REPQA: EVALUATING READABILITY OF LARGE LANGUAGE MODELS IN PATIENT EDUCATION QUESTION ANSWERING

Weikang Qiu[1*†] **Tinglin Huang**[1*] **Ryan Rullo**[2] **Giselle OConnor**[2] **Yuchen Kuang**[4]

Ali Maatouk[1] **S. Raquel Ramos**[2,3] **Rex Ying**[1]

[1]Yale University, Department of Computer Science,
[2]Yale University, School of Nursing,
[3]Yale University, School of Public Health, [4]Northeastern University

ABSTRACT

Large language models (LLMs) have shown great promise in addressing complex medical and clinical problems. However, while most prior studies focus on improving accuracy and reasoning abilities, a significant bottleneck in developing effective healthcare agents lies in the readability of LLM-generated responses, specifically, their ability to answer patient education problems clearly to people without a medical background. In this work, we introduce REPQA[1], a benchmark designed to systematically evaluate the readability of LLMs in patient education question-answering (QA). REPQA comprises 533 expert-reviewed QA pairs sourced from 27 online resources, reflecting common concerns of lay users across 4 categories. REPQA incorporates a proxy multiple-choice QA task to directly assess the informativeness of generated responses, alongside two readability metrics. We present a comprehensive study of 25 LLMs on REPQA, evaluating both instruction-following (answering at requested reading levels) and readability understanding (recognizing the level of questions and texts). We find that current models fail to achieve target levels and struggle to identify appropriate reading levels, revealing a gap between raw reasoning ability and effective communication. We then compare four approaches for improving readability: standard prompting, chain-of-thought prompting, Group Relative Policy Optimization (GRPO), and a token-adapted GRPO. While readability-aware post-training substantially improves readability metrics, it often reduces QA accuracy due to over-simplification, exposing a significant readability–accuracy trade-off. These findings point toward methods for building more user-centered public health agents.

1 INTRODUCTION

Benefiting from extensive training on large-scale corpora, large language models (LLMs) have demonstrated remarkable reasoning capabilities in solving complex problems OpenAI et al. (2024a); DeepSeek-AI et al. (2024); Team et al. (2024b); Qiu et al. (2025); Zhao et al. (2023). One of the most impactful application areas is healthcare, which demands strong domain-specific knowledge and a high reliability standard Singhal et al. (2023); Thirunavukarasu et al. (2023); Nazi & Peng (2024). Previous studies have introduced benchmarks Yao et al. (2024); Jin et al. (2019); Jiang et al. (2024); Wang et al. (2024b); Jin et al. (2019) targeting different levels of medical reasoning tasks.

Despite their promising performance, the **readability** of LLM-generated responses, an essential factor for patient education, i.e., communicating health information to laypeople, has been largely

*Equal contribution.
†Correspondence to weikang.qiu@yale.edu
[1]Code and dataset available at https://anonymous.4open.science/r/RepQA-45A8/

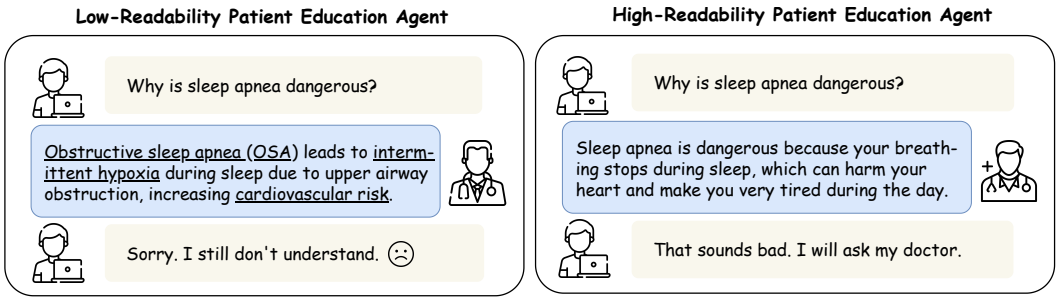

Figure 1: Two examples of user interactions with low- and high-readability patient education agents. Medical jargon is underscored in the responses generated by the low-readability agent.

Table 1: The comparison between REPQA and the existing healthcare-related benchmark datasets.

Dataset	Answer Format	Sources of Questions	Evaluation Metrics	Target Audience
MedQA Jin et al. (2021)	Multiple Choice	Medical Examination	Accuracy	Medical Professionals
MedMCQA Pal et al. (2022)	Multiple Choice	Medical Examination	Accuracy	Medical Professionals
MMLU Hendrycks et al. (2020)	Multiple Choice	Medical Examination	Accuracy	Academic
PubMedQA Jin et al. (2019)	Multiple Choice	PubMed Abstract	Accuracy	Clinical Researchers
LiveQA Abacha et al. (2017)	Long Answer	National Library of Medicine	Human Evaluation	Consumers
MedicationQA Abacha et al. (2019)	Long Answer	MedlinePlus & DailyMed	Human Evaluation	Medicine Consumers
JAMA & Medbullets Chen et al. (2025)	Long Answer & Multiple Choice	JAMA Challenge & USMLE	NLG Metrics & Accuracy	Medical Professionals
REPQA	Long Answer & Multiple Choice	Public Health FAQs & Experts	Readability & Accuracy	Lay Users / Patients

overlooked and remains systematically underexplored. As shown in Figure 1, users without a medical professional background may find low-readability responses filled with medical jargon difficult to comprehend, whereas a plain-language answer can convey the same essential information more accessibly. Moreover, state-of-the-art LLMs are increasingly optimized for complex reasoning Shao et al. (2024), raising the question of whether these gains come at the expense of readability. More advanced reasoning may lead to more technical responses that are harder for lay users to understand.

In this study, we introduce **REPQA** to conduct a comprehensive investigation into the **RE**adability of LLMs in **P**atient **E**ducation question-**A**nswering (QA) tasks. A key strength of REPQA lies in its expert-grounded foundation: it comprises of high-quality test set with 533 QA pairs curated from 27 online public health resources, all reviewed by professionals in nursing, public health, and clinical communication. Different from prior related benchmarks (Table 1), REPQA focuses on frequently asked public health questions from lay users or patients, and introduces a novel proxy multiple-choice QA that directly evaluates the accuracy of generated responses, rather than the underlying factual knowledge of the LLMs. Besides, we collect a large-scale corpus of 36,060 public health questions to serve as a post-training dataset, supporting the exploration of readability-enhancement strategies.

We evaluate 25 LLMs on REPQA using two readability metrics (Flesch-Kincaid grade level Kincaid et al. (1975) and professional score) and one accuracy metric based on our proxy multiple-choice QA. Our study targets three capabilities: (i) answering at a specified reading level, (ii) inferring the appropriate level for a question, and (iii) using readability as a reward signal. We find that current models mostly miss target levels and poorly identify appropriate levels, limiting their applicability for public-health agents. We then benchmark four readability-improvement strategies: two test-time methods (standard prompting Ribeiro et al. (2023); Malik et al. (2024), chain-of-thought Wei et al. (2022)) and two post-training methods (GRPO Shao et al. (2024) and a token-adapted GRPO). Readability-aware post-training yields large gains in readability metrics but reduces QA accuracy due to over-simplification, underscoring a difficult readability–accuracy trade-off. These findings offer insights for future work on developing trustworthy healthcare agents and may be generalized to other educational domains Wang et al. (2024a).

In summary, our contributions are listed as follows:

- We introduce REPQA, a novel benchmark for patient education QA aimed at lay users, with two readability metrics (Flesch–Kincaid grade and professional-term proportion) and a proxy multiple-choice QA protocol for accuracy.

- We conduct a comprehensive study of 25 LLMs on REPQA, revealing that most models struggle to (i) generate at requested reading levels and (ii) identify an appropriate level for a given question.
- We explore four existing strategies to improve LLM readability and propose a novel method, the token-adapted GRPO method. Readability improves substantially, but often at an accuracy cost, highlighting a challenging readability–accuracy trade-off.

2 RELATED WORK

Evaluation of LLM on healthcare QA To evaluate the capabilities of LLMs in solving healthcare and medical problems, multiple benchmarks have been curated from diverse sources and presented in various formats Singhal et al. (2023); Liévin et al. (2024); Thirunavukarasu et al. (2023). For instance, MedQA Jin et al. (2021), MedMCQA Pal et al. (2022), MMLU Hendrycks et al. (2020); Wang et al. (2024b), and PubMedQA Jin et al. (2019) contain multiple-choice questions sourced from medical licensing examinations or PubMed articles, while LiveQA Abacha et al. (2017) and MedicationQA Abacha et al. (2019) focus on long-form question answering for medicine-related problems. Recent efforts have shifted toward more complex clinical tasks requiring multi-step reasoning and decision-making Yao et al. (2024); Chen et al. (2025); Nori et al. (2024); Saab et al. (2024), as well as incorporating additional modalities such as electronic health records Ou et al. (2025); Bardhan et al. (2022); Kweon et al. (2024), radiology Zuo et al. (2025); Liu et al. (2021), and pathology He et al. (2020); Lu et al. (2024). Besides, several studies explore the reliability of LLMs with retrieval augmented generation Xiong et al. (2024). Our proposed benchmark, REPQA, is the first to comprehensively evaluate the readability of LLM-generated responses to healthcare-related questions, using a dataset composed of publicly available medical queries from lay users, rather than professional exam questions or specialized clinical tasks.

Readability control in LLMs There are three main strategies for controlling the complexity or reading level of text generated by large language models. (1) Prompt- or instruction-based methods modulate output difficulty by varying instructions or providing exemplars Tran et al. (2024); Malik et al. (2024); Barayan et al. (2024); Pu & Demberg (2023); (2) Fine-tuning-based methods, such as Chi et al. (2023) fine-tuned language model on same-level paraphrasing datasets, and Ribeiro et al. (2023) employed PPO Schulman et al. (2017) with the measured reading level Kincaid et al. (1975) as a reward signal to optimize the models; (3) Decoding-based methods incorporate constraints during text generation Ribeiro et al. (2023); Luo et al. (2022). Several studies investigated the readability of LLMs' generated responses regarding certain cancers Hershenhouse et al. (2024); Gencer (2024) or the patient education handouts Swisher et al. (2024). While most prior studies focus on text summarization, REPQA specifically targets the readability of LLM in patient education QA tasks.

3 REPQA BENCHMARK

In this section, we introduce REPQA, which comprises a high-quality collection of patient education QA pairs with corresponding multiple-choice questions for evaluation, as well as a large-scale set of questions for training. We detail the dataset curation pipeline in Section 3.1, outline the evaluation metrics in Section 3.2, and describe the studied problems in Section 3.3.

3.1 DATASET CURATION

Expert-Grounded Patient Education QA We collected a diverse set of publicly available healthcare-related QA pairs from 27 authoritative online resources covering a broad range of public health topics, including general health, cardiovascular disease, cholesterol, diabetes, HIV, diet, exercise, alcohol, smoking, sleep, mental health, sexual health, and health equity. Sources were selected for their accessibility to lay audiences and their coverage of frequently asked questions by the general public. All the links can be found in Appendix A.1.

For the evaluation dataset, domain experts further manually reviewed and filtered out QA pairs containing overly technical content, following national, authoritative, evidence-based research and clinical guidelines Doody & Noonan (2016); Chiang-Hanisko et al. (2016); Corry et al. (2013). Questions are phrased in plain language and do not include user profiles. A detailed description of

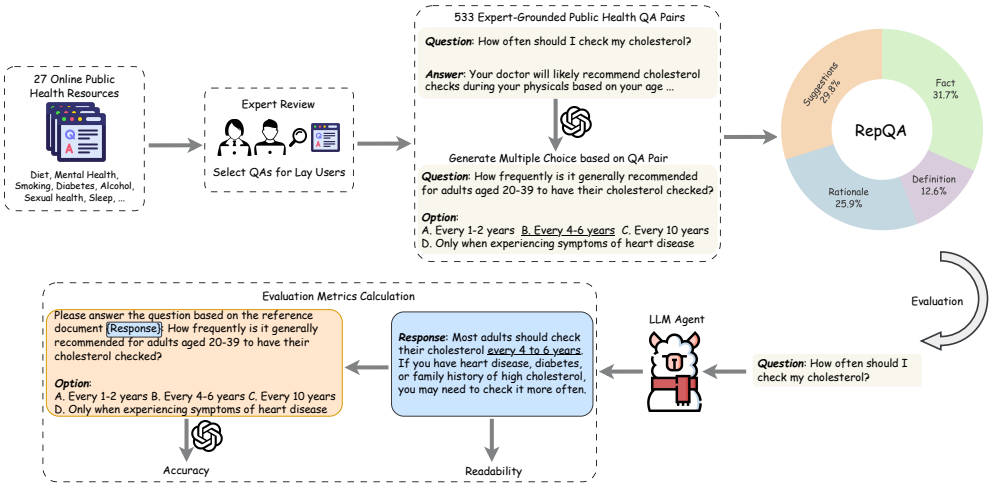

Figure 2: Overview of the REPQA dataset construction and evaluation pipeline. The dataset is curated through expert review and multiple-choice question generation, while the evaluation involves readability assessment and accuracy measurement.

Table 2: Examples and distribution of question categories in REPQA. Each question is accompanied by a corresponding reference answer.

Category	Example & Reference Answer	Proportion
Suggestion	**Q:** How can I quit smoking? **A:** Commit to a quit day, choose a method for quitting (cold turkey or weaning down), talk to your doctor about medications to help quit, make a plan for when you quit on how to resist temptations, remove triggers, and have replacement activities, and then quit on your quit day!	29.8%
Fact	**Q:** What are the chemicals in cigarettes? **A:**There are over 7,000 chemicals in cigarettes including arsenic, formaldehyde, tar, nicotine, and carbon monoxide, all of which are dangerous to your health.	31.7%
Definition	**Q:** What is the Mediterranean Diet? **A:** A Mediterranean diet is considered a healthy option as it emphasizes fruits and vegetables, whole grains, beans and legumes, low-fat dairy products, fish and poultry, and vegetable oils. It also limits sugars, processed foods, and fatty meats.	12.6%
Rationale	**Q:** Why is blood pressure called a "silent killer"? **A:** High blood pressure is sometimes called a silent killer because it can have no physical symptoms; however, the high blood pressure can damage arteries risking serious health events like heart attack or stroke.	25.9%

the curation pipeline and the reviewing team's composition is provided in Appendix A.2. As shown in Table 2, the evaluation dataset includes 533 QA pairs with the following categories:

- *Suggestion*: Questions seeking advice or recommended actions. These responses must be highly readable to ensure that users can easily follow health recommendations without ambiguity.
- *Fact*: Questions requesting objective, verifiable information. Presenting facts in lay-accessible language is essential for avoiding misinterpretation.
- *Definition*: Questions asking for explanations of medical terms or conditions. Definitions require especially high readability, as they often introduce unfamiliar terminology.
- *Rationale*: Questions seeking reasons or justifications behind medical guidance. These responses must balance clarity with accuracy in plain language to build user understanding and trust.

Proxy Multiple Choice QA Each question in the final evaluation dataset is paired with a reference answer. To assess the accuracy of generated responses, prior studies have either employed LLMs as judges to rate the relevance of a response against the reference answer Gu et al. (2024), or directly adopted multiple-choice questions Singhal et al. (2023); Liévin et al. (2024). However, the former approach raises concerns about objectivity, especially in the healthcare domain, while the latter assesses the model's direct knowledge rather than the accuracy of its generated explanations.

Table 3: Examples and distribution of question categories in the collected training datasets. A single question can be categorized into multiple categories (e.g., Diet and Weight: How do dietary habits influence obesity according to research findings?)

Category	Example	Proportion
Sleep	How does sleep apnea during pregnancy potentially affect newborns?	8.1%
Exercise	What are some effective strategies for improving family health through physical activities?	6.3%
Smoking	What are some effective strategies individuals can use to quit smoking and support a smoke-free lifestyle?	4.5%
Weight	How is body mass index (BMI) calculated, and what factors can affect its accuracy?	7.2%
Diet	How can someone maintain a healthy diet while dining out?	10.3%
HIV Care	How can gay men effectively manage their risk of HIV?	14.4%
Mental Health	How can parents support the mental health of their children?	12.6%
Cardiovascular	What is the primary difference between cardiac arrest and a heart attack?	18.7%
Diabetes	How do BMI categories relate to the risk of type 2 diabetes across different age groups?	5.2%
Cholesterol	What is the significance of LDL and HDL cholesterol in relation to heart health?	5.1%
SOGI	How can schools and communities support LGBTQI+ youth to create a safer environment?	12.4%
Others	What are the potential benefits of utilizing specialized healthcare models in patient treatment?	33.7%

To bridge this gap, we propose a proxy multiple-choice selection task, where a generated answer serves as a reference document for an `LLM-Critic` tasked with selecting the correct option, as illustrated in Figure 2. The multiple-choice QA is generated by an LLM (`LLM-Proposer`) using each question and its reference answer. Higher accuracy suggests that the response captures the essential information, making it a reliable proxy for factual correctness.

In practice, we observed that the initial questions generated by the `LLM-Proposer` were often too easy, and that the `LLM-Critic` simply relied on its own knowledge rather than the provided context. To overcome these issues, we integrate the `LLM-Critic` into the problem generation process, using it as a feedback mechanism to iteratively refine the multiple-choice questions, ensuring that they are non-trivial and that their solutions rely on the provided context. In practice, we use GPT-4.1 as the `LLM-Proposer` and GPT-4o-mini as the `LLM-Critic`. The complete algorithm is detailed in Algo. 1, and the prompts for both LLMs are included in Appendix B.2.

Large Scale Healthcare Questions In addition to the evaluation QA pairs, we curated a large-scale collection consisting of 36,060 public health questions. These queries span an extensive range of public health topics, including sleep, exercise, smoking, weight, diet, HIV care, mental health, cardiovascular, diabetes, cholesterol, and SOGI, from 55 publicly available resources (Appendix A.1). The data sources were manually selected by nursing and public health professionals, following the same guidelines used in curating the evaluation dataset. Summary statistics and representative examples for each category are presented in Table 3. Notably, this dataset includes only questions without reference answers, and allows the exploration of different training strategies, such as GRPO Shao et al. (2024), with readability metrics serving as verifiable rewards.

3.2 Evaluation Metrics

Flesch-Kincaid Grade Level We use the Flesch-Kincaid grade level Kincaid et al. (1975) as one of our readability metrics, which is an established standard for assessing reading difficulty, ranging from kindergarten to post-graduate levels. This metric enables fine-grained analysis of language complexity across model-generated responses. Given a text, the calculation is based on the average number of syllables in a word and the average number of words in a sentence:

$$\pi = 0.39 \left(\frac{\text{Total Words}}{\text{Total Sentences}} \right) + 11.8 \left(\frac{\text{Total Syllables}}{\text{Total Words}} \right) - 15.59 \tag{1}$$

The intuition is that sentences with many words or words with numerous syllables are generally more difficult to read and understand, corresponding to a higher Flesch-Kincaid grade level. The range of the grade levels can be categorized into different school levels, such as 6-9 corresponds to the middle school level. More details can be found in Appendix C.1.

Professional Score During communication between the agent and lay users without a medical background, the frequency of professional or technical terms can significantly affect the users' reading experience. In light of this, we incorporate the proportion of health-related professional terms in the

text as one of our readability metrics, reflecting the degree of technicality in a response:

$$\rho = 100\% \cdot \frac{\text{Total Professional Terms}}{\text{Total Words}} \tag{2}$$

We compiled a list of professional terms from the Harvard Medical Dictionary[2], resulting in a vocabulary of 2,058 terms. The proposed professional level metric complements the Flesch-Kincaid Grade Level by incorporating domain-specific terminology, which may affect comprehension for lay audiences but is not captured by traditional readability formulas.

Accuracy via Proxy QA As described in Section 3.1, we propose using a multiple-choice selection task as a proxy for evaluating the factual accuracy of generated answers. Specifically, the generated answer is used as contextual input to guide an LLM-based critic in selecting the correct option from a set of choices. Accuracy is the proportion of cases where the selected option matches the gold label, averaged over the evaluation set. Higher accuracy indicates that the generated answers convey information more accurately and effectively. An ideal healthcare agent should produce responses that are both easy to understand and highly accurate.

3.3 Investigation on LLMs' Readability

Using our curated RepQA, we study LLMs' ability to adapt, assess, and optimize readability. The prompts can be found in Appendix B.2. We pose three research questions:

RQ1: Can LLMs generate responses at a specified reading level? People with different educational backgrounds have varying health literacy. Therefore, healthcare agents should adapt the readability of their answers to a requested level (e.g., 5th grade or middle school). We examine whether LLMs follow explicit readability instructions without sacrificing accuracy. We compare each response's Flesch–Kincaid grade to the requested level and assess accuracy via proxy QA tasks. In addition, we evaluate LLM's readability understanding capability: LLMs are asked to predict the reading level of given responses, and we quantify their accuracy against the Flesch–Kincaid grade.

RQ2: Can LLMs determine what readability is needed for a given question? Oversimplifying medical content can omit qualifiers and lose information, while overly technical language reduces accessibility. Thus, selecting an appropriate reading level for each question is critical. We investigate whether an LLM can infer the target level from the question alone (without user metadata), balancing clarity and accuracy. Concretely, we extend each question in RepQA into three variants that differ in required medical knowledge, and evaluate the accuracy of the models' predictions.

RQ3: Can readability serve as a reward signal? Reinforcement learning has shown promise for steering LLMs toward preferred behaviors. We study whether readability can act as a useful reward to drive responses toward a requested level without sacrificing content quality. Specifically, the reward function is calculated based on the weighted combination of Flesch-Kincaid grade level and the frequency of professional terms:

$$r = -1 \cdot (\alpha \cdot |\pi - 6| + \beta \cdot \rho + \gamma \cdot \nu) \tag{3}$$

where α, β and γ are hyperparameters used to balance components. $\nu = 1$ if the response is considered trivial and 0 otherwise. A Flesch-Kincaid grade level (π) closer to 6, i.e., the middle school reading level, and a lower professional term proportion (ρ) result in a higher reward. We adopt Group Relative Policy Optimization (GRPO) and a token-adapted variant TA-GRPO that redistributes the sequence-level reward across tokens using weights derived from the occurrence of professional terms. More details regarding TA-GRPO can be found in Appendix C.2.

4 Experiments

4.1 Experimental Setup

We present a comprehensive evaluation of 25 LLMs, including 12 proprietary models and 13 open-source models. For proprietary models, we consider GPT-4 series OpenAI et al. (2024c), GPT-4o

[2] https://www.health.harvard.edu/a-through-c

Table 4: Benchmarking both proprietary and open-source LLMs. Metrics: Grade - Flesch-Kincaid grade level; Prof. - Professional score; Acc. - Accuracy via Proxy QA. More readability metrics are included in Appendix E.

Model	Suggestion			Fact			Definition			Rationale		
	Grade	Prof.	Acc.	Grade	Prof.	Acc.	Grade	Prof.	Acc.	Grade	Prof.	Acc.
Proprietary Models												
o4-mini	3.10	2.23%	80.50%	4.70	3.51%	82.84%	3.79	3.74%	86.57%	3.46	3.92%	86.96%
o3-mini	3.58	2.22%	74.21%	4.17	3.26%	75.74%	4.56	3.95%	80.60%	3.99	4.16%	84.06%
GPT-4o	4.30	2.25%	74.21%	5.27	3.22%	79.29%	**5.61**	3.73%	**89.55%**	5.16	3.58%	83.33%
GPT-4o-mini	3.86	2.26%	73.58%	4.63	2.99%	73.96%	5.07	**3.27%**	85.07%	4.50	3.50%	79.71%
GPT-4.1	3.65	2.22%	78.62%	4.66	3.24%	73.37%	4.70	3.77%	86.57%	4.44	4.16%	**88.41%**
GPT-4	4.07	2.45%	75.47%	4.79	3.19%	70.41%	4.87	3.48%	85.07%	4.30	3.61%	78.26%
GPT-4-turbo	4.61	2.19%	72.33%	**5.51**	2.91%	65.09%	5.39	3.48%	76.12%	5.33	3.60%	78.99%
Claude-3.7-Sonnet	8.95	2.49%	**81.32%**	9.43	3.41%	**83.43%**	7.68	3.65%	85.07%	7.72	3.95%	84.78%
Claude-3.5-Sonnet	**6.68**	2.52%	77.36%	6.79	3.17%	76.92%	6.62	3.74%	86.57%	**5.87**	**3.39%**	84.78%
Gemini-1.5-pro	4.37	2.36%	78.62%	4.72	**2.78%**	77.51%	5.15	4.14%	83.58%	4.93	3.63%	82.61%
Gemini-2.0-flash	3.60	2.33%	77.99%	4.34	3.10%	75.74%	4.38	3.43%	83.58%	4.33	3.69%	78.99%
Gemini-1.5-flash	4.02	**2.15%**	77.36%	4.37	2.86%	67.46%	4.25	3.48%	80.60%	4.38	3.80%	77.54%
Open-Source Models												
LLaMA3.1-8B	5.23	2.71%	72.96%	6.59	3.50%	74.56%	6.53	3.77%	86.57%	**5.97**	4.11%	82.61%
Qwen2.5-7B	3.57	2.32%	69.81%	5.01	3.23%	69.23%	4.90	3.74%	85.07%	4.41	3.64%	79.71%
BioMistral-7B	6.18	3.78%	80.50%	6.69	4.66%	65.68%	7.24	4.65%	83.58%	6.34	5.60%	77.54%
Phi-4	4.26	2.00%	72.96%	5.56	2.55%	**75.15%**	**5.97**	2.98%	82.09%	5.85	3.05%	83.33%
Yi-1.5-9B	7.81	3.14%	72.33%	8.09	3.69%	66.86%	8.51	3.80%	**88.06%**	8.34	4.55%	82.61%
Mistal-7B	6.49	3.12%	75.47%	6.94	3.85%	73.37%	6.87	4.04%	85.07%	7.38	4.96%	84.06%
InternLM3-8B	6.58	2.65%	77.36%	6.79	3.10%	72.78%	6.67	3.58%	86.57%	6.82	3.73%	80.43%
Gemma-3-12b	3.29	**1.11%**	73.58%	3.75	**1.65%**	73.96%	3.22	**1.99%**	80.60%	3.72	**2.25%**	81.88%
DeepSeek-V2-Lite	**5.92**	2.92%	74.21%	7.21	3.59%	69.23%	7.77	4.00%	83.58%	7.21	3.93%	78.99%
LLaMA3.1-70B	7.79	3.18%	**81.13%**	8.84	3.81%	74.56%	7.83	4.02%	86.57%	7.94	4.50%	**84.78%**
Qwen-2.5-14B	5.26	3.00%	77.99%	**5.92**	3.26%	71.01%	6.11	3.51%	**88.06%**	5.95	3.92%	81.88%
Qwen-2.5-32B	2.84	2.38%	61.64%	3.64	3.54%	66.27%	4.53	4.33%	77.61%	3.60	4.03%	75.36%
Qwen-2.5-72B	3.13	2.64%	71.07%	3.75	3.50%	72.78%	4.31	3.64%	82.09%	3.91	4.14%	78.26%
Target value	6	0 %	100 %	6	0 %	100 %	6	0 %	100 %	6	0 %	100 %

series OpenAI et al. (2024a), o series OpenAI et al. (2024b), Claude Anthropic (2024), Gemini Team et al. (2024a; 2025) For open source models, we consider LLaMA 3.1 series Grattafiori et al. (2024), Qwen series Team (2024), Phi-4 Abdin et al. (2024), Mistral Jiang et al. (2024), BioMistral Labrak et al. (2024), InternLM3 Cai et al. (2024), Gemma-2 Team et al. (2024b), DeepSeek DeepSeek-AI et al. (2024), Yi-1.5 AI et al. (2025). We use their instruction versions (or chat versions) by default.

To ensure a fair comparison, when benchmarking LLMs, for proprietary models, we use their official OpenAI-compatible APIs. For open-source models, we use vLLM's openai-compatible APIs. When benchmarking post-training methods, we use Huggingface's APIs.

4.2 Investigation on RQ1 and RQ2

Comprehensive study of readability and accuracy For RQ1 (Section 3.3), we explicitly set a target score, i.e., 6 for Flesch-Kincaid grade level and 0 for the professional score, and prompt it to answer the RepQA queries accordingly. Here, π=6 approximates a 5th grade level intended to be broadly accessible, and ρ=0 denotes the absence of domain-specific medical jargon. As shown in Table 4, most models struggle to hit the target, often overshooting complexity or, when simplifying, omitting qualifiers. Across the four categories, *definition* is the easiest, with an average accuracy of 84.06% and the lowest readability scores. More analysis of Table 4 could be found in Appendix D.

Calibration across representative models We evaluate four representative LLMs: two non-thinking models (LLaMA 3.1-8B, Qwen-3-30B) and three thinking models (DeepSeek-R1-Distill-Qwen, Qwen-3-30B-Thinking, GPT-4.1), focusing on their ability to match requested reading levels. We set four target levels, i.e., 5th grade, middle school, high school, and college, corresponding to Flesch–Kincaid grade levels of 6, 9, 12, and 15. As shown in Fig. 3(a,b), all models generally follow the intended direction (Spearman $\rho \approx 0.34 \sim 0.57$) but remain poorly calibrated: their Flesch–Kincaid grades consistently fall short of the targets, with larger deviations at higher levels. In short, while models adjust readability in a relative sense, they fail in absolute alignment.

Readability understanding of generated text We further test whether poor control stems from a weak readability understanding. We sample 200 generated responses per bin (Elementary, Middle School, High School, College; mapped from Flesch–Kincaid) and ask models to classify the bin. As shown in Fig. 3(c), accuracy peaks at Middle School, is moderate at Elementary, and collapses at

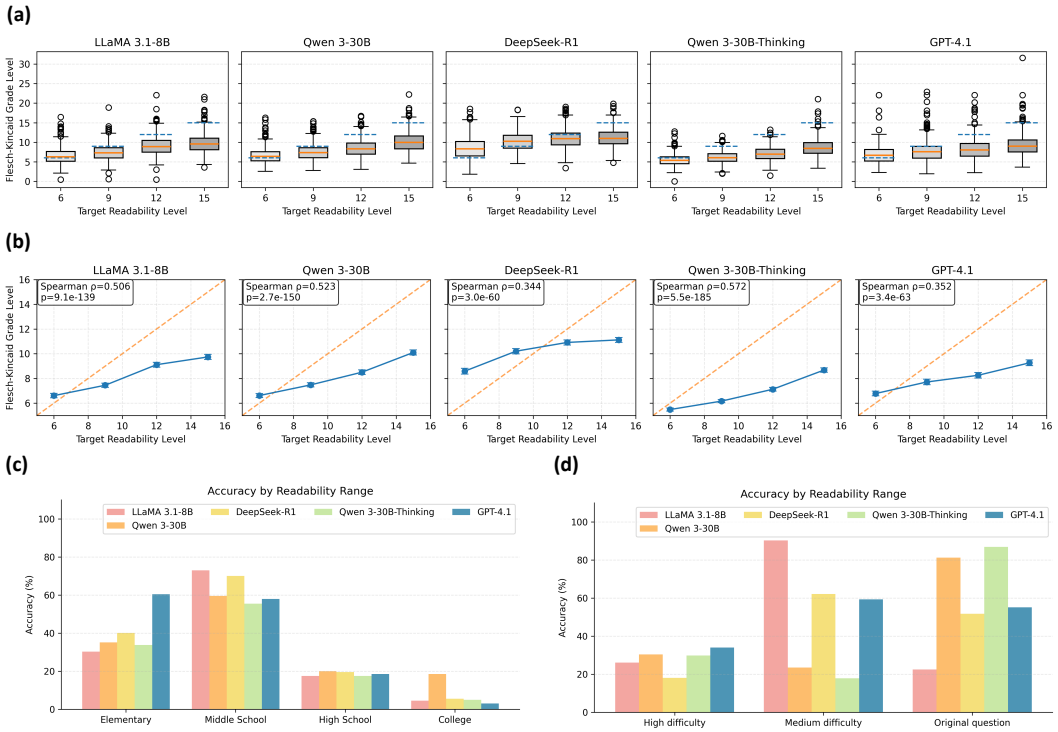

Figure 3: Readability control and understanding across four models (LLaMA 3.1–8B, Qwen-3-30B, DeepSeek-R1-Distill-Qwen, Qwen-3-30B-Thinking, GPT-4.1). **(a)** Achieved Flesch–Kincaid grades when instructed to write at target levels 6/9/12/15 (boxplots). Red horizontal lines mark the targets, and blue dashed lines mark the per-target mean grade. **(b)** Mean Flesch–Kincaid grade vs. target (error bars show 95% CI). **(c)** Accuracy of classifying generated responses into Flesch–Kincaid buckets (Elementary/Middle/High/College). **(d)** Accuracy when classifying Original/Medium/High counterparts of the same question.

High School and College. This systematic bias toward lower buckets mirrors the undershoot observed earlier, indicating that models struggle to recognize high-grade text, not only to produce it.

Sensitivity to required medical knowledge For RQ2, we study whether LLMs understand differences in required medical knowledge. For each question in REPQA, we prompt GPT-4o to generate two counterparts that vary only in required domain knowledge (Medium and High). Some examples can be found in Appendix A.3. Each model then classifies the variant. As shown in Fig. 3(d), performance is uneven across systems, for example, LLaMA 3.1-8B is relatively strong on Medium but weak on Original, whereas Qwen-3-30B-Thinking excels on Original yet remains low on High. Across models, detection of High difficulty is uniformly poor, indicating limited sensitivity to advanced domain knowledge and helping explain the tendency to default to overly simple outputs.

4.3 RQ3: READABILITY AS REWARD SIGNAL

We optimize readability using the reward in Eq. 3 within GRPO and a token-adapted variant (TA-GRPO) that redistributes the sequence-level reward to tokens in proportion to the occurrence of professional terms. We also include standard prompting and chain-of-thought (CoT) prompting as non-RL baselines. Results are reported in Table 5.

Relative to plain prompting, CoT produces less readable text with no meaningful accuracy gain, consistent with our benchmark's emphasis on retrieval and style control rather than step-by-step reasoning. Both GRPO and TA-GRPO improve readability over the original model; TA-GRPO yields the largest gains, reducing the grade level by 41.4% on average and the professional score by 73.3%. Moreover, TA-GRPO outperforms vanilla GRPO on professional readability (53.4% average reduction) while maintaining a comparable grade level (3.0% difference). We report QA accuracy alongside these readability metrics to quantify trade-offs. Despite substantial readability gains from

Table 5: Benchmarking different readability-enhancement strategies with different backbone LLMs.

Model	Suggestion			Fact			Definition			Rationale		
	Grade	Prof.	Acc.	Grade	Prof.	Acc.	Grade	Prof.	Acc.	Grade	Prof.	Acc.
LLaMA3.1-8B	7.64	4.10%	72.33%	8.52	4.69%	71.60%	8.74	5.18%	91.04%	8.94	6.74%	76.81%
+ Prompt	5.23	2.71%	72.96%	6.59	3.50%	74.56%	6.53	3.77%	86.57%	5.97	4.11%	82.61%
+ CoT	10.77	5.07%	79.25%	9.18	5.53%	72.78%	9.03	6.15%	86.57%	9.04	7.02%	81.89%
+ GRPO	4.86	3.12%	69.81%	5.57	3.25%	69.82%	6.87	4.34%	89.55%	5.81	4.49%	76.81%
+ TA-GRPO	5.39	1.37%	64.78%	5.89	1.95%	68.04%	5.65	2.23%	85.07%	6.05	2.34%	75.36%
Qwen2.5-7B	7.89	3.31%	85.53%	9.45	4.08%	79.28%	9.83	4.47%	92.54%	9.70	5.47%	86.23%
+ Lookahead	3.20	1.36%	70.45%	3.84	2.00%	68.09%	4.31	2.91%	91.67%	3.90	2.49%	81.48%
+ Prompt	3.57	2.32%	69.81%	5.01	3.23%	69.23%	4.90	3.74%	85.07%	4.41	3.64%	79.71%
+ CoT	5.09	4.20%	71.07%	5.39	6.63%	71.60%	6.57	7.06%	83.58%	5.44	6.98%	82.61%
+ GRPO	5.66	1.99%	76.73%	5.88	2.19%	73.37%	6.39	2.91%	91.04%	5.69	2.73%	83.33%
+ TA-GRPO	5.29	0.90%	75.47%	5.53	0.89%	76.33%	6.25	1.74%	91.04%	5.05	0.75%	82.61%
BioMistral-7B	8.68	4.00%	67.92%	9.76	4.60%	66.86%	11.45	5.56%	86.57%	10.22	6.14%	76.09%
+ Prompt	6.18	3.78%	80.50%	6.69	4.66%	65.68%	7.24	4.65%	83.58%	6.34	5.60%	77.54%
+ CoT	7.53	4.25%	69.81%	7.46	5.13%	68.05%	7.73	5.57%	86.57%	6.91	5.86%	71.74%
+ GRPO	4.53	1.89%	71.07%	4.56	1.89%	68.64%	4.69	2.54%	86.57%	4.52	2.40%	75.36%
+ TA-GRPO	4.23	0.81%	69.81%	4.73	0.89%	62.13%	5.51	1.86%	82.09%	4.64	0.89%	71.74%
Target value	6	0	100 %	6	0	100 %	6	0	100 %	6	0	100 %

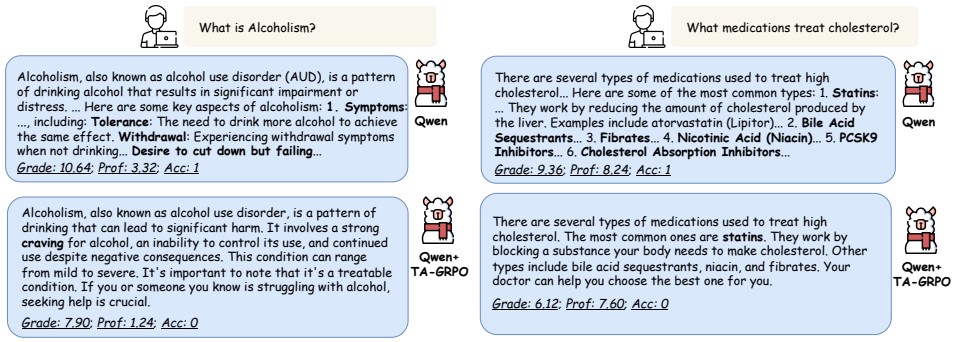

Figure 4: Two examples of improved readability but reduced accuracy.

GRPO/TA–GRPO, QA accuracy often declines, especially on *Suggestion* and *Fact*, highlighting the difficulty of simplifying without losing content.

4.4 Case Study

Table 5 shows that, although GRPO/TA–GRPO with a readability-based reward markedly improves readability, it often reduces QA accuracy. To illustrate this trade-off, Fig. 4 presents two cases where TA–GRPO yields simpler text but misses key facts relative to the base model. In particular, the TA–GRPO model tends to over-simplify, compressing lists and collapsing categories, so only a salient term is retained (e.g., *craving* for alcohol use disorder; *statins* for cholesterol drugs) while other classes (e.g., bile-acid sequestrants, fibrates, niacin, PCSK9 inhibitors) are merged into a single undifferentiated sentence, leading to information loss.

5 Conclusion

In this study, we examine a critical factor in developing effective patient education agents, i.e., the readability of LLM-generated responses, using a collection of expert-grounded patient education QA pairs, RepQA. RepQA comprises 533 QA pairs from 27 online resources, along with a large-scale collection of patient education questions used to explore the impact of various strategies for improving readability. Our evaluation is based on two readability metrics and one accuracy metric assessed via a proxy multiple-choice QA task. We evaluate 25 LLMs on RepQA and benchmark four readability-enhancement strategies across three backbone models. The results reveal a significant limitation of current LLMs: they struggle to follow targeted readability instructions and to recognize the reading levels of both questions and text. While some post-training methods improve readability, these gains often come at the expense of accuracy. Together, these findings provide practical insights for building more accessible and trustworthy LLM-powered healthcare agents. **Limitations** Our evaluation dataset primarily targets English-language content and lay users in the U.S. healthcare context. As such, the generalizability of our findings to other languages, cultures, or health systems remains limited.

REFERENCES

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*, pp. 1–12, 2017.

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pp. 25–29. IOS Press, 2019.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2025. URL https://arxiv.org/abs/2403.04652.

Anthropic. Introducing claude 3.5 sonnet, June 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2024-06-21.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. Analysing zero-shot readability-controlled sentence simplification. *arXiv preprint arXiv:2409.20246*, 2024.

Jayetri Bardhan, Anthony Colas, Kirk Roberts, and Daisy Zhe Wang. Drugehrqa: A question answering dataset on structured and unstructured electronic health records for medicine related queries. *arXiv preprint arXiv:2205.01290*, 2022.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3563–3599, 2025.

Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S Chang. Learning to paraphrase sentences to different complexity levels. *Transactions of the Association for Computational Linguistics*, 11:1332–1354, 2023.

Lenny Chiang-Hanisko, David Newman, Susan Dyess, Duangporn Piyakong, and Patricia Liehr. Guidance for using mixed methods design in nursing practice research. *Applied Nursing Research*, 31:1–5, 2016.

Margarita Corry, Mike Clarke, Alison E While, and Joan Lalor. Developing complex interventions for nursing: a critical review of key guidelines. *Journal of clinical nursing*, 22(17-18):2366–2386, 2013.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024. URL https://arxiv.org/abs/2405.04434.

Owen Doody and Maria Noonan. Nursing research ethics, guidance and application in practice. *British Journal of Nursing*, 25(14):803–807, 2016.

Adem Gencer. Readability analysis of chatgpt's responses on lung cancer. *Scientific Reports*, 14(1): 17234, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie,

Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi,

Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Jacob S Hershenhouse, Daniel Mokhtar, Michael B Eppler, Severin Rodler, Lorenzo Storino Ramacciotti, Conner Ganjavi, Brian Hom, Ryan J Davis, John Tran, Giorgio Ivan Russo, et al. Accuracy, readability, and understandability of large language models for prostate cancer information to the public. *Prostate Cancer and Prostatic Diseases*, pp. 1–6, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwang Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. *Advances in Neural Information Processing Systems*, 37: 124575–124611, 2024.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024. URL https://arxiv.org/abs/2402.10373.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions? *Patterns*, 5(3), 2024.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. Readability controllable biomedical document summarization. *arXiv preprint arXiv:2210.04705*, 2022.

Ali Malik, Stephen Mayhew, Chris Piech, and Klinton Bicknell. From tarzan to tolkien: Controlling the language proficiency level of llms for content generation. *arXiv preprint arXiv:2406.03030*, 2024.

G Harry Mc Laughlin. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646, 1969.

Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, pp. 57. MDPI, 2024.

Harsha Nori, Naoto Usuyama, Nicholas King, Scott Mayer McKinney, Xavier Fernandes, Sheng Zhang, and Eric Horvitz. From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond. *arXiv preprint arXiv:2411.03590*, 2024.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho

Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024a. URL https://arxiv.org/abs/2410.21276.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir

Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024b. URL https://arxiv.org/abs/2412.16720.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024c. URL https://arxiv.org/abs/2303.08774.

Justice Ou, Tinglin Huang, Yilun Zhao, Ziyang Yu, Peiqing Lu, and Rex Ying. Experience retrieval-augmentation with electronic health records enables accurate discharge qa. *arXiv preprint arXiv:2503.17933*, 2025.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Dongqi Pu and Vera Demberg. Chatgpt vs human-authored text: Insights into controllable text summarization and sentence style transfer. *arXiv preprint arXiv:2306.07799*, 2023.

Weikang Qiu, Zheng Huang, Haoyu Hu, Aosong Feng, Yujun Yan, and Rex Ying. Mindllm: A subject-agnostic and versatile model for fmri-to-text decoding. *arXiv preprint arXiv:2502.15786*, 2025.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. Generating summaries with controllable readability levels. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11669–11687, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.714. URL https://aclanthology.org/2023.emnlp-main.714/.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Austin R Swisher, Arthur W Wu, Gene C Liu, Matthew K Lee, Taylor R Carle, and Dennis M Tang. Enhancing health literacy: Evaluating the readability of patient handouts revised by chatgpt's large language model. *Otolaryngology–Head and Neck Surgery*, 171(6):1751–1757, 2024.

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

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew

Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying

Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman,

Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024a. URL https://arxiv.org/abs/2403.05530.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish

Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex

Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández

Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb,

Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda,

Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024b. URL `https://arxiv.org/abs/2408.00118`.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940, 2023.

Hieu Tran, Zonghai Yao, Lingxi Li, and Hong Yu. Readctrl: Personalizing text generation with readability-controlled instruction learning. *arXiv preprint arXiv:2406.09205*, 2024.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*, 2024a.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 6233–6251, 2024.

Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, et al. Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework. *arXiv preprint arXiv:2410.01553*, 2024.

Victor H Yngve. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466, 1960.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

A Dataset Setup

A.1 Online Public Resources

Table 6 lists all the public resources for curating high-quality question-answering pairs in the test set of RepQA. Table 7 lists all the online resources used for curating public health questions. Each link may be associated with multiple health-related categories.

Table 6: Public resources of Public Health for Curating QA Pairs.

Topic	Links
General Health	https://www.aarp.org/health/top-health-questions/
Cardiovascular	https://www.ohsu.edu/knight-cardiovascular-institute/frequently-asked-cardiovascular-questions https://www.multicare.org/vitals/answers-to-common-heart-questions/
Cholesterol	https://nyulangone.org/care-services/center-for-the-prevention-of-cardiovascular-disease/frequently-asked-questions-about-heart-health
Diabetes	https://diabetesaction.org/questions-general-information https://uihc.org/health-topics/diabetes-frequently-asked-questions https://www.nichd.nih.gov/health/topics/diabetes/more_information/other-faqs
HIV	https://www.unaids.org/en/frequently-asked-questions-about-hiv-and-aids
Diet	https://www.uclahealth.org/medical-services/weight-management/patient-resources/frequently-asked-questions https://www.cdc.gov/bmi/faq/index.html https://www.nutrition.gov/expert-q-a
Exercise	https://www.unk.edu/blog/2020/02/exercise-and-physical-activity-faqs.php
Alcohol	https://dmh.lacounty.gov/our-services/employment-education/education/alcohol-abuse-faq/
Smoking	https://resphealth.org/quitting-smoking-faqs/
Sleep	https://www.gwhospital.com/frequently-asked-questions-about-sleep https://medlineplus.gov/ency/quiz/000805_16.htm?quiz=1 https://www.mayoclinichealthsystem.org/hometown-health/speaking-of-health/8-common-sleep-study-questions
Mental health	https://www.wellnessinmind.org/frequently-asked-questions/ https://www.mountsinai.org/care/psychiatry/services/depression-anxiety-disorders/faq https://bbrfoundation.org/faq/frequently-asked-questions-about-depression https://wmich.edu/suicideprevention/basics/faq
Sexual health	https://doh.wa.gov/you-and-your-family/illness-and-disease-z/sexually-transmitted-infections-sti/frequently-asked-questions https://www.vdh.virginia.gov/adolescent-health-hub/sexual-health-faqs/
Health equity	https://lgbtq.unc.edu/resources/frequently-asked-questions/ https://www.unfe.org/know-the-facts/faqs/ https://lpbcc.wordpress.com/wp-content/uploads/2012/02/ref-27-health-inequalities-q-a-final.pdf https://www.heart.org/en/health-topics/consumer-healthcare/why-is-health-insurance-important/faqs-about-health-insurance

Table 7: Public resources of Public Health for Curating Large Scale Healthcare Questions.

Topics	Links
Exercise, Diet, Cardiovascular, Weight, Sleep, Cholesterol	https://www.heart.org/en/healthy-living
Cardiovascular	https://www.heart.org/en/about-us/heart-attack-and-stroke-symptoms
Cardiovascular, Cholesterol, Diabetes, Mental health, Weight	https://www.heart.org/en/health-topics
Exercise, Diet, Sleep, Cardiovascular, Weight, Cholesterol, Smoking	https://www.heart.org/en/healthy-living/healthy-lifestyle/lifes-essential-8
Cardiovascular, Cholesterol, Diabetes	https://professional.heart.org/en/guidelines-statements
Others	https://www.uspreventiveservicestaskforce.org/uspstf/recommendation-topics/information-for-consumers
Exercise, Diet, Mental health, Smoking, Sleep	https://www.sbm.org/healthy-living
Cardiovascular, Sleep, Weight	https://www.nhlbi.nih.gov/health
Cardiovascular, Sleep, Weight	https://www.nhlbi.nih.gov/es/salud
Others	https://www.nhlbi.nih.gov/resources?f%5B0%5D=language%3AEnglish&f%5B1%5D=language%3ASpanish
Diabetes, Weight, Diet	https://www.niddk.nih.gov/health-information
Diabetes, Weight, Diet	https://www.niddk.nih.gov/health-information/informacion-de-la-salud
Diabetes, SOGI	https://www.cdc.gov/diabetes/risk-factors/diabetes-risk-lgbtq.html
Diabetes, SOGI	https://www.lgbtqiahealtheducation.org/wp-content/uploads/2019/07/TFIE-35_LGBT-diabetes-Brief_final2_pages.pdf
Cardiovascular	https://www.ahajournals.org/doi/10.1161/CIR.0000000000001209
Cardiovascular	https://www.ahajournals.org/doi/full/10.1161/CIR.0000000000001003
Mental health	https://psycnet.apa.org/fulltext/2024-79040-001.html
SOGI	https://translegislation.com/
SOGI	https://www.aclu.org/legislative-attacks-on-lgbtq-rights-2024
SOGI	https://medlineplus.gov/lgbtqiahealth.html
SOGI	https://lgbtqhealthcaredirectory.org/
SOGI	https://www.lgbtqiahealtheducation.org/resources/
SOGI	https://www.hrc.org/resources
Mental health	https://www.samhsa.gov/find-help
Mental health	https://www.apa.org/topics
Mental health	https://www.nimh.nih.gov/health
Smoking	https://smokefree.gov/
Cardiovascular, Smoking	https://www.ahajournals.org/doi/10.1161/CIR.0000000000000625
HIV care	https://clinicalinfo.hiv.gov/en/guidelines
HIV care	https://www.hiv.gov/hiv-basics
HIV care	https://www.hivguidelines.org/
HIV care	https://www.cdc.gov/hiv/index.html
HIV care	https://www.cdc.gov/hivpartners/php/index.html
HIV care	https://www.who.int/news-room/fact-sheets/detail/hiv-aids
Sleep	https://www.nhlbi.nih.gov/health/heart-healthy-living/sleep
Others	https://www.drugs.com/
Others	https://www.ncbi.nlm.nih.gov/books/NBK430685/
Diabetes	https://pro.aace.com/clinical-guidance/diabetes
Diabetes	https://professional.diabetes.org/standards-of-care
Others	https://goldcopd.org/2024-gold-report/
Others	https://www.cdc.gov/vaccines/hcp/imz-schedules/index.html
Others	https://www.cancer.org/cancer/types/colon-rectal-cancer/detection-diagnosis-staging/acs-recommendations.html
Others	https://www.asccp.org/clinical-practice/guidelines
Others	https://www.acog.org/clinical/clinical-guidance/practice-bulletin/articles/2017/07/breast-cancer-risk-assessment-and-screening-in-average-risk-women
Others	https://www.asam.org/quality-care/clinical-guidelines/alcohol-withdrawal-management-guideline
Others	https://www.asam.org/asam-criteria/implementation-tools/criteria-intake-assessment-form
Others	https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/
Others	https://www.spanishreadability.com/fernandez-huerta-readability-index
Others	https://www.mccc.edu/nursing/documents/NRS225TherapeuticCommunications.pdf
Others	https://www.reddit.com/r/hivaids/?share_id=asMnalJwXk4SBcVZtxc40&utm_content=1&utm_medium=ios_app&utm_name=ioscss&utm_source=share&utm_term=4
HIV care	https://forums.poz.com/index.php?PHPSESSID=qe2trb4da2kmn0dch13402q9u2&action=forum
HIV care	https://h-i-v.net/forums

A.2 Strategies of Curating Dataset

To curate a high-quality set of patient education question-answer (QA) pairs, the initial content collection was conducted manually by annotators, most of whom are PhD students (Table 8). They identified and extracted QA-relevant content from reputable public health websites, as listed in Table 6. The annotators focused on capturing both explicit QA formats (e.g., FAQ sections) and implicit pairs derived from health articles, clinical guidelines, and symptom descriptions. The extracted content was subsequently organized and labeled by domain experts in nursing and public health to ensure accuracy and domain relevance.

To generate diverse and high-quality healthcare problems, annotators systematically reviewed all primary links and their subpages across each public health website listed in Table 7. The resources

examined included articles, PDF documents, clinical guidelines, and other educational materials. After collecting the relevant content, it was segmented into coherent text chunks based on topical boundaries and structural cues. These chunks were then used as input to an LLM (GPT-4o mini, in our case), which generated multiple healthcare problems designed to be diverse and user-oriented, guided by specific prompt instructions. Each generated problem was subsequently reviewed and validated by domain experts in nursing and public health to ensure factual accuracy, clinical relevance, and clarity.

Table 8: Biographies of all annotators involved in REPQA construction.

ID	Year	Major	Assigned Task	Expert Validator?
1	3rd year PhD	Computer Science	Scrape Website Content	✗
2	3rd year PhD	Computer Science	Scrape Website Content	✗
3	3rd year PhD	Computer Science	Scrape Website Content	✗
4	4th year PhD	Electrical Engineering	Scrape Website Content	✗
5	2nd year Master	Computer Science	Scrape Website Content	✗
6	4th year PhD	Data Science	Scrape Website Content	✗
7	4th year PhD	Nursing & Public Health	Collect Dataset	✓
8	2nd year Master	Public Health	Collect Dataset	✓
9	1st year PhD	Nursing & Public Health	Collect Dataset	✓
10	3st year PhD	Nursing & Public Health	Collect Dataset	✓

A.3 EXAMPLES OF GENERATED QUESTIONS WITH HIGH AND MEDIUM DIFFICULTIES

In Section 3.3 (RQ2), we extend the original questions into two counterparts of medium and high difficulty using the prompt provided in Appendix B.2.2, and apply them for classification. Below, we present several examples of the generated questions:

Example 1

"original_question": "What are the negative effects of smoking?"
"medium_difficulty": "How does smoking affect cardiovascular health?"
"high_difficulty": "What are the carcinogenic mechanisms of tobacco?"

Example 2

"original_question": "Is vaping actually bad for you?"
"medium_difficulty": "What are the long-term health risks of vaping?"
"high_difficulty": "Discuss the pathophysiology of e-cigarette lung injury."

Example 3

"original_question": "How many calories should I eat in a day?"
"medium_difficulty": "How is daily caloric need calculated?"
"high_difficulty": "What is the Harris-Benedict equation for BMR?"

Example 4

"original_question": "What are the types of diabetes?"
"medium_difficulty": "Distinguish between Type 1 and Type 2 diabetes."
"high_difficulty": "Explain the autoimmune pathology of Type 1 diabetes."

Example 5

"original_question": "Can I smoke with high blood pressure?"
"medium_difficulty": "How does smoking affect hypertension management?"
"high_difficulty": "Explain nicotine's effect on arterial blood pressure."

A.4 PROXY MULTIPLE CHOICE QA GENERATION

To quantify the accuracy of the generated responses by LLMs, we introduce a proxy multiple-choice QA task. The multi-choice QA pairs are generated based on the curated QA pairs, following the Algorithm 1.

Algorithm 1: Proxy Multiple Choice QA Generation

Input : Public health QA pairs $\mathcal{D}$

Output : Multiple choice QA sets $\mathcal{D}^{\text{mc}} = \{d_i^{\text{mc}}\}$, where d_i^{mc} is a triplet $(a_i, q_i^{\text{mc}}, a_i^{\text{mc}})$ with a_i the reference answer, q_i^{mc} the generated multiple-choice question, and a_i^{mc} the correct option

Function `Generate(`q, a`)`**:**

```
    // Proposer step
    q^mc, a^mc ←LLM-Proposer(q, a)
    // Step 1:  Sanity check with LLM critic on reference
    s+ ← LLM-Critic(q^mc, a)
    // Step 2:  Sanity check with LLM critic on irrelevant
       context
    s− ← LLM-Critic(q^mc, a_noise)
    if s+ ∧ ¬s− then
        // Provide feedback to proposer and regenerate
        prompt ← updated prompt
        return Generate(q, a)
    end
    return q^mc, a^mc
// Main procedures to generate all multiple-choice QAs
D^mc ← ∅
prompt ← default prompt
foreach QA pair (q_i, a_i) ∈ D do
    q_i^mc, a_i^mc ←Generate(q_i, a_i)
    D^mc ← D^mc ∪ {(a_i, q_i^mc, a_i^mc)}
end
```

B EXPERIMENTAL SETUP

B.1 CONFIGURATION OF MODELS

We set the learning rate to 1×10^{-5} and apply early stopping, using a validation set split from the training data. For post-training, we adopt LoRA Hu et al. (2021) with a rank of 16. In the reward formulation (Equ.4), we set the coefficients as $\alpha = 1$, $\beta = 1$, and $\gamma = 50$. For TA-GRPO (Equ.5), we use a weighting factor of $w = 2$. Each training run is performed on an A40 GPU for 48 hours.

During training with GRPO and TA-GRPO, we use a temperature of $\tau = 0.9$ and set the maximum generation length to 256 tokens. For inference during evaluation, we adopt greedy decoding with a maximum length of 2048 tokens. All other generation hyperparameters follow the default settings in `https://huggingface.co/docs/transformers/en/main_classes/text_generation#transformers.GenerationConfig`.

B.2 Prompt for LLMs

B.2.1 Prompt for Proxy Multiple Choice QA

System Prompt for LLM Proposer

You are a helpful assistant. Given the reference, which is a QA pair, your task is to create a multiple-choice question with one correct answer and three distractors based on the reference QA. Write the question as if it stands alone, without any indication that it was derived from a specific reference. For example, phrases like "according to the reference" are not allowed in your created question. The distractors should be as plausible as possible but either 1) incorrect or 2) correct but cannot be inferred from the reference. Format your output exactly as:
Q: <question>
A. <correct_answer>
B. <distractor1>
C. <distractor2>
... (possibly other options)
Solution: <the letter of the correct answer>. Do not include any text that is not part of the question in any round of the conversation. Your question should be as difficult as possible. Distractors that are obviously wrong should not be considered.

System Prompt for LLM Critic

You are a helpful assistant. Given a multiple-choice question and a reference document, your task is to give the correct answer letter based on the document. The format of the output should exactly be: "the letter of the correct answer", i.e., output the correct letter directly. For example, if the correct answer is A, you should output "A". Output "I do not know" if you cannot find the answer based on the document given. Do not include any other text or explanation in your response. Do not include any intermediate steps or reasoning. Do not use your own knowledge other than what is given in the document to answer the question. That is to say, you should output "I do not know" when you cannot infer the solution from the document given, even if you know the solution.

B.2.2 Prompt for RQ1 and RQ2

System Prompt for Instruction Following

You are a helpful agent in answering questions related to health. Answer clearly and kindly at a {5th grade/middle school/high school/college} reading level (U.S. grade). Maintain a positive sentiment and conversational tone.

System Prompt for Reading Level Inference

You are a careful readability assessor for healthcare and science texts. Your job is to classify the readability/audience level of a given paragraph.

Definitions (pick exactly one):
1) Elementary school: very simple words and ideas; no jargon; everyday concepts only.
2) Middle school: mostly common words; some domain terms with brief explanations; moderately short sentences; basic causal reasoning.
3) High school: frequent domain terms and abstractions; complex sentences and clauses; assumes some prior science knowledge; limited definitions of jargon.
4) College: specialized terminology or references; dense concepts and implicit assumptions; long/complex sentences; minimal or no lay explanations.

Instructions:
1) Judge solely from the paragraph provided. Do not invent context.
2) Choose ONE label from: "Elementary school","Middle school","High school","College".

Output format:
1) You may include optional analysis.
2) The LAST line must be exactly: 'Final Answer: <ONE WORD from "Elementary school","Middle school","High school","College">'
3) No extra words after the label on that last line.

System Prompt for Generating High and Medium Levels Counterpart of Original Questions

You are an expert in public health communication and medical education. Your task is to transform a simple public health question, aimed at the general public, into two more challenging versions: one "medium-difficulty" version and one "high-difficulty" version.

Category Definitions:
1. Medium Difficulty: This question should be something an educated and curious patient might ask their doctor. It is more specific than the original, may use some common medical terms, but remains generally accessible.
2. High Difficulty: This question should be one that a medical student, clinician, or researcher would ask. It must be highly technical, use specific jargon or acronyms, and assume significant background knowledge.

Core Constraints:
1. Length: The two new questions you generate MUST be of a similar length to the original question. Focus on increasing the question's conceptual density, not just its word count.
2. Output Format: Your output must be in a strict JSON format, containing three keys: original_question, medium_difficulty, and high_difficulty.

Examples:

Input Question 1:
"How can I quit smoking?"

Required JSON Output:
{
"original_question": "How can I quit smoking?",
"medium_difficulty": "What are the best smoking cessation aids?",
"high_difficulty": "What are the contraindications for varenicline therapy?"
}

System Prompt for Predicting Needed Readability Level for a Question

You are an expert in public health communication and text analysis. Your task is to classify a given public health question based on the conceptual complexity and the level of readability required to provide a comprehensive answer.
There are three categories:
1. Simple: The question is for the general public, uses common language, and asks about general knowledge. An answer would require a low Flesch-Kincaid Grade Level.
2. Intermediate: The question is for a curious and educated layperson. It is more specific, may use some common medical terms, and requires a more detailed answer than a simple question.
3. Complex: The question is for a medical student, healthcare professional, or researcher. It is highly technical, uses specific medical jargon or acronyms, and assumes significant background knowledge. An answer would require a high Flesch-Kincaid Grade Level.
Read the question below and classify it into one of the three categories. Respond with only one word: 'Simple', 'Intermediate', or 'Complex'.

B.2.3 Prompt for Benchmarking Readability-Enhancement Strategies

Standard System Prompt

You are a helpful agent in answering questions related to health. Please respond to questions based on credible information and use simple terms (limit medical jargon and 5th grade reading level). Answers should be short ($\sim$63 tokens) with simple sentence structure, and words should be easy to understand (fewer syllables). Maintain a positive sentiment and conversational tone.

CoT System Prompt

You are a helpful agent in answering questions related to health. Please respond to questions based on credible information and use simple terms (limit medical jargon and 5th grade reading level). The final answers (excluding reasoning steps) should be short ($\sim$63 tokens) with simple sentence structure, and words should be easy to understand (fewer syllables). Begin by thinking through the problem step by step, clearly showing your reasoning. After completing your reasoning, provide your final conclusion, exactly starting with 'Final Answer:'.

System Prompt for avoiding trivial solution in GRPO and TA-GRPO

Given a response to a question, you are required to judge if it is a useful answer. For example, responses like "I do not have enough information" and "I can provide general information regarding the topic" (but actually no useful information at all) are not useful answers because they give no solution to the question. Some answers that just rephrase the question without giving specific extra information should be considered as not useful as well. Answer yes if it is useful or no otherwise.

C Readability Metrics

C.1 Flesch-Kincaid Grade Level

Table 9 presents the correspondence between the Flesch-Kincaid Grade Level scores and U.S. school grade levels.[3] We use the `readability` Python library[4] to compute this metric and assess the readability of generated content.

[3] https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/
[4] https://pypi.org/project/readability/

Table 9: The relationship between Flesch-Kincaid grade level and grade level.

Flesch-Kincaid Score	Reading Level	School Level	Age Range (US)
0 - 3	Basic	Kindergarten / Elementary	5 - 8
3 - 6	Basic	Elementary	8 - 11
6 - 9	Average	Middle School	11 - 14
9 - 12	Average	High School	14 - 17
12 - 15	Advanced	College	17 - 20
15 - 18	Advanced	Post-grad	20+

C.2 Token-adapted GRPO

The standard GRPO operates at the sentence level, assigning a uniform reward across all the tokens within the generated responses. However, this uniform distribution mitigates the influence of medical jargon on the readability metrics, i.e., the professional score, which depends on the presence of domain-specific terms at certain specific tokens.

In light of this, we propose a token-adapted variant of GRPO that applies a weighted distribution of rewards across tokens. Specifically, at each optimization step, the LLM generates a group of responses $\{\boldsymbol{o}_1, \cdots, \boldsymbol{o}_G\}$ where $\boldsymbol{o}_i$ denotes i-th generated response. For t-th token in i-th response, a token-level reward is computed based on whether it appears in our curated professional vocabulary:

$$r'_{i,t} = \frac{w_{i,t} T}{\sum_{t'}^{T} w_{i,t'}} r_i \quad \text{where} \quad w_{i,t} = 1 + \delta(\boldsymbol{o}_{i,t}) w \tag{4}$$

where r_i is the original sentence-level reward for the response $\boldsymbol{o}_i$, T is the maximum length of the responses, and $\delta(\boldsymbol{o}_{i,t})$ is an indicator function that returns 1 if the token $\boldsymbol{o}_{i,t}$ belongs to a professional term, and 0 otherwise. The hyperparameter w controls the penalty strength for professional terms. Using these token-level rewards, we redefine the advantage function in GRPO as:

$$A_{i,t} = \frac{r'_{i,t} - \text{mean}(\{r_1, \cdots, r_G\})}{\text{std}(\{r_1, \cdots, r_G\})} \tag{5}$$

In our implementation, we apply this token-level reward to the professional score while maintaining the original sentence-level reward formulation for the Flesch-Kincaid grade level metric.

C.3 Professional Score

We construct the vocabulary by collecting professional terms from the Harvard Public Health medical dictionary.[5] Specifically, annotators scraped terminology entries from the site, after which a subset of terms was removed based on a review by public health experts. The vocabulary is available at `https://anonymous.4open.science/r/RepQA-45A8/dataset/harvard_medical_terms.json`.

D More Analysis

Among these models, `Claude-3.7-Sonnet` outputs responses with the highest average grade level, while `Qwen2.5-32B` achieves the lowest grade level, and `LLaMA3.1-8B` is the closest to our target value (i.e., 6). As for the professional score, `BioMistral-7B` produces the most specialized language, likely due to its fine-tuning on medical corpora. In contrast, `Gemma3-12b` yields the lowest professional word usage, which is preferable for readability. As for accuracy, `o4-mini` attains the highest score at 83.68%.

Across the four evaluated categories (*suggestion*, *fact*, *definition*, *rationale*), the *definition* category is the easiest, with an average accuracy of 84.06%, whereas the *fact* category is the most challenging, with an average accuracy of 72.80%. Notably, the *suggestion* category consistently exhibits both the

[5] `https://www.health.harvard.edu/a-through-c`

Table 10: Benchmarking both proprietary and open-source LLMs.

Model	Flesch-Kincaid	Dale Chall	SMOG	Yngve Depth
Proprietary Models				
o4-mini	3.79	7.03	5.53	2.84
o3-mini	3.99	6.72	5.56	2.55
GPT-4o	5.00	7.56	6.97	2.63
GPT-4o-mini	4.42	7.07	6.62	2.56
GPT-4.1	4.31	6.74	5.93	2.68
GPT-4	4.46	6.89	6.28	2.57
GPT-4-turbo	5.18	6.69	5.97	2.66
Claude-3.7-Sonnet	8.63	8.43	7.60	3.60
Claude-3.5-Sonnet	6.50	7.59	6.44	2.99
Gemini-1.5-pro	4.73	6.87	6.82	2.63
Gemini-2.0-flash	4.12	6.45	6.06	2.53
Gemini-1.5-flash	4.25	7.10	6.47	2.44
Open-Source Models				
LLaMA3.1-8B	6.27	8.22	8.19	2.89
Qwen2.5-7B	4.49	6.59	5.72	2.57
Phi-4	5.30	7.21	7.49	2.76
Yi-1.5-9B	8.12	8.56	8.96	2.98
Mistal-7B	6.91	9.22	7.55	2.84
InternLM3-8B	6.72	8.31	7.60	3.05
Gemma-3-12b	3.54	7.87	6.69	2.49
DeepSeek-V2-Lite	6.89	8.29	8.30	2.82
LLaMA3.1-70B	8.17	8.78	8.63	3.21
Qwen-Qwen2.5-14B	5.76	7.35	6.55	2.73
Qwen-Qwen2.5-32B	3.50	6.34	5.03	2.41
Qwen-Qwen2.5-72B	3.68	6.73	5.05	2.49

lowest Flesch-Kincaid grade level and professional score, which aligns with intuition: suggestions typically do not require complex language compared to the explanation of medical facts or rationales.

Among proprietary models, we observe that more advanced versions within a series consistently demonstrate improved accuracy. For instance, accuracy increases with model upgrades such as `o3-mini` → `o4-mini` and `GPT-4-turbo` → `GPT-4` → `GPT-4.1`. These trends indicate that newer model iterations are more effective in generating accurate responses.

E More readability metrics

Table 10 reports additional readability scores. Alongside the Flesch–Kincaid score, we include Dale–Chall Tanprasert & Kauchak (2021), SMOG Mc Laughlin (1969), and Yngve Depth Yngve (1960). These metrics are all strongly correlated with Flesch–Kincaid (Dale–Chall: 0.800; SMOG: 0.806; Yngve Depth: 0.823), the primary readability measure used in the main text. The close agreement across metrics further supports the robustness of our conclusions.

F LLM Usage

Large language models were used only for editorial polishing (grammar, style, and readability). They did not generate content or conduct analyses. All changes were reviewed by the authors.