, pages 301–313.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.
- Guoqing Luo, Yu Tong Han, Lili Mou, and Mauajama Firdaus. 2023. Prompt-based editing for text style transfer. *arXiv preprint arXiv:2301.11997*.
- Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.
- Amanda M McCarthy and Frank C Keil. 2023. A right way to explain? function, mechanism, and the order of explanations. *Cognition*, 238:105494.
- Chancharik Mitra, Mihran Miroyan, Rishi Jain, Vedant Kumud, Gireeja Ranade, and Narges Norouzi. 2024. Retllm-e: Retrieval-prompt strategy for question-answering on student discussion forums. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23215–23223.
- Randall Munroe. 2015. *Thing explainer: complicated stuff in simple words*. Hachette UK.
- Sanghee Oh, Jung Sun Oh, and Chirag Shah. 2008. The use of information sources by internet users in answering questions. *Proceedings of the American Society for Information Science and Technology*, 45(1):1–13.
- John O’hayre. 1966. *Gobbledygook has gotta go*. US Department of the Interior, Bureau of Land Management.
- Ben Prystawski, Michael Li, and Noah Goodman. 2023. Why think step by step? reasoning emerges from the locality of experience. *Advances in Neural Information Processing Systems*, 36:70926–70947.
- Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2024. [Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure](#). Preprint, arXiv:2410.03781.
- Rod D Roscoe and Michelene TH Chi. 2008. Tutor learning: The role of explaining and responding to questions. *Instructional science*, 36:321–350.
- Alexis Ross and Jacob Andreas. 2024. [Toward in-context teaching: Adapting examples to students’ misconceptions](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13283–13310, Bangkok, Thailand. Association for Computational Linguistics.
- Adena Schachner, Liqi Zhu, Jing Li, and Deborah Kelemen. 2017. Is the bias for function-based explanations culturally universal? children from china endorse teleological explanations of natural phenomena. *J. Exp. Child Psychol.*, 157:29–48.
- Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2024. Ruffle&riley: Insights from designing and evaluating a large language model-based conversational tutoring system. In *International Conference on Artificial Intelligence in Education*, pages 75–90. Springer.
- Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. Chacha: Leveraging large language models to prompt children to share their emotions about personal events. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20.
- Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*.
- John Stamper, Ruiwei Xiao, and Xinying Hou. 2024. Enhancing llm-based feedback: Insights from intelligent tutoring systems and the learning sciences. In *International Conference on Artificial Intelligence in Education*, pages 32–43. Springer.
- Wadim Strielkowski, Veronika Grebennikova, Alexander Lisovskiy, Guzalbegim Rakhimova, and Tatiana Vasileva. 2024. Ai-driven adaptive learning for sustainable educational transformation. *Sustainable Development*.
- Justin Sulik, Jeroen van Paridon, and Gary Lupyan. 2023. Explanations in the wild. *Cognition*, 237(105464):105464.
- Lihui Sun and Liang Zhou. 2024. [Does generative artificial intelligence improve the academic achievement of college students? a meta-analysis](#). *Journal of Educational Computing Research*, 62(7):1896–1933.



- Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, Mengting Wan, Tara Safavi, Jennifer Neville, Chirag Shah, Ryen W White, Reid Andersen, et al. 2024. The use of generative search engines for knowledge work and complex tasks. *arXiv preprint arXiv:2404.04268*.
- Clemmie Telford. 2021. *But Why?: How to answer tricky questions from kids and have an honest conversation with yourself*. Headline Home.
- Barbara Tizard and Martin Hughes. 2008. *Young children learning*. John Wiley & Sons.
- Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T Hickey, Ronghuai Huang, and Brighter Agyemang. 2023. What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education. *Smart learning environments*, 10(1):15.
- Bailin Wang, Zi Wang, Xuezhi Wang, Yuan Cao, Rif A Saurous, and Yoon Kim. 2023. Grammar prompting for domain-specific language generation with large language models. *Advances in Neural Information Processing Systems*, 36:65030–65055.
- Adrian F Ward. 2021. People mistake the internet’s knowledge for their own. *Proceedings of the National Academy of Sciences*, 118(43):e2105061118.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Lyumanshan Ye, Jiandong Jiang, Danni Chang, and Pengfei Liu. 2024. Storypark: Leveraging large language models to enhance children story learning through child-ai collaboration storytelling. *arXiv preprint arXiv:2405.06495*.
- Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.
- Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. 2024. Mathemyths: leveraging large language models to teach mathematical language through child-ai co-creative storytelling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–23.
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. *Wildchat: Im chatgpt interaction logs in the wild*. *Preprint*, arXiv:2405.01470.

## A Related Work

Previous research has explored the novelty of LLM-generated explanations (for various types of question-answering tasks), assessing how much additional information an explanation provides beyond what is already contained in the question (Chen et al., 2023; Joshi et al., 2023). Recent advancements in LLMs integrated them into educational applications such as conversational tutoring systems (CTSs) (Schmucker et al., 2024), intelligent tutoring systems (ITSs) (Stamper et al., 2024), and other AI-driven tutoring frameworks tailored especially for younger learners (Strielkowski et al., 2024). Learners spend more time with these systems, integrating them into daily life, sharing even their deep emotions and daily experiences (Seo et al., 2024). This deep bond can make these systems the first point of reference for learners, especially younger ones, whose questions increase in the presence of an easily accessible responder (Tizard and Hughes, 2008). Meanwhile, teachers and parents may struggle with some of these questions (Telford, 2021), or learners may find their explanations unsatisfactory (Corriveau and Kurkul, 2014). Corriveau and Kurkul (2014) showed that children as young as 3 to 5 could ask complex questions about physics, biology, and social science, detect circular explanations, and reject them. They establish causal relationships (Kurkul and Corriveau, 2018) and enhance their theory-building abilities (Callanan and Oakes, 1992; Chouinard et al., 2007) by exploring the world through ‘why’ and ‘how’ questions. While teachers leverage LLMs to explore diverse teaching strategies (Feldhus et al., 2024), accommodate different learning styles (Kolb et al., 2007), and tailor instruction based on insights from conversation logs (Kim et al., 2024a), they also express concerns that learners may encounter concepts misaligned with educational goals (Kim et al., 2024a).

Learners, in turn, often find LLM responses unsuitable due to context ignorance (Lovato et al., 2019) or excessive length (Lee et al., 2023; Bertrand et al., 2023). For example, when asked, “Why do polar bears have white fur?” an AI might respond: “Polar bears have white fur to blend into their environment. Their coat is so well camouflaged in Arctic settings that it can sometimes pass as a snowdrift. Interestingly, their fur contains no white pigment; rather, a polar bear’s skin is black, and its hairs are hollow.” (Lee et al., 2023)

To address these challenges, researchers have explored interactive dialogues (Lee et al., 2023), tutoring mechanisms (Roscoe and Chi, 2008) (such as learning-by-teaching (Schmucker et al., 2024)), and co-creative storytelling (Zhang et al., 2024; Ye et al., 2024) to foster effective pedagogical interactions. Lee et al. (2023) proposed 23 guidelines for tailoring explanations for children, including the use of examples, personifications, and prompts for critical thinking. Recently, Kim et al. (2024b) envisioned a human-AI interaction framework where user-specified criteria guide LLM outputs.

Our research bridges this gap by identifying mismatches between desired and generated explanations for the groups of interest, investigating their causes, and proposing solutions. This is crucial as LLMs become more integrated into daily life, particularly for children who interact with AI through text-based interfaces and voice assistants like AI agents in home (Lovato et al., 2019). Improving explanation-tailoring methods could enable a dedicated for children mode in such devices, ensuring more age-appropriate and effective responses.

## B ELI-WHY Dataset

### B.1 Generation questions from GPT-4

Table 2 shows the 50 “Why?” questions (from Sulik et al. (2023)) that were used as seed examples to guide our question generation for ELI-WHY. Refer to Table 3 for the configuration and prompt used for generating ‘Why’ questions.

### B.2 Filters after question generation

After generating the questions, we manually reviewed all 30,671 instances to filter out invalid or toxic entries, including deduplicating “similar” questions. The filtering process took  $\sim 12$  hours. This human annotation process, rather than a rule-based system, ensured that only high-quality, non-toxic and non-hallucinated questions were retained. Additionally, we further filter questions if they are *too niche* for a given domain. This is evaluated using an “answerability” task performed by participants recruited from Amazon Mechanical Turk. Annotator’s were posed the following question – *Given “Why” questions about world phenomena, you have to judge whether you can answer the question on your own with some help from external resources.* We determined that questions which cannot be answered by lay annotators without any external resources would be too niche, and thus

not suitable to be grade-tailored. Each question is annotated by three annotators. We only keep questions where all annotators agree that they would be able to answer the question. As detailed in Table 4, these “Why” questions were excluded from the final dataset because annotators determined that they were too niche.

### B.3 Distribution of different academic disciplines in ELI-WHY

For classifying the question fields, we used a classification task with a reference list of fields derived from the “Outline of Academic Disciplines” (using its second subheader)<sup>12</sup>. We run this task with Llama-3.2-3B-Instruct and Llama-3.3-70B-Instruct with the prompt in Table 5; We run both models 20 times each; if Llama did not assign a question to one of the provided fields, we prompted it to return the most suitable field it could find, and manually classified it to the closest field.

Figure 7 shows all the fields classified by Llama-3.3-70B-Instruct. We performed a sanity check of the field classification by performing a manual check of 50 questions.

## C Generating Explanations for ELI-WHY

### C.1 Prompts used for explanation generation

Table 6 presents the complete set of prompts used for dataset collection and explanation generation. GPT-4 has a proprietary license. While all our analysis in this paper are based on GPT-4, we also generate explanations from Llama-3.2-3B-Instruct to conduct automatic evaluation Appendix E.4.

### C.2 Model details

Table 7 shows the configurations of GPT-4 and Llama-3.2-3B-Instruct used during explanation generation.

## D Human Experiments

Participants were recruited from Prolific<sup>13</sup>, who consented to our study. Participants were given an option to exit the study at any point. For Perceived Background Match evaluation in Section 4.1, participants were screened based on location (United States), education level (high school

<sup>12</sup>Subheadings outlined in [https://en.wikipedia.org/wiki/Outline\\_of\\_academic\\_disciplines](https://en.wikipedia.org/wiki/Outline_of_academic_disciplines)

<sup>13</sup><https://www.prolific.com/>

ID	Domain	Discipline	Question Text
1	STEM	Physics	Why does thunder make a noise?
2	STEM	Biology	Why do flies like poop?
3	Non-STEM	Linguistics	Why are there so many languages in the world?
4	Non-STEM	Earth science	Why are there waves in the ocean?
5	STEM	Biology	Why do we need sleep?
6	STEM	Biology	Why do leaves change color in fall?
7	Non-STEM	Psychology	Why do we dream?
8	STEM	Biology	Why is human birth more difficult than for other animals?
9	STEM	Biology	Why do our nails grow?
10	Non-STEM	Economics	Why are coins round?
11	STEM	Physics	Why is glass transparent since it is made from the same thing as sand?
12	Non-STEM	Psychology	Why do people bite their nails?
13	Non-STEM	History	Why is number 13 considered unlucky?
14	STEM	Biology	Why are eggs egg-shaped?
15	STEM	Biology	Why did the dodo die out?
16	STEM	Engineering and technology	Why are manhole covers round?
17	STEM	Physics	Why does it echo if we yell in a cave but not a regular room?
18	STEM	Biology	Why do some animals live longer than others?
19	STEM	Biology	Why are polar bears white?
20	Non-STEM	Geography	Why are there so many countries in the world?
21	STEM	Biology	Why do we itch?
22	Non-STEM	Sociology	Why do fashions change?
23	Non-STEM	Sociology	Why do people get divorced?
24	STEM	Physics	Why is the sky blue?
25	STEM	Engineering and technology	Why do fridges hum?
26	STEM	Psychology	Why do people do drugs?
27	STEM	Physics	Why are snowflakes hexagonal?
28	Non-STEM	Anthropology	Why do we shake our heads for "no"?
29	STEM	Architecture and design	Why are most clocks round?
30	STEM	Astronomy	Why does Saturn have rings?
31	STEM	Earth science	Why did the dinosaurs die out?
32	STEM	Biology	Why do we hiccup?
33	STEM	Biology	Why do lions roar?
34	Non-STEM	Sociology	Why do some people want tattoos?
35	STEM	Biology	Why are dogs loyal?
36	Non-STEM	Psychology	Why does tickling make us laugh?
37	Non-STEM	Psychology	Why do some people bully?
38	STEM	Biology	Why do our noses run when we eat spicy food?
39	STEM	Biology	Why are lemons sour?
40	STEM	Biology	Why are we awake during the day and sleepy at night?
41	Non-STEM	Psychology	Why are women often more emotional than men?
42	STEM	Physics	Why is water transparent?
43	STEM	Biology	Why are honeycombs hexagonal?
44	Non-STEM	Architecture and design	Why are jeans blue?
45	STEM	Biology	Why are flowers colorful?
46	STEM	Engineering and technology	Why are flags rectangular?
47	Non-STEM	Psychology	Why do people fall in love?
48	Non-STEM	Psychology	Why do people lie about small things?
49	Non-STEM	Psychology	Why do we look around when we hear a noise?
50	STEM	Medicine and health	Why do people die?

Table 2: A set of 50 seed questions categorized by domain (STEM or Non-STEM) and academic discipline. Among these, 32 questions belong to STEM disciplines, while 18 fall under Non-STEM disciplines.

---

**Model:** GPT-4-0613, ~1.8 trillion parameters

---

**max\_tokens:** 4096**temperature:** 1.0

---

**Prompt:** Generate 100 non-STEM "why" questions.

For example, STEM "why" questions can be questions about Physics, Chemistry, Biology and Neuroscience. Non-STEM "why" questions can be about Humanities, Liberal Arts, Psychology, Law, Sociology or Socio-Cultural topics to name a few domains.

Some examples of non-STEM questions are -

1. Why do people bite their nails?
  2. Why are there so many languages in the world?
- 

Table 3: Model Configuration and Prompt Used for Overgenerating “Why” Questions. The domain (STEM and Non-STEM) and the in-context examples are changed multiple times to get diverse “Why” questions.

Question
Why is P always less than or equal to NP in complexity theory?
Why does Black-Scholes model matter to finance?
Why is quantum entanglement paradoxical?
Why was Euclid’s fifth postulate so controversial?
Why do carbon atoms form four bonds in organic chemistry?
Why do DNA strands run from the 3’ to 5’ direction?

Table 4: Examples of niche questions filtered out based on annotators’ answerability judgments.

or higher), and active participation to ensure high-quality responses. For informativeness simulation, participants were also recruited through Prolific and screened based on location, educational background, discipline of study (for graduate-level participants), and active participation on the platform. To ensure high-quality responses, all participants received detailed task instructions and were provided with examples to clarify expectations. Additionally, participants submitted natural language justifications for their selections, allowing further insight into their reasoning. Each participant annotated five questions, and each question received five independent annotations for each educational background. There were a total of 811 unique participants in all our studies.

## D.1 Annotator Filtering Criteria

We applied task-specific filtering criteria during participant recruitment. Table 8 details the screener settings for each group—Elementary School, High School, Physics Graduates, and Psychology Graduates. All annotators were required to reside in the US and have English as their primary language.

## D.2 Annotators Demographic Distribution

We collected demographic information, including highest education level, age, sex, and ethnicity, to ensure a representative sample. All annotators were required to have a country of residence in the United States and a primary language of English. Figure 8 presents the overall distribution of these demographics, confirming the diversity of our 811 unique annotators and enhancing the generalizability of our findings.

## D.3 Overview of the User studies

Figure 9 and Figure 10 provide an overview of the two human experiments, including instructions, examples, and the evaluation interface.

## D.4 Example annotations for perceived background match user study

Table 9 presents the annotations for the perceived background match user study, with different annotators’ justifications separated by semicolons. It shows examples where annotators mark the explanation’s perceived educational background based on their evaluation of the language and detail, and provided their justifications of choices. Note that annotators weren’t aware of the intended readability level.

Table 10 displays the annotations for the explanation informativeness user study, where the explanation column shows the final explanation shown to participants. In the last two rows of Table 10, annotators entered the “need more complex explanation” path despite the explanation being at a graduate level, resulting in no available explanations and both the informativeness and the match being marked as “no.”



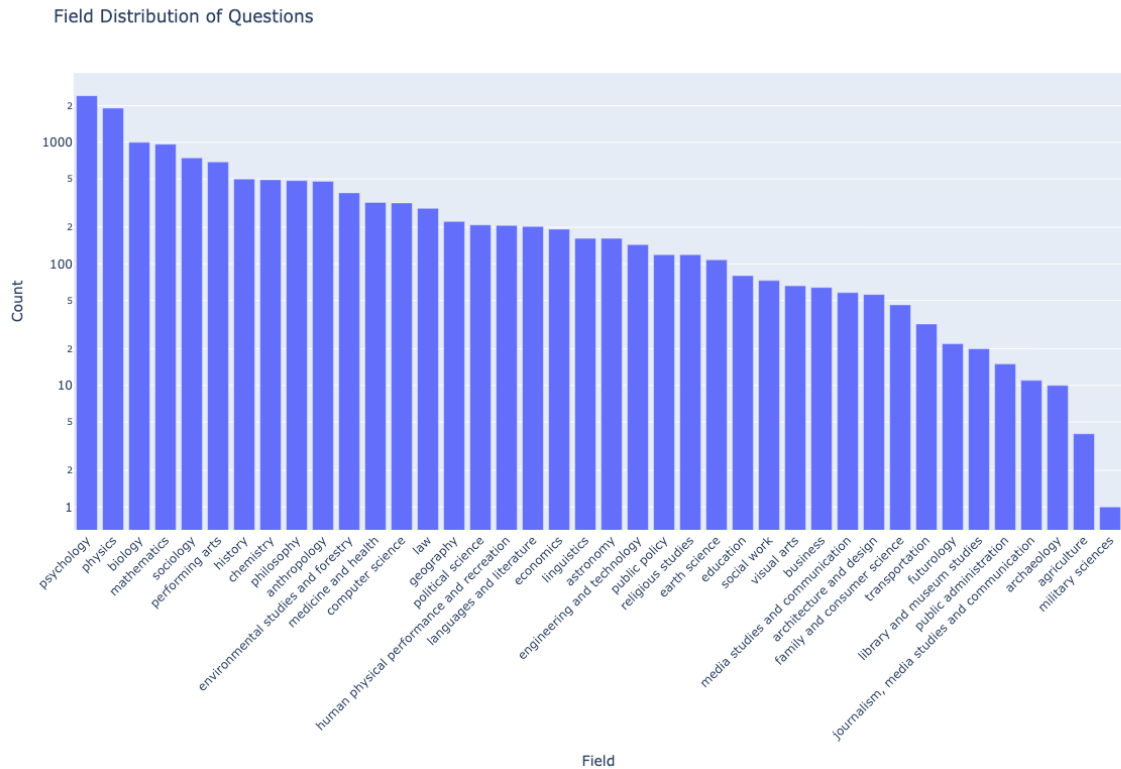


Figure 7: The distribution of question fields.

---

**model: Llama-3.2-3B-Instruct-4bit and Llama-3.3-70B-Instruct-4bit**

---

**max\_tokens:** 50

**temperature:** 0.2

**seed:** 0

**GPU:** Apple M4 Max

**Inferring time:** 8 hours

---

**Prompt:** You are a helpful assistant that categorizes questions into their relevant domains. Please classify each question into the most related domain in the following domains exactly as written (in lowercase): [discipline list]. Respond with only the domain name.

---

**Fallback Prompt:** You are a helpful assistant that categorizes questions into their relevant domains. Please classify each question into the most related domain in the following domains exactly as written (in lowercase): [discipline list]. Respond with only the domain name. If the domain is not in the list, please provide your best guess in one word or a short phrase (provide the field only without extra words).

---

Table 5: Model Configuration and Prompting Details for Discipline Classification. The Fallback Prompt will be used on Llama-3.3-70B-Instruct if we prompted both Llama-3.2-3B-Instruct and Llama-3.3-70B-Instruct 20 times each and none responds with a class in the list.

## D.5 Annotation Feedback Processing and Analysis

Additionally, we processed the annotators’ general feedback using Llama-3.3-70B-Instruct. For each feedback text, a function queries the LLM with a prompt that instructs it to output comma-

separated labels in the format “Aspect: direction” (e.g., “Vocabulary Complexity: complex” or “Sentence Structure: simple”). The query explicitly limits responses to our five predefined aspects and their valid directions (Vocabulary Complexity, Sentence Structure, Depth of Explanation, Technical

<p><b>Dataset Collection Prompt:</b>  Generate 100 [STEM/Non-STEM] "why" questions.  For example, STEM "why" questions can be questions about Mathematics, Deductive Reasoning, Logical Reasoning, Physics, Chemistry, Biology and Neuroscience. Non-STEM "why" questions can be about Humanities, Liberal Arts, Psychology, Law, Sociology or Socio-Cultural topics to name a few domains.</p>
<p><b>Graduate School Prompt:</b>  You will be asked a "Why" question. You are an expert in the domain of the why question you are asked. The user asking you the question also has a PhD in the domain of the question they asked.  Your job as an expert is to provide a concise explanation to the PhD holder. Make sure that the explanation is useful to the user - they will use it to validate and cross check important information. They may also use the explanation to teach that topic to a class.  Just provide the explanation as is - do not add any additional text like greetings or ornamental words.</p>
<p><b>High School Prompt:</b>  You will be asked a "Why" question. You are an expert in the domain of the why question you are asked. The user asking you the question is someone who holds a basic american high school education. You can assume they are a "layperson" to the domain of the question asked.  Your job as an expert is to provide a concise explanation to the user. They asked you the question as they were curious about the topic, so make sure that the explanation is useful to them.  Just provide the explanation as is - do not add any additional text like greetings or ornamental words.</p>
<p><b>Elementary School Prompt:</b>  You will be asked a "Why" question. You are an expert in the domain of the why question you are asked. The user asking you the question is someone who holds a basic american elementary school education.  Your job as an expert is to provide a concise explanation to the user.  Just provide the explanation as is - do not add any additional text like greetings or ornamental words.</p>
<p><b>Default Prompt:</b>  You will be asked a "Why" question. You are an expert in the domain of the why question you are asked.  Your job as an expert is to provide a concise explanation to the user.  Just provide the explanation as is - do not add any additional text like greetings or ornamental words.</p>

Table 6: Prompts used for Dataset Collection and Explanation Generation

Terms Usage, Overall Suitability). The function repeats the query if necessary (up to 10 times) until the returned labels match our predefined set. These labels are then used to quantify the feedback—calculating, for each aspect, the normalized difference between positive and negative mentions—which helps explain the perceived educational background in our analysis.

Figure 11 shows the normalized difference (positive minus negative counts, divided by total feedback) to our five predefined aspects. Each subplot corresponds to a specific combination of the intended (ground truth) and majority perceived educational levels (Elementary, High School, Graduate). A positive bar indicates that feedback leaned toward more complex/advanced language (e.g., “complex,” “in-depth,” “technical”), whereas a negative bar suggests simpler characteristics. These results align with our expectations and show that the tailored explanations exhibit distinct linguistic features corresponding to the perceived educational levels.

**E Automated Metrics for Explanation Evaluation**

**E.1 Metric Details**

**Surface-form Metrics.** Given generated explanations, we calculate the number of sentences, the average number of words per sentence, and the average reading time (which assumes 14.69 ms for each character read) for each explanation (Demberg and Keller, 2008). Additionally, we also calculate the Thing Explainer Out of Vocabulary (TE) Score proposed by August et al. (2024), which counts the ratio of words outside the top 2,000 most common words in English<sup>14</sup>.

**Readability Tests.** Given our goal of tailoring explanations to readers with varying educational backgrounds, the difficulty of the text plays a crucial role in ensuring comprehension. Prior research in psychology has developed a variety of *readability*

<sup>14</sup>This list of words was presented in the Thing Explainer book (Munroe, 2015) to explain scientific concepts using simple language

Config	GPT-4 Assignment	Llama Assignment	Qwen Assignment	Deepseek R1 Distilled Assignment
model	<b>gpt-4-0613</b> Number of parameters: ~1.8 trillion	<b>Llama-3.2-3B-Instruct</b> Number of parameters: 3 billion	<b>Qwen 2.5 14B Instruct</b> Number of parameters: 14 billion	<b>DeepSeek R1 Distill Llama 8B</b> Number of parameters: 8 billion
max_tokens	4096	4096	4096	4096
temperature	1	0.1	0.1	0.1
seed	random	0	0	0
GPU	N/A, openai api call	A100	A100	A100
Inferring time	N/A	2 hours	4 hours	4 hours

Table 7: Model Configurations for Explanation Generation

Config	Elementary School	High School	Physics Graduates	Psychology Graduates
	<b>Screener</b>	<b>Screener</b>	<b>Screener</b>	<b>Screener</b>
Current Country of Residence	United States	United States	United States	United States
Primary Language	English	English	English	English
Age	n/a	18-24	n/a	n/a
Approval Rate	98-100	98-100	n/a	n/a
Number of previous submissions	1000-10000	1000-10000	n/a	n/a
Highest education level completed	No formal qualifications	High school diploma/A-levels, Technical/community college	Graduate degree (MA/MSc/MPhil/other), Doctorate degree (PhD/other)	Graduate degree (MA/MSc/MPhil/other), Doctorate degree (PhD/other)
Degree subject	n/a	n/a	Natural Sciences	Psychology

Table 8: Prolific annotator filtering

*tests*, designed to quantify the difficulty of a piece of text. These tests typically rely on linguistic features such as sentence length, word complexity, and the presence of “easy” versus “hard” words (Flesch, 1948; O’hayre, 1966; Dale and Chall, 1948). Such metrics have traditionally been employed to assess the readability of textbooks, instructional materials, and technical documentation, making them highly relevant to our evaluation of explanations.

In this work, we leverage three widely used readability tests: Flesch-Kincaid Reading Ease, Linsear Write Formula, and Dale-Chall Readability Score. These were selected due to their interpretability, established validity in prior research, and diverse methodological approaches to estimating text difficulty. Moreover, many other readability metrics are often correlated with these three, making them representative choices for our analysis.

The **Flesch-Kincaid Reading Ease** metric assesses the readability of a text by considering the number of syllables, words, and sentences. Higher scores indicate text that is easier to read. The score is computed as:

$$206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right). \quad (1)$$

In this formula, a higher ratio of words per sentence and syllables per word reduces the score, signaling

greater difficulty.

The **Linsear Write Formula** evaluates readability by assigning points based on word difficulty. “Easy words,” with two syllables or less, earn 1 point, while “hard words,” with three syllables or more, earn 3 points. The total points are then divided by the number of sentences in the sample to produce a score  $r$ . If  $r > 20$ , the formula adjusts to  $Lw = r/2$ ; otherwise,  $Lw = r/2 - 1$ . The resulting score represents the grade level. It emphasizes sentence structure and word complexity, providing a score that correlates with U.S. grade levels. Lower scores indicate text that is easier to read. This metric is particularly effective for identifying the complexity of short instructional or educational texts.

The **Dale-Chall Readability Score** incorporates a curated list of 3,000 common “easy” words. Words not on this list are considered “difficult,” and the proportion of these difficult words, combined with sentence length, determines the readability score. Unlike the Flesch-Kincaid metric, lower scores indicate easier text, with thresholds provided to map scores to grade levels. It is calculated as:

$$0.1579 \left( \frac{\text{difficult words}}{\text{words}} \times 100 \right) + 0.0496 \left( \frac{\text{words}}{\text{sentences}} \right). \quad (2)$$

Texts with a high proportion of difficult words or

Question	Explanation	Intended Educational Background	Perceived Educational Background (Majority)	Justification
Why does chocolate melt in the mouth?	Chocolate melts in the mouth because it's made with cocoa butter, a type of fat that melts at or just below human body temperature. So when you put chocolate in your mouth, the heat from your body is enough to melt the cocoa butter and that's why the chocolate softens and dissolves.	Elementary	Elementary	Sentences are very easy to understand as well as concepts.; Uses basic words and simple sentence structure. Doesn't go overboard with detail; It uses simple sentence structure and wording and add extra detail to make it easily understood.
Why does the sun rises in the east and sets in the west?	The sun rises in the east and sets in the west because of the Earth's rotation. The Earth rotates on its axis, an imaginary line that runs from the North Pole to the South Pole. It rotates in a counterclockwise or eastward direction. Because of this rotation, it appears to us like the Sun is moving across the sky from east to west when in reality, we're the ones moving. This phenomenon is called apparent solar motion, and it's the reason for the sun's apparent path in the sky.	High School	Elementary	The wording and details suggests elementary school.; Wording, details and depth; The examples given are straight to the point and do not go into too much depth. The details of the reasoning are not too long or difficult to follow. The sentence structure is short. The use of words like axis, North Pole, South Pole, imaginary, etc. is not of an high school level, but does show that they have knowledge of science.
Why are people's blood types different?	Blood types are different due to the presence or absence of certain molecules, called antigens, on the surface of red blood cells. These antigens are essentially identifiers which your immune system use to recognize its own cells. The different blood types, A, B, AB, and O, refer to these specific antigens. Your blood type is determined by the genes you inherit from your parents.	Elementary	High School	The wording was simple as well as the sentence structure and details; This explanation uses more advanced terminology like molecules and antigens. But it does not go into too much detail about how they are different.
Why is there less gravity on the moon?	Gravity at a specific location is determined by the mass of the object and the distance from its center. The Moon is much smaller than Earth and has less mass. This reduces the gravitational pull that it has on objects on its surface. So, when you're on the Moon, you weigh less than you do on Earth because the Moon's gravitational pull is weaker due to its smaller mass.	High School	High School	This explanation uses terms that would be understandable to a high school level reader but too advanced for elementary reading. The comparison of the moon's density to earth's and the explanation of how gravity operates less on mass is a good detailed response for a high school level reader.; The reading level appears to be that of high school level students. It doesn't go deeply into mathematical details beyond a high school level. It doesn't provide advanced example of why you weigh less on the moon compared to earth.
Why do our bodies produce sweat when we're hot?	Sweating is the body's essential cooling mechanism. When your core body temperature rises, your hypothalamus, which is the body's thermostat, signals your sweat glands to start producing sweat. It's predominantly composed of water, but also includes minor amounts of various salts and urea. As the sweat evaporates from your skin, it helps to dissipate the heat and cool your body down. This process, known as evaporative cooling, maintains the body within a narrow, optimal temperature range, critical for the normal functioning of physiological processes and homeostasis.	Graduate	High School	The sentence structure is digestible, and topic is simple. Terminology like "predominantly," "dissipate," "critical," "physiological processes and homeostasis" indicate more advanced than elementary. However, not enough detail/sources to reach graduate level; Word usage is above elementary level, but should be fully comprehended by high school students.
Why is the human brain folded?	The human brain is folded into wrinkles, also known as sulci and gyri, to increase the surface area for more neurons. More neurons mean more processing power and cognitive abilities. The folding allows our brains to fit a larger number of neurons within the confined space of our skulls. Like squeezing more electronic circuits into a smaller computer chip, the folds allow for greater complexity and speed in our thinking process.	High School	Graduate	This article introduced specialized terminology related to the brain, like gyri. The article assumed that the reader would easily understand how increased surface area would lead to increased processing using only an analogy of a computer chip, which also requires preexisting electronics knowledge.; The sentences contain terminologies that are appropriate for professional audience or specialists
Why do fruits turn sweet when they ripen?	Fruits ripen due to the action of a plant hormone called ethylene. During the ripening process, ethylene accelerates the conversion of starches to sugars in the fruit, leading to an increased sweetness. Complex carbohydrates are broken down into simpler sugars; for instance, sucrose is converted into glucose and fructose. This is a part of the plant's strategy to make the fruit appealing to animals, so they will eat it and subsequently disperse the seeds in the process. Additionally, the process of ripening also often involves other changes, such as color, texture, aroma development, or an increase in nutritional content, which are conducive to seed dispersal and survival.	Graduate	Graduate	Detailed examples sounds like someone who may have advanced schooling in the subject.; I picked grad level, because it assumes prior knowledge of higher level science concepts.

Table 9: Annotations for perceived background match user study, different annotator's justifications are separated by semicolons.



Question	Explanation	Intended Educational Background	Informative	Matched	Justification
<b>Elementary School</b> Background Annotators					
Why do we yawn?	We yawn as a way to wake up. When you're tired or bored, your body takes in less oxygen because your breathing slows down. Yawning helps you take in a lot of air and uses muscles which sends signals to your brain to wake it up. It also helps cool the brain. It's even contagious because it's a primitive reflex humans have adapted for social bonding and empathy.	Elementary	Yes	Yes	I knew we yawn for empathetic reasons from some videos on psychology. And of course, I knew we yawn when we're tired or bored. I did not, however, know that we take in less oxygen from our slowed breathing when we're tired, and that is why we yawn. I also didn't know it sent signals to the brain to wake it that way.
Why does acid rain occur?	Acid rain occurs due to the release of sulfur dioxide (SO <sub>2</sub> ) and nitrogen oxide (NO <sub>x</sub> ) into the atmosphere. These gases are primarily released by burning fossil fuels like coal, oil, and gas in power plants, cars, and factories. Once in the atmosphere, they react with water, oxygen, and other substances to form sulfuric acid and nitric acid. These acids can then be carried over long distances by wind and eventually fall as acid rain. This process can be harmful to the environment, damaging vegetation, aquatic habitats, and even buildings.	High School	Yes	No	While I knew acid rain can be caused from fossil fuel burning, I did not know the types of acid it produces and how it is actually formed from mixing with water, oxygen and other substances. It reinforces the damage that it can cause to the environment that I was also aware of.
<b>High School</b> Background Annotators					
Why does DNA have a double helix structure?	DNA has a double helix structure because of the nature of the molecules it's made of and the physical and chemical interactions between them. It consists of two strands of nucleotides, each made up of a sugar, a phosphate, and a base. The sugar and phosphate form the backbone of the DNA, and the bases (adenine, cytosine, guanine, thymine) from each strand bond together. Adenine always pairs with thymine, and cytosine with guanine. These pairs are of almost equal shape and size, generally fitting neatly between the two spiraling backbones, maintaining a constant distance, thus ensuring the helical structure. Furthermore, the double helix provides protection to the genetic information and is more stable, which is crucial for genetic inheritance.	High School	Yes	Yes	I had learned much about DNA in compulsory education but had forgotten the real reason for the double helix shape. This introduced to me the bases adenine, cytosine, guanine and thymine being in the mix.
Why do we have day and night?	Day and night occur due to the Earth's rotation on its axis. The Earth spins around an imaginary line that passes through the North and South Poles, known as the Earth's axis. It completes one full rotation in approximately 24 hours. During this rotation, only the half of the Earth that's facing the sun experiences day, as it's being illuminated by sunlight. Simultaneously, the other half of Earth faces away from the sun and is shrouded in darkness, and so it experiences night. As the Earth continues rotating, the regions exposed to sunlight shift, so different parts of the world pass through day and night cycles. This constant rotation creates the 24-hour day and night cycle we experience.	Graduate	No	No	The explanation does not introduce a completely new concept to me, as I was already familiar with the idea that Earth's rotation causes day and night. However, it reinforced my understanding by clearly describing the Earth's axis as an imaginary line passing through the North and South Poles. The explanation also effectively connected the idea of Earth's continuous rotation to the 24-hour cycle we experience. While the core concept was familiar, the phrasing and details helped refine my understanding.
<b>Graduate School</b> (Psychology) Background Annotators					
Why do we sleep?	Sleep serves multiple purposes that are essential to the human brain and body. One of the key reasons we sleep is for restoration. During sleep, the body has the chance to repair muscle tissue and synthesize proteins, restore energy, and release the growth hormone which is crucial for growth and development. In the brain, sleep helps in clearing out waste products, some of which are potentially neurotoxic, that brain cells generate as a byproduct of their activity. This clearance is carried out by the glymphatic system which becomes especially active during sleep. Sleep also has a significant role in memory consolidation. The neural connections that form our memories are strengthened during sleep, particularly during REM sleep. During this time, the brain reorganizes and restructures memories, which may result in more creativity. Lastly, lack of sleep can cause numerous health problems, including weakened immunity, increased risk of chronic diseases, poor mood, and cognitive difficulties. Thus, sleep is essential for maintaining optimal health and well-being.	Graduate	No	No	It seems like common knowledge that I have heard many times in the past. I have read many empirical scientific studies and listened to many professionals exclaim the many benefits of sleep hygiene over the years. This is crucial knowledge to understand in terms of vitality and longevity of our existence and mental fitness in everyday life.

Table 10: Annotations for the Explanation Informativeness User Study. The Explanation column displays the final explanation shown to participants when they viewed the explanations (see Figure 5 for details on the annotator's interaction). In the last two rows, annotators entered the "need more complex explanation" path despite the explanation being at a graduate level, resulting in no available explanations and both informative and matched marked as "no."

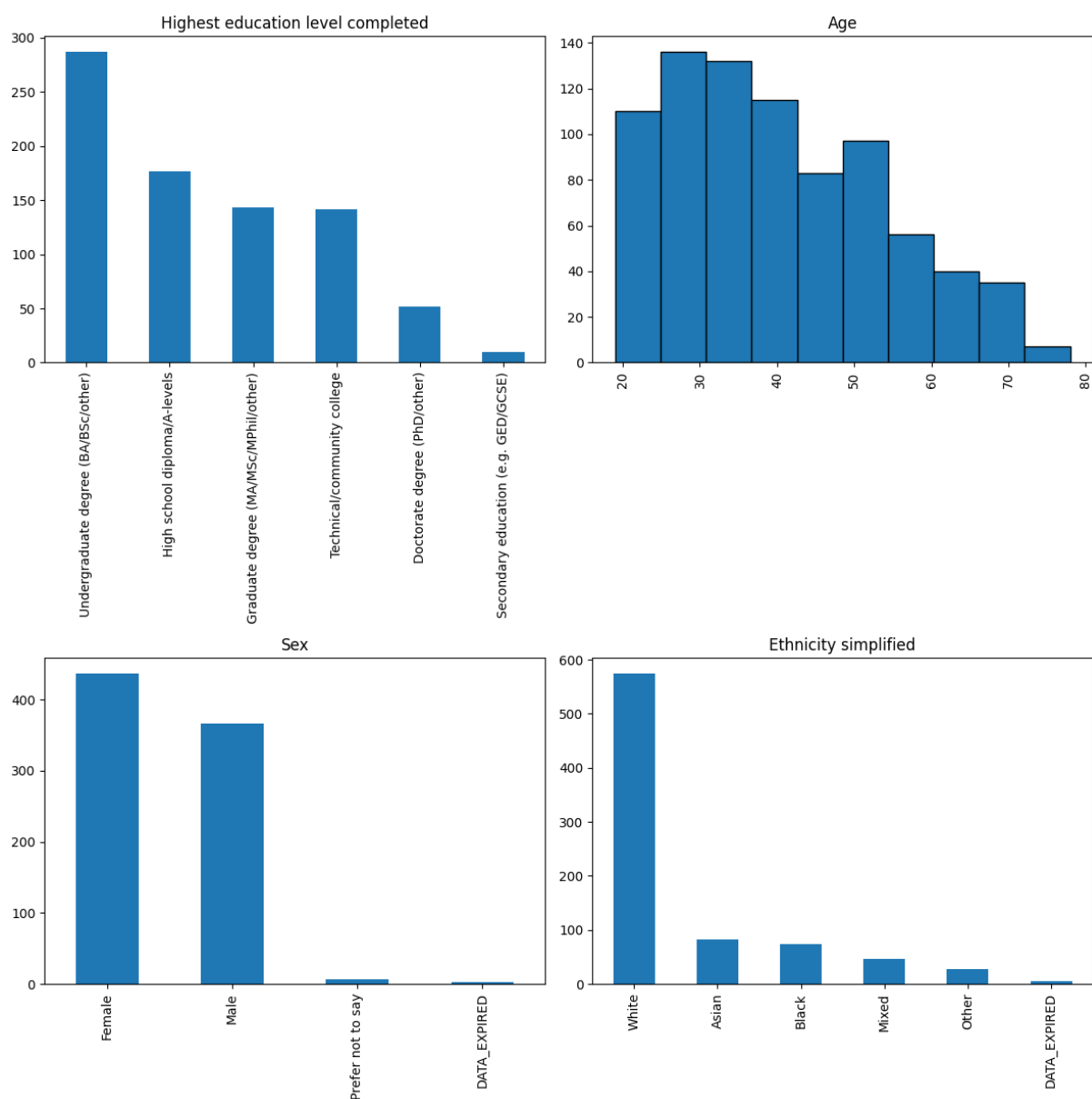


Figure 8: Annotators Demographic Distribution

longer sentences result in higher scores, reflecting increased reading difficulty.

The computed scores for all three readability metrics—Flesch-Kincaid Reading Ease, Linsear Write Formula, and Dale-Chall Readability Score—are mapped to U.S. grade levels in Table 11, which provides a detailed breakdown of score ranges and their interpreted readability levels (Flesch, 1979; O’hayre, 1966; Dale and Chall, 1948).

**Reasoning Type.** In addition to surface form metrics and readability tests, we also characterize the kind of reasoning used to answer these “why” questions. “Why” questions can be answered in

multiple ways, but the two predominant manners are mechanistic (explaining the *process* or *mechanism* of how a phenomenon came to be) or teleological (explaining the *purpose* or *function* of the phenomenon) (Lombrozo and Gwynne, 2014), and this reasoning style has often been shown to distinguish explanations preferred by people from various educational backgrounds like children or students (McCarthy and Keil, 2023; Kelemen, 1999). E.g., “Why is the sky blue?” can be answered either by explaining rayleigh scattering (mechanistic) or that the blue sky has a prettier contrast (teleological).

For this classification task, we employ Llama-3.2-3B-Instruct using the prompt and configura-

Test	Score Range	Interpretation
Flesch Reading Ease	90–100	5th grade
	80–90	6th grade
	70–80	7th grade
	60–70	8th & 9th grade
	50–60	10th to 12th grade
	30–50	College
	10–30	College graduate
	0–10	Professional
Linsear Write Formula	0–1	Pre-Kindergarten to 1st grade
	1–5	1st to 5th grade
	5–8	5th to 8th grade
	8–11	8th to 11th grade
	11+	11th grade to college
Dale-Chall Readability	0–5	4th grade or lower
	5–6	5th–6th grade level
	6–7	7th–8th grade level
	7–8	9th–10th grade level
	8–9	11th–12th grade level
	9–10	College level

Table 11: Readability Score Interpretations

tions in Table 12.

## E.2 Significance tests in automated metric evaluation

Beyond average values, we also examine the distribution of scores for each metric (see Appendix E.4), and observe significant differences in the scores. We evaluate two aspects: (1) a two-tailed test assessing whether explanation distributions are significantly distinct across these groups (Overall) and (2) a one-sided test verifying whether metric values follow the expected complexity ordering (Order Preserved), i.e., Elementary School < High School < Graduate School. In both tests, across all explanation pairs, we observe statistically significant differences ( $p < 0.001$ ), confirming that GPT-4’s explanations not only vary in complexity but also follow a systematic progression in difficulty. We conduct significance tests for all three pairs of explanation types (Elementary School vs. High School, High School vs. Graduate School, and Elementary School vs. Graduate School) using the Kolmogorov-Smirnov (KS) test. We opted for the KS test because it is a nonparametric method that compares entire distributions without

assuming normality. Given that many of our distributions are not bell-shaped, a t-test which assumes normality would not be suitable here.

## E.3 Correlation between automated metrics and user perceived educational backgrounds

We computed the *correlation ratio* across different automatic readability metrics to assess the correlation of annotator judgments with automated metrics. The correlation ratio measures the categorical-continuous association (Fisher, 1970). Higher values indicate stronger association between the automated readability scores and the perceived educational background.

Table 13 summarize the correlation ratios for various readability metrics. Notably, the Flesch Reading Ease score shows a medium correlation (0.425) with the annotator-perceived educational background, indicating that automated assessments align well with human judgments.

Config	Assignment
model	<b>Llama-3.2-3B-Instruct-4bit</b> Number of parameters: 3 billion
max_tokens	10
temperature	0.1
GPU	Apple M4 Max
Inferring time	1 hour per 13k explanations

**Prompt:** You are a helpful assistant. Below is a question followed by an explanation. For the explanation, classify whether it is mechanistic or teleological. A mechanistic explanation describes how something happens, focusing on the processes, systems, or mechanisms involved. A teleological explanation describes something in terms of its goal, purpose, or intended outcome.

Question: [Question]

Explanation: [Explanation]

Please provide your classification in one word: mechanistic or teleological.

Table 12: Model Configuration and Prompting Details for Reasoning Type Classification.

Metric	Correlation Ratio $\eta$
Num Sentences	0.360
Avg Words Per Sentence	0.177
Reading Time	0.511
TE Score	0.441
Flesch Reading Ease	0.425
Linsear Write Formula	0.291
Dale Chall Readability Score	0.324

Table 13: Correlation ratios between automated readability metrics and annotator-perceived educational background.

#### E.4 Readability and Explanation Type Distributions

We report the distributions of several automated metrics computed over the generated explanations. These metrics include surface-form attributes such as sentence count, words per sentence, and estimated reading time, as well as traditional readability tests including Flesch-Kincaid Reading Ease, Linsear Write Formula, and Dale-Chall Readability Score.

Table 14 presents the aggregated scores across all explanations, while Table 15 and Table 16 provide a breakdown for STEM and non-STEM questions, respectively.

Figure 12 shows the overall distribution of read-

ability scores with explanations generated by GPT-4, with trends consistent with our expectations. We observe that explanations tailored for audiences with higher educational backgrounds exhibit longer sentences and increased reading times, leading to lower Flesch-Kincaid scores and higher Linsear Write and Dale-Chall scores.

Figure 13 displays the readability distribution evaluated using Llama-3.2-3B-Instruct. The overall trend is the same. In rare cases (<1%) the model generated repetitive phrases (e.g., “gentrification-induced gentrification-induced gentrification-induced...”), which significantly impacted the variance of certain readability metrics (up to 50%). To mitigate this effect, we applied a rough filtering step to remove these anomalous outputs prior to analysis.

Figure 14 and Figure 15 displays readability distributions for Qwen 2.5 14B Instruct and DeepSeek R1 Distill LLaMA 8B respectively.

When splitting the data into STEM (Table 15) and non-STEM (Table 16) domains, we observe the same overall trends in readability across educational levels. However, the reasoning type distribution shows a divergence between domains. For STEM questions, the mechanistic reasoning proportion is significantly higher, approaching 90% for both GPT-4 and Llama-3.2-3B-Instruct models’ explanations, whereas for non-STEM questions, the proportion for mechanistic reasoning is generally around 30%.

#### E.5 Informational overlap between tailored and baseline explanations

To analyze the informational overlap between explanations, we introduce the TE Set Difference (TESDiff). While the TE Score captures the proportion of simple, commonly used words in an explanation, TESSDiff measures *how much* of the TE word set from one explanation is *missing in another*. Specifically, it quantifies the proportion of TE words in one explanation that do not appear in another, normalized by the total number of unique TE words in the first explanation:

$$\text{TESDiff}(x, y) = \frac{|(x \setminus y) \cap TE|}{|x \cap TE|} \quad (3)$$

A higher TESSDiff score indicates that the explanation  $x$  contains significantly more unique information compared to  $y$ , whereas a lower TESSDiff score suggests greater overlap between the two explana-



Type of Metric	Metric	GPT-4				Llama-3.2-3B-Instruct				
		Elementary School	High School	Graduate School	Default	Elementary School	High School	Graduate School	Default	Web-Retrieved
Surface-form	# Sentences	04.63 ± 01.34	07.08 ± 02.53	08.46 ± 02.62	05.07 ± 01.63	03.29 ± 01.63	06.70 ± 02.97	09.10 ± 03.33	04.24 ± 02.63	02.30 ± 00.90
	Avg. # Words / Sentence	18.43 ± 03.47	19.17 ± 03.36	20.00 ± 03.38	19.35 ± 03.57	20.39 ± 05.10	21.30 ± 03.81	23.12 ± 03.69	23.74 ± 04.88	17.26 ± 06.90
	Avg. Reading Time (s)	06.36 ± 01.75	10.57 ± 03.65	13.93 ± 04.05	07.81 ± 02.41	04.61 ± 02.14	10.97 ± 05.00	17.05 ± 06.30	07.93 ± 04.71	02.93 ± 01.04
	TE Score	00.43 ± 00.09	00.49 ± 00.09	00.55 ± 00.09	00.50 ± 00.09	00.37 ± 00.11	00.47 ± 00.09	00.54 ± 00.10	00.53 ± 00.11	00.44 ± 00.12
Readability Tests	Flesch-Kincaid Reading Ease (↓)	53.82 ± 14.52	45.51 ± 14.26	34.70 ± 14.43	41.00 ± 15.73	60.91 ± 17.14	46.39 ± 15.11	34.56 ± 14.51	33.68 ± 16.87	53.35 ± 18.52
	Linear Write Formula (↑)	11.67 ± 02.77	12.15 ± 02.70	13.21 ± 02.70	13.16 ± 02.82	12.29 ± 03.94	13.23 ± 03.09	14.88 ± 02.93	16.43 ± 03.93	11.10 ± 05.21
	Dale-Chall Readability Score (↑)	09.31 ± 01.16	09.84 ± 01.10	10.55 ± 01.10	10.35 ± 01.23	08.73 ± 01.47	09.49 ± 01.19	09.88 ± 01.13	10.80 ± 01.38	09.94 ± 01.56
Reasoning Type	% Mechanistic Reasoning	54.76%	57.28%	63.54%	58.51%	55.04%	57.81%	65.65%	63.03%	65.17%

Table 14: Comparison of surface-form, readability, and reasoning-type metrics across different education levels for GPT-4 and Llama-3.2-3B-Instruct, along with retrieved explanations.

Type of Metric	Metric	GPT-4				Llama-3.2-3B-Instruct				
		Elementary School	High School	Graduate School	Default	Elementary School	High School	Graduate School	Default	Web-Retrieved
Surface-form	# Sentences	04.39 ± 01.28	06.19 ± 02.12	07.44 ± 02.37	04.65 ± 01.43	03.24 ± 01.54	05.68 ± 02.55	07.64 ± 03.14	03.39 ± 01.81	02.52 ± 00.86
	Avg. # Words / Sentence	18.82 ± 03.78	19.83 ± 03.62	20.87 ± 03.58	19.88 ± 03.90	19.56 ± 05.24	20.80 ± 04.11	23.09 ± 04.11	23.97 ± 05.22	17.86 ± 05.96
	Avg. Reading Time (s)	05.88 ± 01.70	09.10 ± 03.60	12.14 ± 03.75	06.97 ± 02.11	04.14 ± 01.68	08.57 ± 03.96	13.51 ± 05.63	06.03 ± 02.94	03.12 ± 00.79
	TE Score	00.42 ± 00.09	00.46 ± 00.09	00.52 ± 00.09	00.47 ± 00.10	00.37 ± 00.10	00.45 ± 00.10	00.51 ± 00.11	00.49 ± 00.11	00.44 ± 00.12
Readability Tests	Flesch-Kincaid Reading Ease (↓)	60.61 ± 13.82	53.30 ± 13.47	42.89 ± 13.76	49.41 ± 14.99	66.97 ± 15.60	55.03 ± 14.10	42.81 ± 14.10	42.48 ± 16.57	57.61 ± 17.38
	Linear Write Formula (↑)	11.40 ± 03.00	12.10 ± 02.84	13.16 ± 02.87	12.91 ± 03.09	11.39 ± 04.02	12.57 ± 03.25	14.43 ± 03.23	16.09 ± 04.18	11.13 ± 04.50
	Dale-Chall Readability Score (↑)	08.81 ± 01.10	09.23 ± 01.04	09.92 ± 01.08	09.69 ± 01.17	08.39 ± 01.36	08.93 ± 01.17	09.39 ± 01.17	10.22 ± 01.41	09.54 ± 01.52
Reasoning Type	% Mechanistic Reasoning	88.49%	89.91%	93.09%	91.58%	89.26%	92.59%	95.26%	94.40%	93.92%

Table 15: Comparison of surface-form, readability, and reasoning-type metrics across different education levels for GPT-4 and Llama-3.2-3B-Instruct, on the STEM questions split.

Type of Metric	Metric	GPT-4				Llama-3.2-3B-Instruct				
		Elementary School	High School	Graduate School	Default	Elementary School	High School	Graduate School	Default	Web-Retrieved
Surface-form	# Sentences	04.83 ± 01.36	07.85 ± 02.61	09.35 ± 02.51	05.44 ± 01.70	03.34 ± 01.70	07.59 ± 03.03	10.36 ± 02.96	04.97 ± 02.98	02.12 ± 00.88
	Avg. # Words / Sentence	18.09 ± 03.13	18.60 ± 03.01	19.25 ± 02.99	18.88 ± 03.18	21.10 ± 04.86	21.72 ± 03.47	23.15 ± 03.30	23.55 ± 04.55	16.72 ± 07.60
	Avg. Reading Time (s)	06.77 ± 01.69	11.84 ± 03.62	15.49 ± 03.64	08.53 ± 02.42	05.02 ± 02.39	13.05 ± 04.88	20.10 ± 05.14	09.57 ± 05.29	02.77 ± 01.20
	TE Score	00.44 ± 00.09	00.52 ± 00.08	00.58 ± 00.07	00.53 ± 00.08	00.37 ± 00.11	00.49 ± 00.09	00.56 ± 00.09	00.57 ± 00.10	00.45 ± 00.12
Readability Tests	Flesch-Kincaid Reading Ease (↓)	47.93 ± 12.38	38.75 ± 11.15	27.62 ± 10.80	33.73 ± 12.38	55.63 ± 16.67	38.89 ± 11.53	27.44 ± 10.55	26.12 ± 13.01	49.60 ± 18.69
	Linear Write Formula (↑)	11.91 ± 02.54	12.19 ± 02.57	13.26 ± 02.55	13.37 ± 02.53	13.07 ± 03.69	13.81 ± 02.83	15.27 ± 02.58	16.72 ± 03.67	11.07 ± 05.76
	Dale-Chall Readability Score (↑)	09.74 ± 01.03	10.38 ± 00.83	11.09 ± 00.79	10.93 ± 00.96	09.03 ± 01.50	09.97 ± 00.97	10.31 ± 00.90	11.31 ± 01.12	10.29 ± 01.50
Reasoning Type	% Mechanistic Reasoning	25.55%	29.02%	37.97%	29.88%	25.26%	27.63%	40.05%	36.05%	39.92%

Table 16: Comparison of surface-form, readability, and reasoning-type metrics across different education levels for GPT-4 and Llama-3.2-3B-Instruct, on the Non-STEM questions split.

tions, meaning they are more similar in terms of TE word usage.

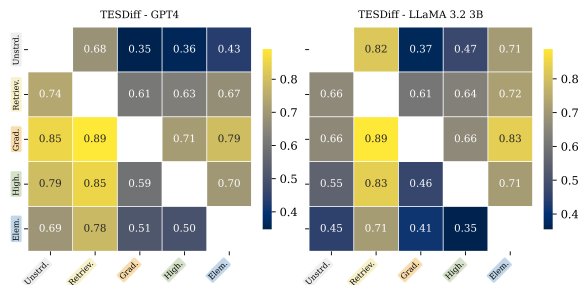


Figure 16: Heatmap of TESDiff scores across different explanation types for GPT-4 (left) and Llama-3.2-3B-Instruct (right). Higher TESDiff scores indicate greater divergence from general-purpose responses, with “Graduate School” explanations consistently exhibiting the highest TESDiff values across models.

Figure 16 presents a heatmap of TESDiff scores across different explanation types. We observe that explanations generated for **Graduate School** have consistently **high** TESDiff scores when compared to other explanations, suggesting that they introduce the most new information and diverge the

most from general-purpose responses. In contrast, explanations generated for **High School** and **Elementary School** audiences exhibit lower TESDiff scores between them, indicating that they are more similar to each other and that GPT-4 does not make as strong a distinction between these two levels.

While looking at **Default** and other explanations, we find that **Elementary School** and **High School** prompts produce the most similar explanations to the **Default** setting, as indicated by their lower TESDiff scores when compared to default explanations. This suggests that when GPT-4 generates explanations *without* an explicitly assigned educational background, the resulting explanations tend to align more with these middle-ground audiences. On the other hand, **Web-Retrieved** explanations show greater similarity to **High School** and **Graduate School** explanations, likely because web sources often contain more technical detail than **Default** responses.

Automated metrics for evaluating explanation complexity and similarity indicate that GPT-4 gen-

erates explanations with distinct scores for different audience roles. Surface-form attributes, readability tests, and the TESDiff score all show systematic differences across educational levels, suggesting that explanations vary in length, complexity, and word choice. However, these numerical differences do not necessarily imply that the explanations are *useful or well-suited for their intended audiences*. While GPT-4 can produce explanations that differ in measurable ways, whether these differences actually improve comprehension or align with user needs requires further human evaluation, which we explore in the Human Experiment section.

## **F Case study with ELI5 subreddit questions**

To examine why explanation complexity varies, we conducted a case study using the *Explain Like I'm Five* (ELI5) subreddit. For each question where we performed human annotations on explanation readability, we retrieved the most similar question from ELI5 using the Reddit API. We queried up to 25 relevant posts and identified the closest match using the SentenceTransformer all-MiniLM-L6-v2.

Table 17 presents a selection of our findings, including the original question, its most similar counterpart in ELI5, the similarity score, and the intended-to-perceived educational mapping. Our analysis reveals a consistent trend: when a question has a highly similar counterpart in ELI5, explanations tend to be oversimplified, particularly for prompts originally intended for higher educational levels. This suggests that when tasked with explaining such questions, the model is more likely to produce simpler explanations, potentially due to their prevalence in training data.

Instructions

In this task, you will be presented with "why" questions. You will then be shown explanations for these questions. Your task is to assess who can be an **appropriate reader** of these explanations.

More concretely, given an explanation, you have to choose what the educational background of an appropriate reader is, from the following options followed by a justification:

- Elementary School Level Education:** Typically covers up to American Grade 4, focusing on fundamental concepts in reading, writing, and math. Adults with this education level may have limited theoretical knowledge but often possess practical skills beyond a child's understanding.
- High School Level Education:** Extends through American Grade 12 to approximately the sophomore year of undergraduate studies, introducing concepts across various fields and enabling connections between theories. Adults at this level have a foundational grasp of academic subjects but may need guidance to deepen understanding.
- Graduate School Level Education:** Graduate school in America follows a bachelor's degree, offering advanced, specialized education at the master's or doctoral level. Adults with this education have few knowledge gaps and possess advanced expertise in specific areas without needing foundational instruction.

You will then have to provide **justifications** for the choice you have made. Keep the following pointers in mind while providing the justification:

- Wordings/Terminology:** Use of scientific words, abbreviations, or complex synonyms influences readability levels.
- Examples/Analogies:** Mark only if examples or analogies specifically impact readability level, not if included just for detail.
- Sentence Structure:** Consider digestibility, multiple concepts, and sentence length as factors affecting readability.
- Details and Depth:** Extra details in explanations that aid in determining readability level should be marked.

**IMPORTANT NOTE:** Please go over the examples before proceeding.

(a) Instruction for perceived background match user study: We asked annotators to assess the appropriate reader of explanations.

Question

Why is voting an important responsibility in democratic nations?

Explanation:

Voting is morally special in its own right regardless of the fact that citizens may also be bound to discharge other duties of aid. Because governments are powerful entities that distribute and shape access to basic social goods as no other organization does, the mechanism to install them is unique and deserves moral attention separately. Other political, non-electoral ways of influencing government matter, but they lose all relevance if elections are absent.

**Note:** If you need to revisit the instructions and examples again, feel free to press on the tabs above the question.

## Readability: High School

For this question, what do you think is an appropriate educational level for this explanation? If you think your choice lies in the middle (e.g. if you think an explanation requires between High School and Graduate Degree to understand), use your best judgement for your selection and add that in the justification box. You can think about what an **average** person of that level would understand.

- Elementary:** Typically covers up to American Grade 4
- High School:** Extends through American Grade 12 to approximately the sophomore year of undergraduate studies
- Graduate:** Graduate school in America follows a bachelor's degree, offering advanced, specialized education at the master's or doctoral level.



## Justify your selection below.

- Keep the following pointers in mind while providing the justification:
- Wordings/Terminology:** Use of scientific words, abbreviations, or complex synonyms influences readability level.
  - Examples/Analogies:** Mark only if examples or analogies specifically impact readability level, not if included just for detail.
  - Sentence Structure:** Consider digestibility, multiple concepts, and sentence length as factors affecting readability.
  - Details and Depth:** Extra details in explanations that aid in determining readability level should be marked.

(c) Question for perceived background match user study: Annotators select the readability level using a scrollbar and justify their choice based on factors such as wording, examples, sentence structure, and depth of detail.

## Example 1:

Question 1:

Why is the sky blue?

Explanation 1 (rate this!):

When sunlight reaches Earth's atmosphere, it is scattered in all directions by all the gases and particles in the air. Blue is scattered by air molecules more than other colors because it travels as shorter, smaller waves. This is why we see a blue sky most of the time.

- Answer:** Elementary School Education
- Justification:** The terminology used in this explanation like "waves" and "blue light scattering" implies some basic knowledge of science. This seems to be less than a high school explanation because of the depth in which it is explained.

Explanation 2 (rate this!):

The blue colour of the sky is nowadays usually explained by Rayleigh scattering, an explanation that is, however, not entirely correct. In his 1953 publication entitled "Explanation of the Brightness and Color of the Sky, Particularly the Twilight Sky", Edward Olson Hulburt (1890–1982) demonstrated that for specific illumination and viewing conditions, Rayleigh scattering plays only a second-order role for the blue colour of the sky. Based on simplified radiative transfer simulations and single-scattering approximation, Hulburt (1953) concluded that for a solar zenith angle (SZA) of 90. only 1/3 of the blue colour of the sky of the zenith is caused by Rayleigh scattering and 2/3 by absorption of solar radiation in the Chappuis bands of O3.

- Answer:** Graduate Education
- Justification:** This explanation includes technical terms like "Rayleigh Scattering" and references research papers that is intended for audience that are specialists in this domain (Physics / Atmospheric Science).

(b) Example for perceived background match user study: We selected 2 questions and 5 explanations to clarify the task details.

Figure 9: Overview of the perceived background match user study.





## Feedback Difference by Aspect

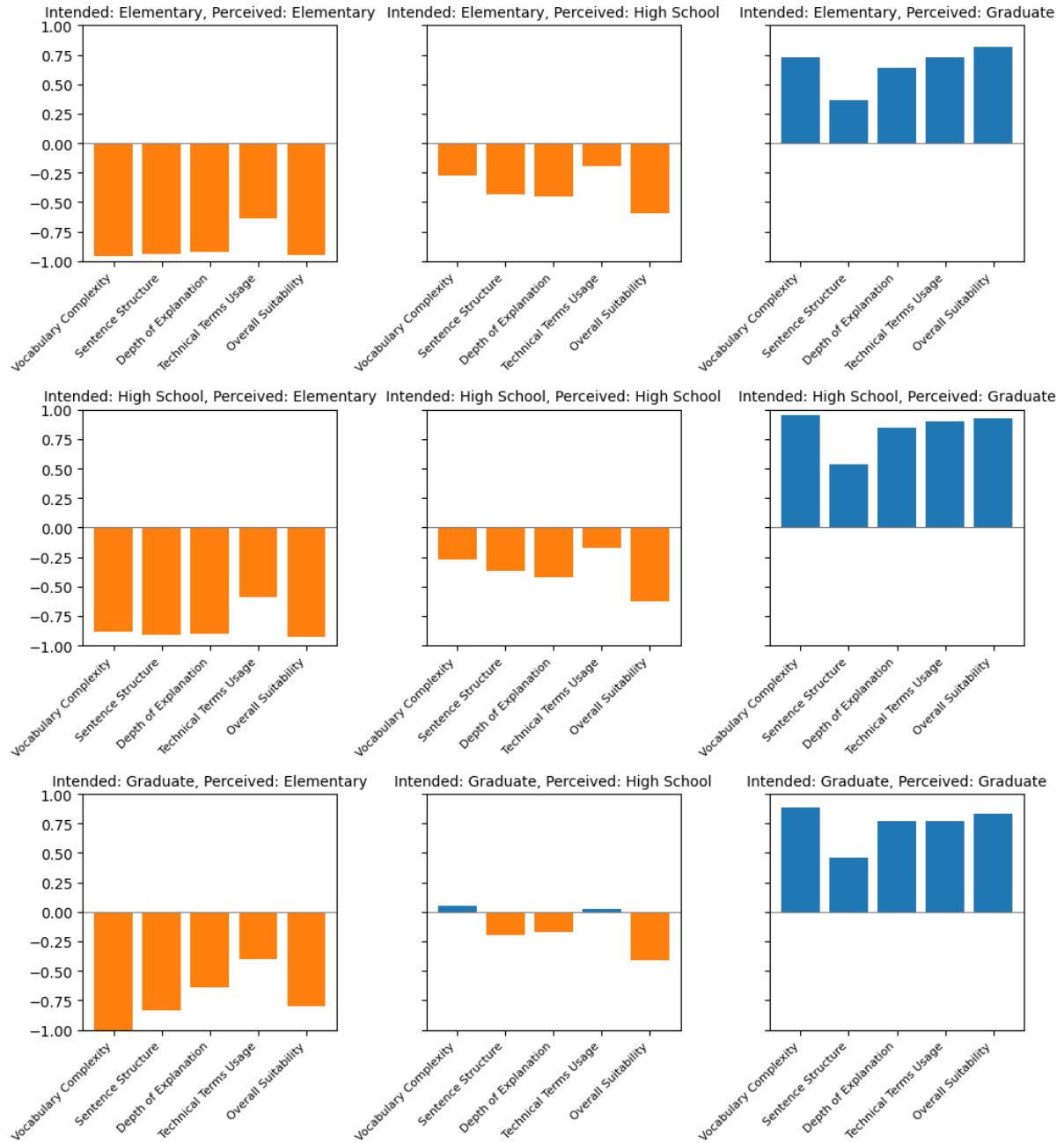


Figure 11: The normalized differences (positive minus negative counts, divided by total feedback) to our five predefined aspects (Vocabulary Complexity, Sentence Structure, Depth of Explanation, Technical Terms Usage, Overall Suitability). Each subplot corresponds to a specific combination of the intended (ground truth) and majority perceived educational levels (Elementary, High School, Graduate). A positive bar indicates that feedback leaned toward more complex/advanced language (e.g., “complex,” “in-depth,” “technical”), whereas a negative bar suggests simpler characteristics.

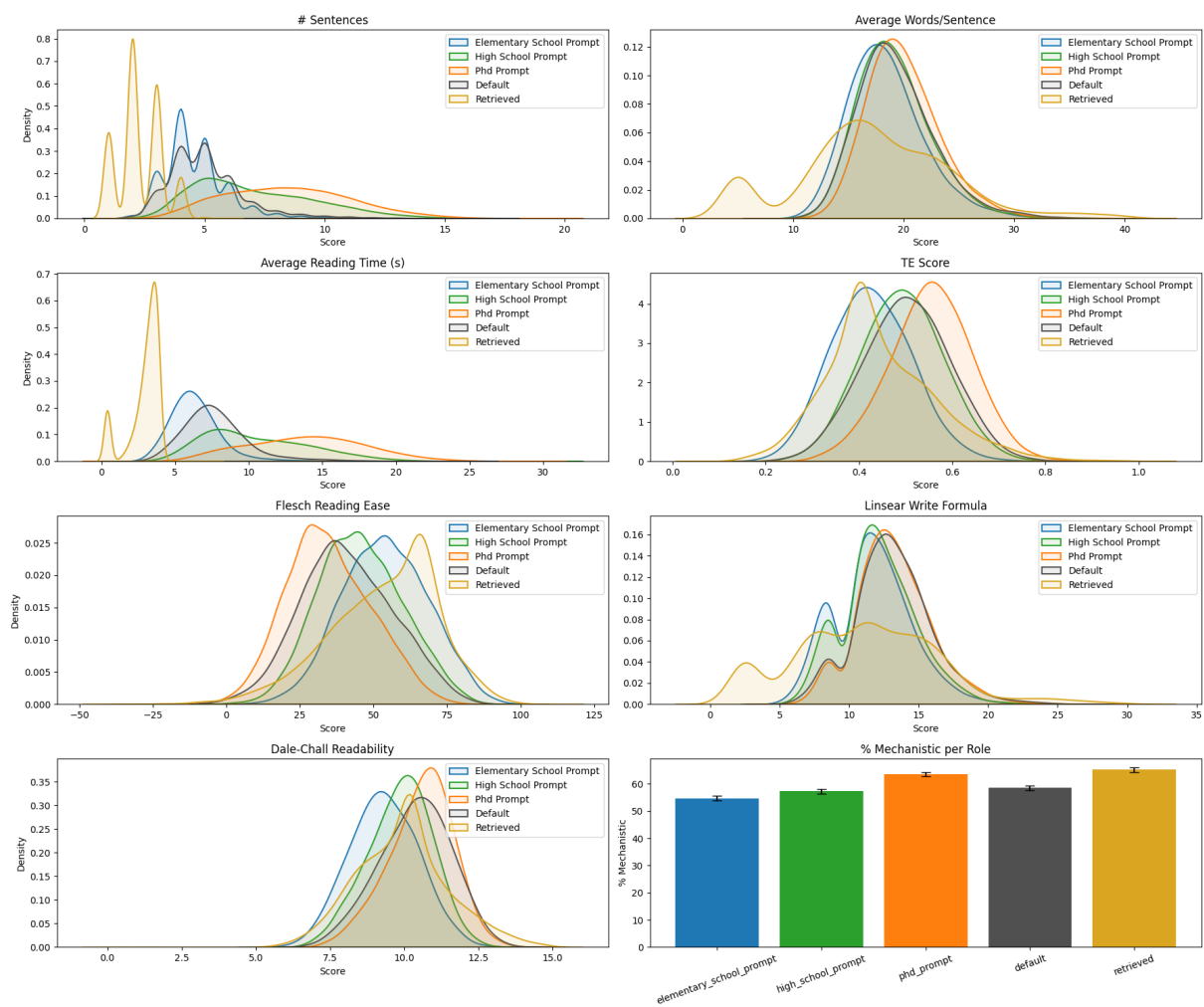


Figure 12: Overall distribution of readability scores with explanations generated by GPT-4

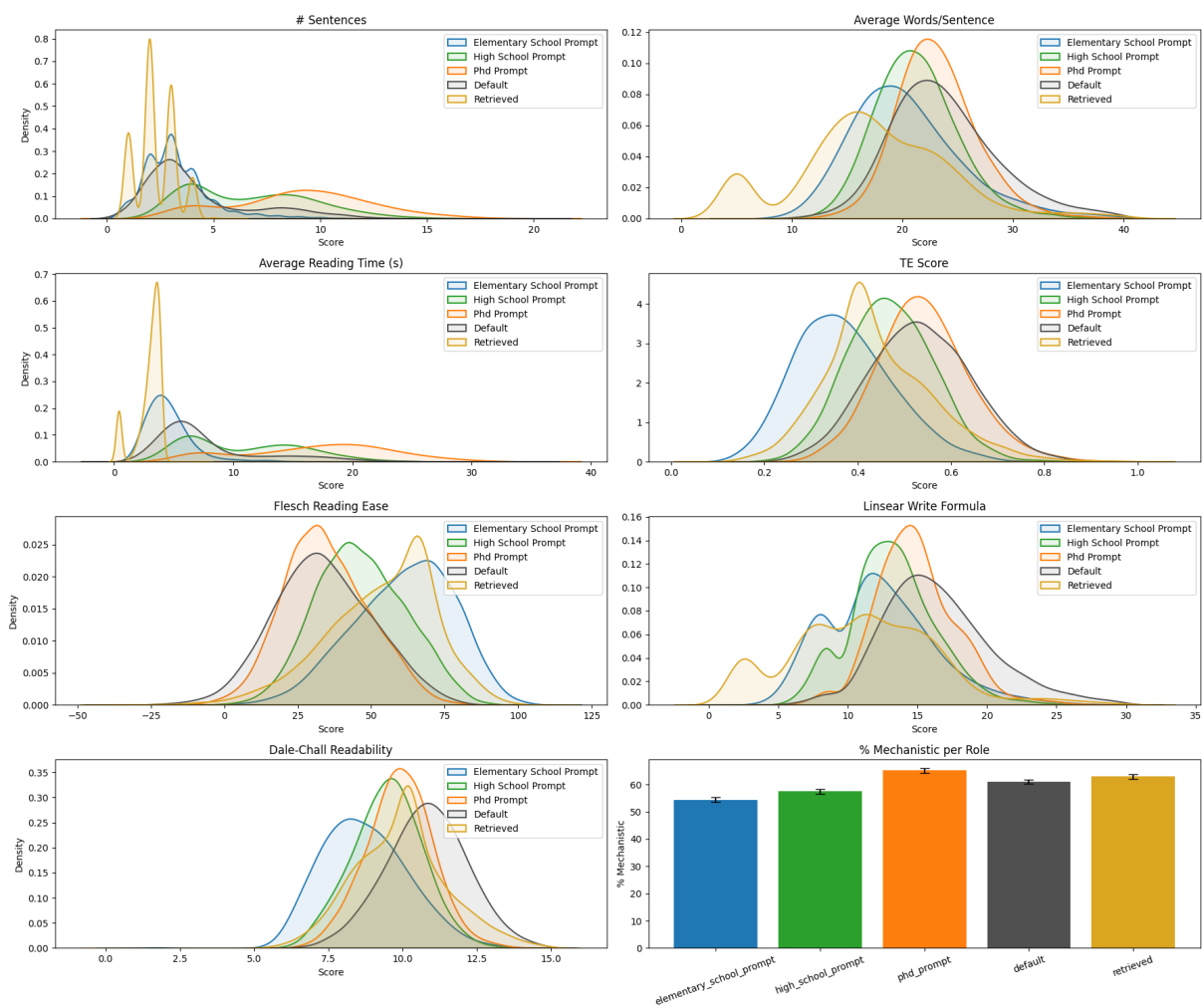


Figure 13: Overall distribution of readability scores with explanations generated by Llama-3.2-3B-Instruct.

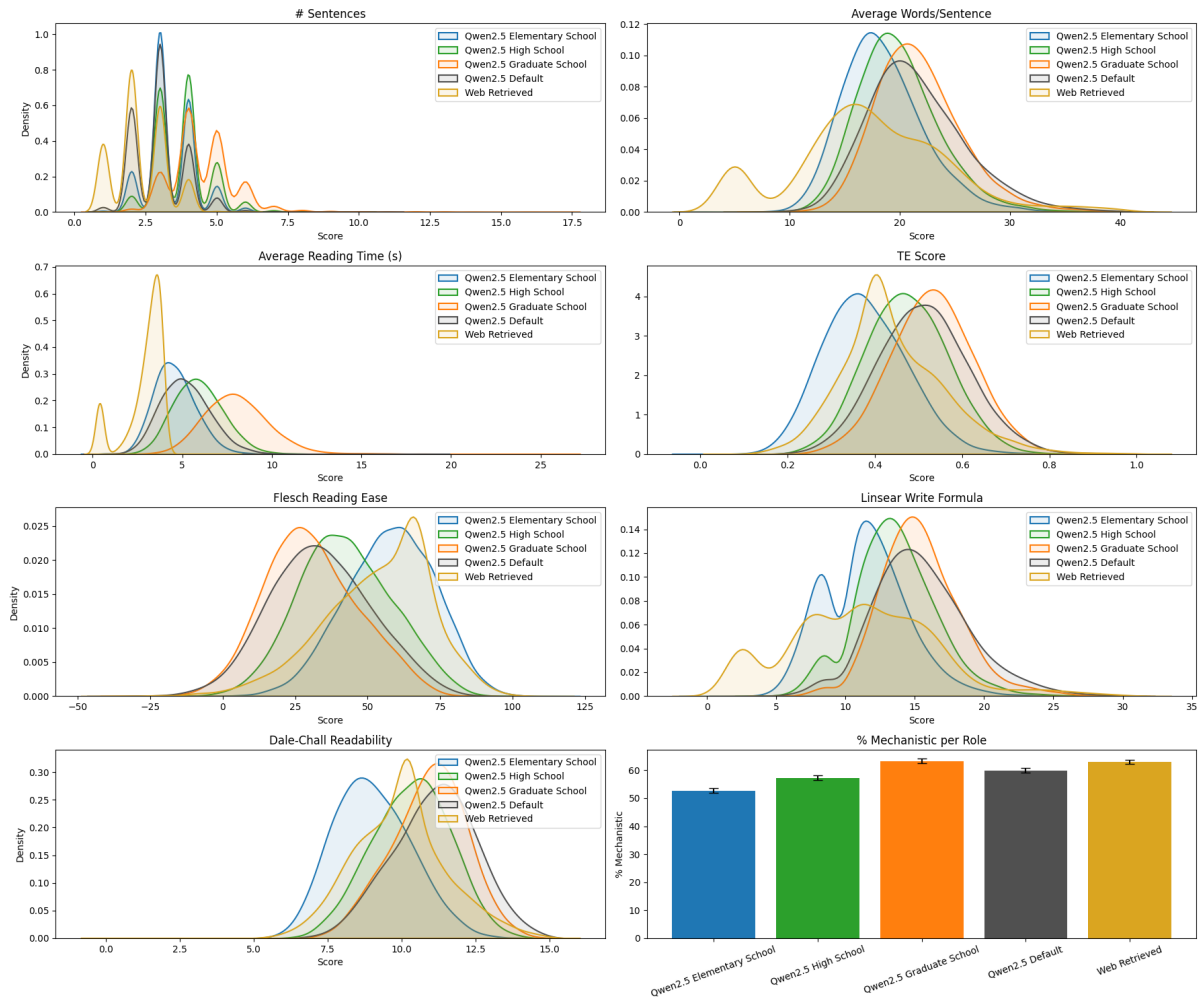


Figure 14: Overall distribution of readability scores with explanations generated by Qwen 2.5 14B Instruct.



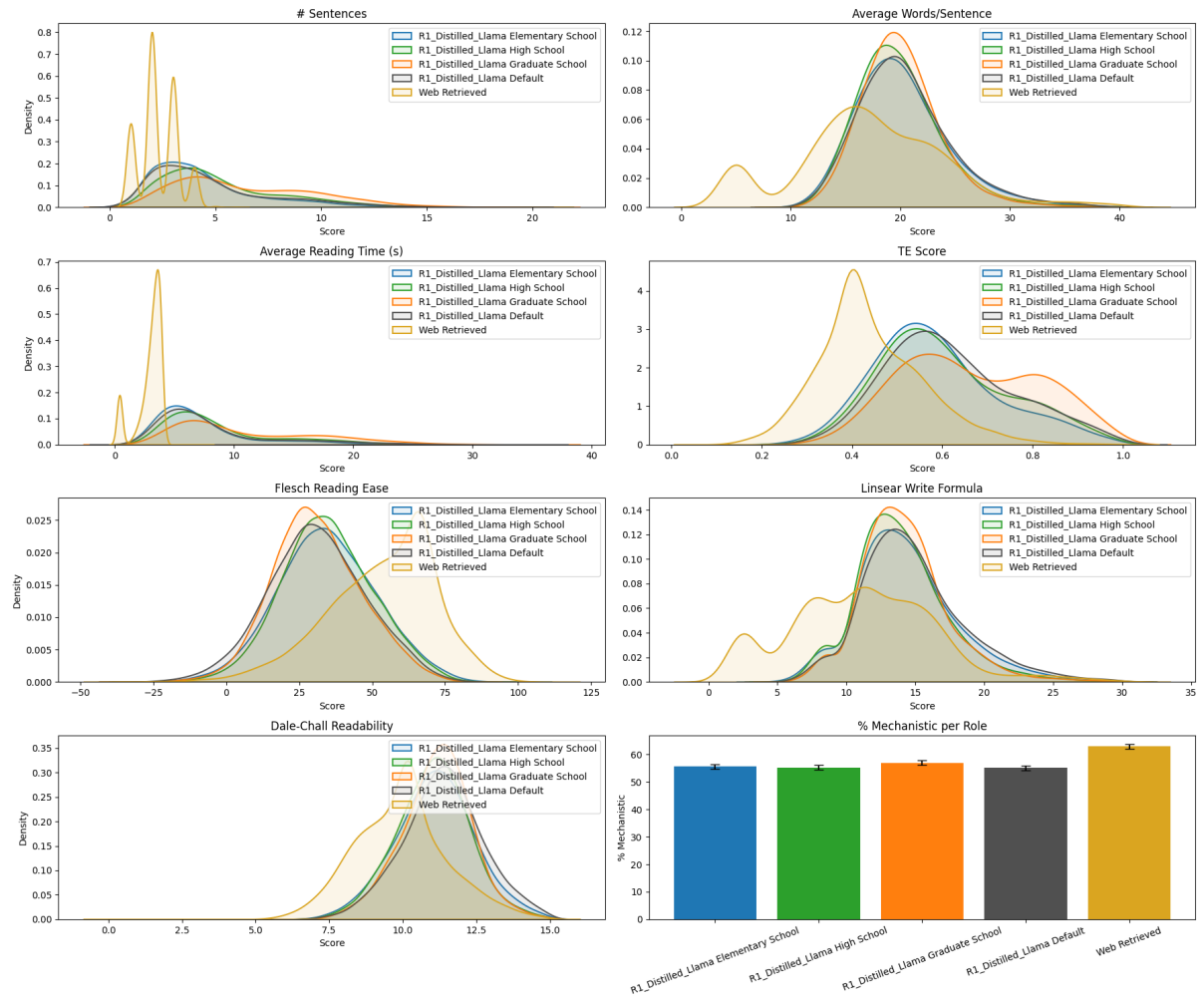


Figure 15: Overall distribution of readability scores with explanations generated by DeepSeek R1 Distill Llama 8B.

Question	Best Reddit Question	Match ELI5	Reddit ELI5 link	Similarity Score	Intended -> Perceived Mapping
Why do carnivorous plants eat insects?	Why do carnivorous plants eat meat?		<a href="https://www.reddit.com/r/explainlikeimfive/comments/trx3rw/eli5_why_do_carnivorous_plants_eat_meat/">https://www.reddit.com/r/explainlikeimfive/comments/trx3rw/eli5_why_do_carnivorous_plants_eat_meat/</a>	0.83	Elementary -> Elementary , High School -> High School , Graduate -> Elementary
Why do we perceive different colors?	Why do we see colours other than red, green and blue?		<a href="https://www.reddit.com/r/explainlikeimfive/comments/rihdyj/eli5_why_do_we_see_colours_other_than_red_green/">https://www.reddit.com/r/explainlikeimfive/comments/rihdyj/eli5_why_do_we_see_colours_other_than_red_green/</a>	0.85	Elementary -> Elementary , High School -> Elementary , Graduate -> Elementary
Why do balloons pop when exposed to sharp objects?	Why do bubbles pop when otherwise undisturbed?		<a href="https://www.reddit.com/r/explainlikeimfive/comments/p7kowq/eli5_why_do_bubbles_pop_when_otherwise_undisturbed/">https://www.reddit.com/r/explainlikeimfive/comments/p7kowq/eli5_why_do_bubbles_pop_when_otherwise_undisturbed/</a>	0.55	Elementary -> Elementary , High School -> High School , Graduate -> High School
Why is it beneficial for an organism to adapt?	Why would viruses and bacteria ever try to harm their host?		<a href="https://www.reddit.com/r/explainlikeimfive/comments/1optwy/eli5_why_would_viruses_and_bacteria_ever_try_to/">https://www.reddit.com/r/explainlikeimfive/comments/1optwy/eli5_why_would_viruses_and_bacteria_ever_try_to/</a>	0.43	Elementary -> High School , High School -> High School , Graduate -> High School
Why do we mourn the loss of celebrity we never met?	What does this poem mean?		<a href="https://www.reddit.com/r/explainlikeimfive/comments/1902oc/what_does_this_poem_mean/">https://www.reddit.com/r/explainlikeimfive/comments/1902oc/what_does_this_poem_mean/</a>	0.16	Elementary -> Elementary , High School -> High School , Graduate -> Graduate
Why have "perennial topics" like love, death, and identity remained key themes in literature?	Why is Death Valley one of the hottest places on earth despite being far from the equator?		<a href="https://www.reddit.com/r/explainlikeimfive/comments/1hfcnv1/eli5_why_is_death_valley_one_of_the_hottest/">https://www.reddit.com/r/explainlikeimfive/comments/1hfcnv1/eli5_why_is_death_valley_one_of_the_hottest/</a>	0.12	Elementary -> High School , High School -> High School , Graduate -> Graduate

Table 17: Examples of ELI5 question matching with similarity scores. When a question has a highly similar variant in the ELI5 subreddit, there is a tendency for oversimplification for prompts originally intended for higher educational levels.