

1 **Pilot Study of Large Language Models as an Age-Appropriate Explanatory Tool for**

2 **Chronic Pediatric Conditions**

3

4 Cameron C. Young MPhil^{1,2}, Elizabeth Enichen BA^{1,2}, Arya Rao BA^{1,2}, Sidney Hilker MD^{1,3},

5 Alex Butler MD, MS^{1,3}, Jessica Laird-Gion MD^{1,3}, Marc D. Succi MD^{1,2,4}

6

7 **Affiliations**

8 ¹Harvard Medical School, Boston, MA,

9 ²Medically Engineered Solutions in Healthcare Incubator, Innovation in Operations Research

10 Center, Mass General Brigham, Boston, MA

11 ³Boston Children's Hospital, Boston, MA

12 ⁴Department of Radiology, Massachusetts General Hospital, Boston, MA

13

14 **Corresponding Author**

15 Marc D. Succi, MD

16 Mass General Brigham

17 55 Fruit Street Boston, MA 02114

18 Phone: 617-935-9144

19 Email: msucci@mgh.harvard.edu

20

21 **Key Words:** large language models; artificial intelligence; pediatric chronic conditions;

22 medical communication

23

24 **Conflicts of Interest:** The authors have no relevant financial or non-financial interests to

25 disclose.

26 **Funding/Role of the Funder:** The project described was supported in part by award Number
27 T32GM144273 from the National Institute of General Medical Sciences. The content is
28 solely the responsibility of the authors and does not necessarily represent the official views of
29 the National Institute of General Medical Sciences or the National Institutes of Health.

30 **Author Contributions:** Cameron Young conceptualized and designed the study, drafted the
31 initial manuscript, designed the data collection instruments, carried out analyses, and
32 critically reviewed and revised the manuscript.

33 Ellie Enichen and Arya Rao conceptualized and designed the study and critically reviewed
34 and revised the manuscript.

35 Drs Sidney Hilker, Alex Butler, and Jessica Laird-Gion collected data and critically reviewed
36 and revised the manuscript.

37 Dr Marc Succi conceptualized and designed the study, coordinated and supervised data
38 collection, and critically reviewed and revised the manuscript for important intellectual
39 content.

40 All authors approved the final manuscript as submitted and agree to be accountable for all
41 aspects of the work.

42 **Ethics Approval:** This study was compliant with all applicable Health Insurance Portability
43 and Accountability Act regulations and did not require Institutional Review Board review.

44 **Data Statement:** All data will be made available for any research purpose by contacting the
45 corresponding author.

46

47 **Abstract**

48 There exists a gap in existing patient education resources for children with chronic
49 conditions. This pilot study assesses large language models' (LLMs) capacity to deliver
50 developmentally appropriate explanations of chronic conditions to pediatric patients. Two
51 commonly used LLMs generated responses that accurately, appropriately, and effectively
52 communicate complex medical information, making them a potentially valuable tool for
53 enhancing patient understanding and engagement in clinical settings.

54 **Introduction**

55 The ability to translate complex medical terminology into commonly understood
56 phrases is one of the numerous promising applications of artificial intelligence (AI),
57 particularly large language models (LLMs), in the healthcare field.¹⁻⁸ LLMs are advanced AI
58 models designed to understand and generate human-like text by leveraging vast amounts of
59 data and complex algorithms. Communicating medical information to children with chronic
60 conditions presents a unique challenge for providers as developmental stages, perspectives,
61 and understanding vary considerably across ages and disease processes.⁹ Previous studies
62 have shown that how providers communicate can affect both health outcomes and patient and
63 caregiver satisfaction;^{10,11} particularly, ineffective communication can result in negative
64 outcomes for children and families.^{12,13} Therefore, ensuring children comprehend health
65 information empowers active participation in their medical care, increasing knowledge and
66 treatment adherence, while reducing adverse events.^{14,15}

67 There exists a gap in educational materials for pediatric patients with chronic
68 conditions due to the lack of standardized approaches, particularly for rare diseases,
69 indicating a scarcity of research in this area. Current materials often fail to cater to the
70 specific needs of pediatric patients, neither being written in age-appropriate, plain language
71 nor considering the complexities of multisystemic diseases, or focus on educating the parents,
72 rather than the patient.¹⁵ Recent studies emphasize the significance of tailoring educational
73 programs to meet the unique needs of pediatric patients with chronic conditions. For instance,
74 a component-based educational program was successful in improving self-efficacy and
75 treatment satisfaction among children with rare chronic diseases.¹⁶

76 LLMs offer a novel solution to this challenge. Given this potential, we hypothesize
77 that LLMs can serve as effective tools for providing age-appropriate explanations of chronic
78 conditions, thereby enhancing the communication between healthcare providers, caregivers,

79 and pediatric patients. This study evaluates the ability of two commonly used LLMs to
80 generate accurate, complete, and developmentally appropriate explanations of chronic
81 diseases to children of different ages. By integrating these AI tools into pediatric healthcare
82 communication, we aim to bridge the gap between clinical knowledge and patient
83 comprehension, fostering better engagement and adherence to treatment among young
84 patients.

85

86 **Methods**

87 Two generalist LLMs (GPT-4 [OpenAI] and Gemini 1.0 Ultra [Google]; accessed
88 January 16, 2024) were asked to respond to the following prompt: “act as a pediatrician and
89 explain a diagnosis of [CONDITION] to a [AGE]-year-old in language they can understand.”
90 Responses were generated for five common chronic conditions (asthma, anaphylactic allergy
91 [peanut allergy], epilepsy, sickle cell disease, and type I diabetes) for children of odd ages
92 between 5 and 17 (5-year-old, 7-year-old, 9-year-old, 11-year-old, 13-year-old, 15-year-old,
93 and 17-year-old). Representative responses from GPT-4 and Gemini can be found in

94 **Supplementary Table 1.**

95 A total of 70 LLM responses (35 from each model) were assessed for accuracy,
96 completeness, age-appropriateness, possibility of demographic bias, and overall quality,
97 based on an existing framework for the human evaluations of the clinical application of
98 LLMs and prior literature.¹⁷ Demographic bias was defined as whether implementing the
99 response in clinical practice would favor or disadvantage particular groups based on
100 demographic characteristics such as race, age, gender, socioeconomic status, or geographic
101 location. Three pediatric physicians (S.H., A.B., and J.L.) rated the responses based on how
102 well they aligned with these five criteria using a Likert scale from 1 (highly disagree) to 5
103 (highly agree). Numeric ratings were treated as continuous variables and summarized as

104 means and 95% confidence intervals. A Welch two sample t-test was used to assess
105 differences in means. $P < 0.05$ was considered statistically significant. Intra-rater reliability
106 was assessed by calculating Pearson correlation coefficients between individual raters.
107 Additionally, Pearson correlation coefficients were computed to assess the degree of
108 correlation between evaluation criteria Analyses were performed in R version 4.2.2.

109

110 **Results**

111 Across both LLMs, responses were rated as highly accurate (GPT-4: 4.37 [4.27-4.47];
112 Gemini: 4.55 [4.45-4.65]), highly complete (GPT-4: 4.25, [4.16-4.34]; Gemini: 4.39, [4.28-
113 4.50]), moderately age-appropriate (GPT-4: 3.95, [3.81-4.09]; Gemini: 3.26, [3.09-3.43]), of
114 moderate quality (GPT-4: 3.88, [3.75-4.01]; Gemini: 3.43, [3.26-3.60]), and with low
115 possibility of demographic bias (GPT-4: 1.61, [1.49-1.73]; Gemini: 1.16, [1.11-1.21]).
116 Gemini responses had a significantly lower possibility of demographic bias ($p < 0.001$), while
117 responses from GPT-4 were of significantly higher quality ($p = 0.004$) and age-appropriateness
118 ($p < 0.001$) (**Table 1**). Across both models, age-appropriateness and overall quality tended to
119 increase with age, while other criteria remained similar (**Table 2**). There were no differences
120 in ratings across chronic conditions (**Supplementary Table 2**). Intra-rater reliability was
121 high, with an average Pearson correlation coefficient of 0.72 (**Supplementary Table 3**).

122 The use of metaphors to explain biological concepts was common throughout
123 responses (red blood cells are “delivery trucks” around the body, insulin is the “key” to
124 unlocking the door for glucose to enter cells, a “glitch” in the brain causes an epileptic
125 seizure). References to superheroes (15.7% of responses), food (12.9% of responses), and
126 weather (12.9% of responses) were most frequent among all responses. Additionally, the
127 mention of videogames, sports, and cartoons were common. Some of these responses were
128 confusing in the context that they were provided (“villains blocking pipes” in a videogame

129 may not be easily understandable by all children), could be interpreted as problematic by the
130 patient (a “glitch in the brain” may seem that something is wrong that can never be fixed), or
131 risk demographic bias (referring to a child as “kiddo” or “buddy”).

132

133 **Discussion**

134 LLMs can generate accurate, complete, age-appropriate chronic disease explanations
135 with low possibility of demographic bias for children of different ages and chronic
136 conditions, providing a potential additional source of patient educational materials. These
137 models are flexible, easy-to-use, and can be implemented at the point of care by clinicians or
138 at home by parents or caregivers and personalized to a patient’s specific condition and
139 demographics. Further, technology-based interventions can positively impact pediatric
140 health-related outcomes,¹⁸ further highlighting the potential utility of these tools.

141 Additionally, the use of AI chatbots is popular among children and adolescents through their
142 integration into social media platforms, such as Snapchat’s My AI¹⁹ and as educational
143 tools.²⁰ Further, a survey of parents showed an openness towards AI-driven technologies in
144 pediatric healthcare, with quality, convenience, and cost positively influencing their
145 openness, but concerns about privacy, the need for human interaction in care, and shared
146 decision-making were noted.²¹

147 Despite these positive findings and likelihood of translatability, there are several
148 limitations related to the findings. The use of words like “kiddo” or “buddy” as well as
149 references to sports and videogames may risk biasing patients and decreasing effectiveness of
150 explanations.¹⁴ Further, differences in age-appropriateness, possibility of demographic bias,
151 and overall quality were noted between GPT-4 and Gemini. This discrepancy in LLM
152 responses could be due to variations in training data and model architecture.²² Therefore,
153 clinicians should be cognizant of these potential differences, and evaluate multiple LLM

154 output before sharing responses with patients and caregivers. Finally, these responses were
155 reviewed by pediatric clinicians, rather than children, who may interpret these responses
156 differently. Evaluation of children’s interactions with LLMs for pediatric healthcare
157 represents a promising area of future research.

158 This pilot study shows that LLMs offer a promising tool to explain complex chronic
159 diseases to children of different ages, with room for improvement. Developing custom-built,
160 specialty LLMs curated by clinicians and child development experts that incorporate patient-
161 specific details may improve these LLMs ability to act as an explanatory tool.⁹ However,
162 LLMs have the potential to aid in closing the existing gap in education materials for pediatric
163 patients with chronic conditions.

References

- 164
165
166 1. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language
167 models in medicine. *Commun Med (Lond)*. Oct 10 2023;3(1):141. doi:10.1038/s43856-023-
168 00370-1
- 169 2. Koranteng E, Rao A, Flores E, et al. Empathy and Equity: Key Considerations for
170 Large Language Model Adoption in Health Care. *JMIR Med Educ*. Dec 28 2023;9:e51199.
171 doi:10.2196/51199
- 172 3. Rao A, Kim J, Lie W, et al. Proactive Polypharmacy Management Using Large
173 Language Models: Opportunities to Enhance Geriatric Care. *J Med Syst*. Apr 18
174 2024;48(1):41. doi:10.1007/s10916-024-02058-y
- 175 4. Rao A, Pang M, Kim J, et al. Assessing the Utility of ChatGPT Throughout the Entire
176 Clinical Workflow: Development and Usability Study. *J Med Internet Res*. Aug 22
177 2023;25:e48659. doi:10.2196/48659
- 178 5. Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an Adjunct for Radiologic
179 Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. *J Am Coll Radiol*. Oct
180 2023;20(10):990-997. doi:10.1016/j.jacr.2023.05.003
- 181 6. Rao A, Pang M, Kim J, et al. Assessing the Utility of ChatGPT Throughout the Entire
182 Clinical Workflow. *medRxiv*. Feb 26 2023;doi:10.1101/2023.02.21.23285886
- 183 7. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an
184 Adjunct for Radiologic Decision-Making. *medRxiv*. Feb 7
185 2023;doi:10.1101/2023.02.02.23285399
- 186 8. Young CC, Enichen E, Rao A, Succi MD. Racial, Ethnic, and Sex Bias in Large
187 Language Model Opioid Recommendations for Pain Management. *PAIN*. 2024;
- 188 9. Shah NH, Entwistle D, Pfeffer MA. Creation and Adoption of Large Language
189 Models in Medicine. *JAMA*. Sep 5 2023;330(9):866-869. doi:10.1001/jama.2023.14217

- 190 10. Espinel AG, Shah RK, Beach MC, Boss EF. What parents say about their child's
191 surgeon: parent-reported experiences with pediatric surgical physicians. *JAMA Otolaryngol*
192 *Head Neck Surg.* May 2014;140(5):397-402. doi:10.1001/jamaoto.2014.102
- 193 11. Hsiao JL, Evan EE, Zeltzer LK. Parent and child perspectives on physician
194 communication in pediatric palliative care. *Palliat Support Care.* Dec 2007;5(4):355-65.
195 doi:10.1017/s1478951507000557
- 196 12. Dimatteo MR. The role of effective communication with children and their families in
197 fostering adherence to pediatric regimens. *Patient Educ Couns.* Dec 2004;55(3):339-44.
198 doi:10.1016/j.pec.2003.04.003
- 199 13. Hallman ML, Bellury LM. Communication in Pediatric Critical Care Units: A Review
200 of the Literature. *Crit Care Nurse.* Apr 1 2020;40(2):e1-e15. doi:10.4037/ccn2020751
- 201 14. Bell J, Condren M. Communication Strategies for Empowering and Protecting
202 Children. *J Pediatr Pharmacol Ther.* Mar-Apr 2016;21(2):176-84. doi:10.5863/1551-6776-
203 21.2.176
- 204 15. Falcao M, Allocca M, Rodrigues AS, et al. A Community-Based Participatory
205 Framework to Co-Develop Patient Education Materials (PEMs) for Rare Diseases: A Model
206 Transferable across Diseases. *Int J Environ Res Public Health.* Jan 5
207 2023;20(2)doi:10.3390/ijerph20020968
- 208 16. Niemitz M, Schrader M, Carlens J, et al. Patient education for children with
209 interstitial lung diseases and their caregivers: A pilot study. *Patient Educ Couns.* Jun
210 2019;102(6):1131-1139. doi:10.1016/j.pec.2019.01.016
- 211 17. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge.
212 *Nature.* Aug 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

- 213 18. McMullan M, Millar R, Woodside JV. A systematic review to assess the effectiveness
214 of technology-based interventions to address obesity in children. *BMC Pediatr.* May 22
215 2020;20(1):242. doi:10.1186/s12887-020-02081-1
- 216 19. Pratt N, Madhavan R, Weleff J. Digital Dialogue-How Youth Are Interacting With
217 Chatbots. *JAMA Pediatr.* Mar 18 2024;doi:10.1001/jamapediatrics.2024.0084
- 218 20. Gill SS, Xu M, Pastros P, et al. Transformative effects of ChatGPT on modern
219 education: Emerging Era of AI Chatbots. *Internet of Things and Cyber-Physical Systems.*
220 2024;4:19-23. doi:10.1016/j.iotcps.2023.06.002
- 221 21. Sisk BA, Antes AL, Burrous S, DuBois JM. Parental Attitudes toward Artificial
222 Intelligence-Driven Precision Medicine Technologies in Pediatric Healthcare. *Children*
223 *(Basel).* Sep 20 2020;7(9)doi:10.3390/children7090145
- 224 22. Lee GG, Latif E, Shi L, Zhai X. Gemini Pro Defeated by GPT-4V: Evidence from
225 Education. *arXiv.* 2023;2401.08660doi:10.48550/arXiv.2401.08660
- 226
- 227

Table 1 – Overall and age-stratified average reviewer ratings of GPT-4 and Gemini across five evaluation criteria

Large Language Model	Accuracy, mean (95% CI)	Completeness, mean (95% CI)	Age-Appropriateness, mean (95% CI)	Possibility of Demographic Bias, mean (95% CI)	Overall Quality, mean (95% CI)
<i>GPT-4</i>	4.37 (4.27, 4.47)	4.25 (4.16, 4.34)	3.95 (3.81, 4.09)	1.61 (1.49, 1.73)	3.88 (3.75, 4.01)
<i>Gemini</i>	4.55 (4.45, 4.65)	4.39 (4.28, 4.50)	3.26 (3.09, 3.43)	1.16 (1.11, 1.21)	3.43 (3.26, 3.60)
<i>P-value</i>	0.08	0.15	<0.001	<0.001	0.004

CI = confidence interval

Table 2 – Age-stratified average reviewer ratings of GPT-4 and Gemini responses across five evaluation criteria

Large Language Model	Accuracy, mean (95% CI)	Completeness, mean (95% CI)	Age-Appropriateness, mean (95% CI)	Possibility of Demographic Bias, mean (95% CI)	Overall Quality, mean (95% CI)
<i>GPT-4</i>					
<i>5-Year-Old</i>	4.20 (3.76, 4.64)	4.07 (3.67, 4.47)	3.47 (2.76, 4.18)	1.53 (1.07, 1.99)	3.47 (2.87, 4.07)
<i>7-Year-Old</i>	4.40 (4.08, 4.72)	4.20 (3.99, 4.41)	4.07 (3.62, 4.52)	1.53 (1.15, 1.91)	3.93 (3.63, 4.23)
<i>9-Year-Old</i>	4.47 (4.21, 4.73)	4.27 (3.97, 4.57)	4.07 (3.71, 4.43)	1.60 (1.28, 1.92)	3.93 (3.57, 4.29)
<i>11-Year-Old</i>	4.40 (3.94, 4.86)	4.27 (3.97, 4.57)	4.00 (3.57, 4.43)	1.33 (1.08, 1.58)	3.80 (3.32, 4.28)
<i>13-Year-Old</i>	4.27 (3.91, 4.63)	4.13 (3.75, 4.51)	3.87 (3.33, 4.41)	1.73 (1.24, 2.22)	3.93 (3.41, 4.45)
<i>15-Year-Old</i>	4.40 (3.98, 4.82)	4.40 (4.08, 4.72)	3.67 (2.91, 4.43)	1.93 (1.34, 2.52)	3.93 (3.31, 4.55)
<i>17-Year-Old</i>	4.47 (4.09, 4.85)	4.40 (4.08, 4.72)	4.53 (4.27, 4.79)	1.60 (1.07, 2.13)	4.13 (3.81, 4.45)
<i>Gemini</i>					
<i>5-Year-Old</i>	4.47 (4.01, 4.93)	4.27 (3.82, 4.72)	2.53 (1.79, 3.27)	1.33 (1.02, 1.64)	2.87 (2.18, 3.56)
<i>7-Year-Old</i>	4.53 (4.11, 4.95)	4.40 (3.98, 4.82)	2.53 (1.90, 3.16)	1.07 (0.94, 1.20)	3.07 (2.32, 3.82)
<i>9-Year-Old</i>	4.60 (4.14, 5.06)	4.47 (4.09, 4.85)	3.00 (2.37, 3.63)	1.20 (0.99, 1.41)	3.20 (2.51, 3.89)
<i>11-Year-Old</i>	4.60 (4.28, 4.92)	4.40 (4.03, 4.77)	3.00 (2.49, 3.51)	1.07 (0.94, 1.20)	3.07 (2.48, 3.66)
<i>13-Year-Old</i>	4.67 (4.42, 4.92)	4.27 (3.91, 4.63)	3.80 (3.32, 4.28)	1.13 (0.95, 1.31)	4.00 (3.57, 4.43)
<i>15-Year-Old</i>	4.60 (4.23, 4.97)	4.47 (4.01, 4.93)	3.80 (3.52, 4.08)	1.20 (0.99, 1.41)	3.87 (3.41, 4.33)
<i>17-Year-Old</i>	4.47 (4.01, 4.93)	4.27 (3.82, 4.72)	2.53 (1.79, 3.27)	1.33 (1.02, 1.64)	2.87 (2.18, 3.56)

CI = confidence interval

