Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Nobody writes for the reader: Reading level in paediatric health guidance, real health queries and AI chatbot responses

Conal Cunningham

conal.cunningham2@childrenshealthireland.ie

Children's Health Ireland at Crumlin

Colm Breatnach[1]

Children's Health Ireland at Crumlin

Research Article

Keywords: Health literacy, Readability, Large language models, Artificial intelligence, Patient education, Child health

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11576943/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Nobody writes for the reader: reading level in paediatric health guidance, real health queries and AI chatbot responses

Conal Cunningham[1], Colm Breatnach[1]

[1] Paediatric Intensive Care Unit, Children's Health Ireland at Crumlin, Dublin, Ireland

Corresponding author: Conal Cunningham, conal.cunningham2@childrenshealthireland.ie

ORCID: Conal Cunningham: 0009-0009-8037-2560, Colm Breatnach: 0000-0002-1453-8050

Word count: 2990 (main text, excluding title page, abstract, references and captions)

Acknowledgements We thank the writers of the original queries.

Abstract

Purpose To compare the reading level of paediatric health guidance, parents' own questions and artificial intelligence (AI) chatbot responses to those questions, and to test whether chatbots adapt to the reading level of the person asking.

Methods Cross-sectional study, reported following the Chatbot Assessment Reporting Tool (CHART), of 18,392 guideline pages (17,516 family-facing, 876 clinician-facing) from 86 organisations, 1,000 parent-authored questions sampled from 23,477 public forum posts, and 5,000 responses to those questions from five commercially deployed chatbots. Flesch–Kincaid Grade Level (FKGL) was the primary outcome, with six secondary metrics. In a sub-study, 100 questions were resubmitted with an instruction to match the reader's level or to answer at sixth-grade level (1,500 responses).

Results Median FKGL was 12.51 for clinician-facing guidance, 9.08 for family-facing guidance, 8.37 for chatbot responses and 6.15 for parents' questions. The American Medical Association sixth-grade standard was exceeded by 99.3% of clinician-facing pages, 92.6% of family-facing pages and 89.2% of chatbot responses. Every chatbot answered above the reading level of the question asked (median difference +1.21 to +3.57 grades). Given an explicit sixth-grade instruction, the median gap narrowed in every model and three of five answered below the reader's level.

Conclusion Paediatric health information from institutions and chatbots alike is written above the level of its readers. Chatbots do not adapt to the reader by default, but can when instructed.

What is Known:

• Paediatric health information, from institutions and AI chatbots alike, is often written above recommended reading levels.

• Chatbot readability studies have used investigator-written prompts rather than real parents' questions.

What is New:

• All five chatbots tested answered 1,000 real parent questions above the reading level of the question asked.

• All five simplified when explicitly instructed: the capability exists but is unused by default.

Keywords Health literacy; Readability; Large language models; Artificial intelligence; Patient education; Child health

Abbreviations

AI	artificial intelligence
AMA	American Medical Association
API	application programming interface
CHART	Chatbot Assessment Reporting Tool
FKGL	Flesch–Kincaid Grade Level
IQR	interquartile range
LLM	large language model
NIH	National Institutes of Health

50 SMOG Simple Measure of Gobbledygook

51

1 Introduction

Generative artificial intelligence (AI) has become a mainstream source of health information: ChatGPT alone handled more than 2.6 billion messages per day by mid-2025 [1], with about 40 million people using it for health purposes daily [2], and in 2026 roughly a third of adults had used a large language model (LLM) for a health query [3]. People rate AI-generated medical advice as comparable to physician advice even when its accuracy is low [4], and current models still answer genuine patient questions unsafely at non-trivial rates [5].

Whether such an answer can be *understood* is a separate question from whether it is correct. Readability, quantified through formulae such as the Flesch–Kincaid Grade Level (FKGL) [6, 7], is not sufficient for comprehension but is a prerequisite: a reader who cannot decode a sentence cannot evaluate its claim. The American Medical Association (AMA) recommends that patient material be written at sixth-grade level and the National Institutes of Health (NIH) at eighth-grade level. The stakes are high in paediatrics, where the reader is usually a parent making a time-pressured decision for a child: parental health literacy influences child health outcomes, medication errors and emergency department use [8, 9], yet 47.6% of adults across eight countries had inadequate or problematic health literacy in the European Health Literacy Survey [10].

Recommended reading levels for patient materials are long established [11], yet audits over three decades have found materials pitched well above them [12, 13]. AI-generated material is no different: ChatGPT's English-language vaccination guidance measured grade 12.84 against the Centers for Disease Control and Prevention's 12.13 [14], and in a randomised evaluation lay users assisted by an LLM did not reliably outperform controls at identifying conditions or choosing a course of action [15]. A 2026 review of 140 studies across 21 chatbots reached the

same conclusion for AI-generated health text, but found a field of single-specialty audits of investigator-written prompts too heterogeneous for meta-analysis [16].

What has not been established is whether conversational LLMs can, and do, adapt to the individual reader. The question itself gives a conversational system a sample of the reader's own written language. Whether that signal is used is testable but untested at scale, because existing studies supply investigator-written prompts and so exclude the reader's language by construction. We aimed to establish whether LLMs adapt their reading level to that of the question asked and, if not, whether this reflects default behaviour or an inability to match the reader.

2 Methods

2.1 Design

We performed a cross-sectional comparison of three text corpora, reported following the Chatbot Assessment Reporting Tool (CHART) statement [17] (checklist in Online Resource 2). FKGL was the pre-specified primary outcome, with six further metrics as secondary outcomes. The corpora were built for a parallel study validating an automated LLM jury against expert clinician ratings.

2.2 Guideline corpus

We web-scraped paediatric guidance from organisations in Ireland, the UK, Australia, New Zealand, Canada and the US that were accredited paediatric hospitals or health systems, professional medical societies, government health agencies or nationally recognised medical charities; guidance had to be publicly accessible without login. Of 26,076 candidate pages, 24,486 remained after removing non-content pages, pages lacking substantive clinical content, extraction artefacts, non-English pages and exact duplicates. Pages were classified as family-

facing or clinician-facing using source-level defaults, URL-path rules and title-keyword matching, with manual review of ambiguous cases. Automated structural denoising was followed by a two-phase LLM pass (gpt-oss-20b) reproducing article text verbatim (Online Resource 1, Methods S1). The working corpus comprised 18,392 pages (17,516 family-facing; 876 clinician-facing) from 86 organisations (Fig. 1); source-level composition decisions are in Online Resource 1, Methods S2.

2.3 Parent query corpus

All 2005–2024 posts from seven sub-forums identified in advance as likely sources of paediatric clinical questions were retrieved from the PushShift Reddit archive [18], yielding 3,127,084 candidate posts. Regular-expression filtering, BioBERT clinical-content classification [19] and an LLM audit confirming a clinical question with a child as the primary subject reduced these to 23,477 queries. Each was assigned to one of 29 subspecialties by sentence embeddings and BERTopic clustering [20], and a stratified random sample of 1,000 (minimum five per subspecialty) was drawn. Public forum posts are an established proxy for the questions patients put to AI [21], unlike the examination questions and synthetic vignettes used in much previous work [22, 23].

2.4 AI response corpus

Each query was submitted verbatim to five commercially deployed LLMs, four proprietary and one open-weight (DeepSeek V4), through their consumer web interfaces rather than developer application programming interfaces (APIs), to capture responses as users receive them. Each was submitted in a fresh session with no reading-level instruction or conversational context, between 9 and 16 June 2026 from Dublin, Ireland, yielding 5,000 responses. Version identifiers were checked manually at the start and close of collection and were unchanged for all five systems.

No response was near-empty (<10 words) or malformed. Within-model variability was assessed by resubmitting a stratified 50-query subsample to each model twice (CHART item 9a).

2.5 Adaptation sub-study

A stratified subsample of 100 queries, balanced across subspecialty and question reading level, was submitted to all five models under three conditions: question alone (arm A); question plus "Please answer at the same reading level as my question" (arm B); and question plus "Please answer at a 6th-grade reading level" (arm C), giving 1,500 responses from the same model versions. The outcome was the post-instruction gap: response reading level minus that of the question. Because a gap closed by deleting content is not an improvement, we recorded the proportion of responses losing more than 40% of their arm-A word count.

2.6 Readability metrics and text normalisation

Seven validated metrics were computed for every text: FKGL (primary) [7], Flesch Reading Ease [6], Simple Measure of Gobbledygook (SMOG) [24], Gunning Fog [25], Coleman–Liau [26], Automated Readability Index [27] and Dale–Chall [28, 29], using the Python textstat library [30]. Six derive difficulty mainly from sentence length; Dale–Chall alone scores vocabulary familiarity against a fixed word list.

Classical formulas depend on words per sentence, and sentence detection depends on terminal punctuation. Unpunctuated list and bullet text, common in clinical guidance and LLM output alike, is merged by automated splitters into artificially long "sentences", inflating grade level independent of linguistic complexity. Every non-empty line was therefore given terminal punctuation if it lacked one, after stripping list markers, identically across all three corpora (Online Resource 1, Methods S3).

2.7 Statistical analysis

Distributions were non-normal (Shapiro–Wilk $p < 0.001$), so non-parametric methods were used: Kruskal–Wallis [31] with Dunn's post-hoc tests [32] and Holm correction [33], Mann–Whitney U for two-group comparisons, and Wilcoxon signed-rank tests [34] matched by query. Effect sizes are epsilon-squared for omnibus tests and matched rank-biserial correlation for pairwise comparisons [35, 36], with alpha 0.05. Given the large samples, a pre-specified practical-significance threshold flagged significant differences below one grade level (or one Dale–Chall point). The proportion of each group at or below sixth grade (AMA) and eighth grade (NIH) [11, 12] was a co-primary descriptive outcome. Because queries are much shorter than responses, the query–response comparison was repeated within query-length strata. Guideline pages are nested within organisations, so the audience comparison was refitted as a linear mixed model with organisation as a random intercept, with cluster bootstrap and rank-transformed sensitivity analyses (Online Resource 1, Methods S4).

2.8 Ethics

The Children's Health Ireland Research Ethics Committee determined that this study did not require ethics review, as it involved no human participants and no recruitment. No author or child was identified, no verbatim post text is reproduced, no account names or URLs are published, and all reporting is aggregate.

3 Results

3.1 No source of paediatric health information meets the recommended standard

The analytic sample comprised 18,392 guideline pages (17,516 family-facing; 876 clinician-facing) from 86 organisations, 1,000 parent-authored questions and 5,000 AI responses. Every group failed the recommended reading level, most by a wide margin (Table 1; Online Resource

1, Fig. S1). The AMA sixth-grade standard was exceeded by 92.6% of family-facing guidance, 89.2% of AI responses and 99.3% of clinician-facing guidance, and the NIH eighth-grade standard by 68.1%, 56.9% and 94.7%. Only for the parents' questions did a majority (82.4%) fall at or below eighth grade, suggesting the standards approximate the public's reading level.

3.2 Reading level by group and by model

Median FKGL was highest for clinician-facing guidance, 12.51 (interquartile range, IQR, 10.74–14.25), followed by family-facing guidance, 9.08 (7.55–10.97), and AI responses, 8.37 (7.02–9.83); parents' questions were markedly lower at 6.15 (4.87–7.45). The three-group comparison was significant (Kruskal–Wallis $H = 1423.6$, $df = 2$, $p < 0.001$, $\varepsilon^2 = 0.061$), and all pairwise comparisons survived Holm correction (Fig. 2). Clinician-facing guidance was harder than family-facing guidance (median difference 3.43 grades, $r = 0.598$), and AI responses easier than clinician-facing guidance (4.14 grades, $r = 0.754$). AI responses versus family-facing guidance was significant but below the practical-significance threshold (0.71 grades, $r = 0.189$).

That aggregate similarity conceals substantial between-model variation. Medians ranged from 7.13 (DeepSeek) to 9.61 (Grok), and 67.2% of DeepSeek's responses but 18.2% of Grok's reached eighth grade. ChatGPT produced the narrowest distribution of any group (IQR width 1.98 grades): 46.5% of its responses were at or below eighth grade but only 5.3% at or below sixth.

3.3 The adaptation gap: questions versus answers

Parents' questions were written at a markedly lower reading level (median FKGL 6.15) than every model's responses to them (7.13 to 9.61). The query-matched Wilcoxon signed-rank comparison was significant for every model (all $p < 0.001$), with effect sizes from medium (DeepSeek, $r = 0.440$) to large (Grok, $r = 0.907$) and median response-minus-question

differences from +1.21 (DeepSeek) to +3.57 grades (Grok). The pattern held across metrics: 34 of 35 model-by-metric comparisons were significant and above the practical-significance threshold, the exception being Claude's Dale–Chall (+0.94 against a 1-point threshold) (Online Resource 1, Fig. S2).

Restricted to the longest tertile (n = 329, >220 words) or to queries of at least 200 words (n = 384), the gap persisted for every model (ChatGPT +1.8, Claude +0.8 to +0.9, DeepSeek +0.5 to +0.6, Gemini +2.9 to +3.0, Grok +3.1 to +3.2 grades; all Holm-adjusted $p < 0.01$), although effect sizes were smaller than in the shortest tertile.

3.4 The models don't adapt, but they can

Instruction moved every model in the same direction, and an explicit target closed the gap more than an implicit one (Table 2, Fig. 3). With the explicit sixth-grade instruction, the median gap fell in every model: Claude (+2.82 → −0.30), DeepSeek (+1.33 → −1.57) and Grok (+2.53 → −2.30) crossed below the reader's own level, whereas ChatGPT (+4.34 → +2.18) and Gemini (+4.40 → +0.84) did not. For ChatGPT, Claude, DeepSeek and Gemini, content loss was modest, with 2–12% of responses losing more than 40% of their arm-A length (Table 2), indicating simplification rather than deletion. Grok achieved the largest gap closure but also the largest content loss, with 30% of responses losing more than 40% of their length.

3.5 Where the variation comes from: institution and vocabulary

FKGL varied across the 74 institutions contributing at least five pages (Kruskal–Wallis $H = 6876$, $p < 0.001$, $\varepsilon^2 = 0.372$), six times the variance explained by the family-versus-clinician distinction ($\varepsilon^2 = 0.061$). The intraclass correlation for organisation was 0.39, an effective sample of roughly 27 to 32 independent clusters. All seven metrics agreed in direction and were independently significant (ε^2 0.050 to 0.067; all $p < 0.001$; Online Resource 1, Table S1).

Decomposing FKGL into its sentence-length and word-difficulty terms attributed 87% of the mean clinician-versus-family gap to word difficulty and 13% to sentence length.

3.6 Sensitivity to text preprocessing

With sentence-boundary normalisation disabled (Table 3), guideline text and parents' questions barely changed (family-facing +0.14 grades, clinician-facing +0.39, questions +0.40), as expected for punctuated prose. AI responses rose by 1.36 grades overall, driven by an extreme outlier: ChatGPT shifted by +6.13 grades (8.12 → 14.26), against +0.42 to +0.84 for the other four models.

4 Discussion

4.1 Principal findings

Across, to our knowledge, the largest audience-classified paediatric guideline corpus yet assembled for readability analysis, 1,000 real parent questions and 5,000 chatbot responses to them, the dominant finding is a shared failure rather than a difference between groups. Nine in ten AI responses and nine in ten family-facing guideline pages exceed the reading level the AMA recommends.

4.2 The adaptation gap is a choice

The most consequential finding is that no model adjusted its output to the reader. A guideline page cannot know its reader; it is written once, for a heterogeneous audience. A conversational system receives, in the question itself, a direct sample of the reader's written language, and prompting alone substantially reduces the grade level of patient material [37, 38]. Yet every model returned an answer harder than the question that prompted it, although every model could narrow the gap when instructed.

A recent synthesis recommends that systems let a user or clinician specify a target readability band, such as FKGL 6–8 [16]. In every model, naming an explicit sixth-grade target closed the gap more than asking the model to match the reader, so a specified band is the right mechanism; but a user-specified control helps only users who know to ask. Our data support a system-level default target, set against a stated standard rather than inferred from the user's language and applied without being requested. It remains imprecise: two of five models stayed above the reader's level even with the target given directly.

4.3 Reading level is a property of individual systems, not AI generally

The 2.48-grade spread between DeepSeek and Grok on identical questions shows that readability is a property not of large language models in general but of a specific deployed configuration; findings from single-model studies should not be generalised. An audit of five consumer chatbots likewise found Grok the least readable and most likely to produce problematic content [39], but ranked DeepSeek among the harder models where we found it the easiest, plausibly because it tested earlier versions (Gemini 2.0 and DeepSeek V3). FKGL also varied significantly within model families when the analysis was repeated on a subset two months later, after all vendors had released updated versions. This argues for continuous benchmarking rather than periodic audit, and for regulators specifying a threshold rather than endorsing a vendor.

4.4 Implications

For guideline producers. Institutional house style outweighs intended audience as a determinant of reading level (ε^2 0.372 versus 0.061), so an organisational readability standard applied at publication is likely to achieve more than page-by-page revision. That vocabulary carries 87% of the audience gap is encouraging, because terminology substitution is local, verifiable and

unlikely to sacrifice content, a real concern given that prompted simplification has reached target reading levels partly by deleting up to 83% of source content [37].

For model developers. Calibrating reading level to the user's own language is technically feasible, matters for equity and is currently absent; it should be evaluated as a capability, not assumed. ChatGPT shows why sentence length alone is insufficient: it writes the shortest sentences of any model tested, yet only 5.3% of its responses reach sixth grade.

For clinicians and policymakers. Chatbot answers are pitched roughly where a hospital's family-facing webpage is pitched, above where many parents read, so directing a family to a chatbot is not, on readability grounds, materially different from directing them to a website. Readability standards for patient information are decades old [11, 12] but are not applied to AI-generated health text now delivered tens of millions of times a day [2].

4.5 Preprocessing is an unreported confound in this literature

Our normalised grade levels sit three to six points below pooled published values for the same model families; unnormalised, they converge (Table 3), so much of the apparent difference is a measurement choice. Chatbot output is heavily bullet-formatted, and unpunctuated list items defeat the sentence splitters classical formulas depend on; this one decision moved a model's median grade level by 6.13 points. Unnormalised scoring penalises models that format content for easy scanning.

Two contemporaneous audits converge with our conclusions: Tiller et al. reported Flesch Reading Ease of 30–50 across five chatbots, with 49.6% of responses problematic on quality grounds [39], and Carlson et al. found FKGL 10.1–14.2 across five systems answering urological

questions [40]. Neither found a system meeting the standard, and Carlson et al. found their most accurate chatbot the least readable [40].

5 Limitations

Readability formulas on short texts. Queries are short relative to the texts these formulas were validated on: 94.6% fall below SMOG's recommended 30 sentences. Figures here are corpus- and stratum-level aggregates, not statements about individual queries, and the adaptation gap persists in direction and significance for every model at every query length.

Corpus representativeness. Reddit posts from 2005–2024 are a proxy for questions put to a chatbot; most predate consumer LLMs and may come from a population whose literacy differs from that of today's chatbot users. Each query was submitted once, over seven days. Findings are English-language and Anglosphere only.

What readability does not capture. Readability formulas measure surface features as proxies for comprehension difficulty, not comprehension, organisation, accuracy or actionability; they can be gamed by shortening sentences without improving clarity [41, 42], and different formulas disagree on the same text by several grade levels [43]. Our claim is therefore bounded: paediatric health text, institutional and AI-generated alike, sits above recommended reading levels and does not adapt to the reader. This does not establish that parents fail to understand it; instruments assessing understandability and actionability [44], and comprehension testing with parents, are the necessary complement.

6 Conclusions

Paediatric health information is written above the level of the people who read it, whether produced by institutions or by the AI chatbots now answering tens of millions of people daily. Every one of the five leading chatbots tested answered above the level of the question asked, yet each could adapt when asked directly. The capability is present but unused, which makes reading level a product decision, and a missed opportunity to give parents advice they can understand.

References

1. Chatterji A, Cunningham T, Deming DJ, Hitzig Z, Ong C, Shan CY et al (2025) How people use ChatGPT. NBER Working Paper No. 34255. National Bureau of Economic Research, Cambridge, MA

2. OpenAI (2026) AI as a healthcare ally: how Americans are navigating the system with ChatGPT. OpenAI, San Francisco, CA

3. Mendel T, Singh N, Mann DM, Wiesenfeld B, Nov O (2025) Laypeople's use of and attitudes toward large language models and search engines for health queries: survey study. J Med Internet Res 27:e64290. https://doi.org/10.2196/64290

4. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P (2025) People overtrust AI-generated medical advice despite low accuracy. NEJM AI 2:AIoa2300015. https://doi.org/10.1056/AIoa2300015

5. Draelos RL, Afreen S, Blasko B et al (2026) Large language models provide unsafe answers to patient-posed medical questions. npj Digit Med 9:241. https://doi.org/10.1038/s41746-026-02428-5

6. Flesch R (1948) A new readability yardstick. J Appl Psychol 32:221–233. https://doi.org/10.1037/h0057532

7. Kincaid JP, Fishburne RP, Rogers RL, Chissom BS (1975) Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Research Branch Report 8-75. Naval Technical Training Command, Millington, TN

8. DeWalt DA, Hink A (2009) Health literacy and child health outcomes: a systematic review of the literature. Pediatrics 124(Suppl 3):S265–S274. https://doi.org/10.1542/peds.2009-1162B

9. Sanders LM, Federico S, Klass P, Abrams MA, Dreyer B (2009) Literacy and child health: a systematic review. Arch Pediatr Adolesc Med 163:131–140. https://doi.org/10.1001/archpediatrics.2008.539

10. Sørensen K, Pelikan JM, Röthlin F, Ganahl K, Slonska Z, Doyle G et al (2015) Health literacy in Europe: comparative results of the European Health Literacy Survey (HLS-EU). Eur J Public Health 25:1053–1058. https://doi.org/10.1093/eurpub/ckv043

11. Weiss BD (2003) Health literacy: a manual for clinicians. American Medical Association Foundation and American Medical Association, Chicago, IL

12. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ et al (2021) Readability of patient education materials from high-impact medical journals: a 20-year analysis. J Patient Exp 8:2374373521998847. https://doi.org/10.1177/2374373521998847

13. Okuhara T, Furukawa E, Okada H, Yokota R, Kiuchi T (2025) Readability of written information for patients across 30 years: a systematic review of systematic reviews. Patient Educ Couns 135:108656. https://doi.org/10.1016/j.pec.2025.108656

14. Joshi S, Ha E, Amaya A, Mendoza M, Rivera Y, Singh VK (2024) Ensuring accuracy and equity in vaccination information from ChatGPT and CDC: mixed-methods cross-language evaluation. JMIR Form Res 8:e60939. https://doi.org/10.2196/60939

15. Bean AM, Payne RE, Parsons G, Kirk HR, Ciro J, Mosquera-Gómez R et al (2026) Reliability of LLMs as medical assistants for the general public: a randomized preregistered

344 study. Nat Med 32:609–615. https://doi.org/10.1038/s41591-025-04074-y (Publisher Correction:
345 https://doi.org/10.1038/s41591-026-04404-8)

346 16. Olszewski R, Brzeziński J, Watros K, Rysz J (2026) Quantifying readability in chatbot-
347 generated medical texts using classical linguistic indices: a review. Appl Sci 16:1423.
348 https://doi.org/10.3390/app16031423

349 17. The CHART Collaborative (2025) Reporting guideline for chatbot health advice studies: the
350 Chatbot Assessment Reporting Tool (CHART) statement. Ann Fam Med 23:389–398.
351 https://doi.org/10.1370/afm.250386

352 18. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J (2020) The Pushshift Reddit
353 dataset. Proc Int AAAI Conf Web Soc Media 14:830–839.
354 https://doi.org/10.1609/icwsm.v14i1.7347

355 19. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH et al (2020) BioBERT: a pre-trained
356 biomedical language representation model for biomedical text mining. Bioinformatics 36:1234–
357 1240. https://doi.org/10.1093/bioinformatics/btz682

358 20. Grootendorst M (2022) BERTopic: neural topic modeling with a class-based TF-IDF
359 procedure. arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794

360 21. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB et al (2023) Comparing
361 physician and artificial intelligence chatbot responses to patient questions posted to a public
362 social media forum. JAMA Intern Med 183:589–596.
363 https://doi.org/10.1001/jamainternmed.2023.1838

22. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C et al (2023) Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2:e0000198. https://doi.org/10.1371/journal.pdig.0000198

23. Nori H, King N, McKinney SM, Carignan D, Horvitz E (2023) Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375. https://doi.org/10.48550/arXiv.2303.13375

24. McLaughlin GH (1969) SMOG grading: a new readability formula. J Read 12:639–646

25. Gunning R (1952) The technique of clear writing. McGraw-Hill, New York

26. Coleman M, Liau TL (1975) A computer readability formula designed for machine scoring. J Appl Psychol 60:283–284. https://doi.org/10.1037/h0076540

27. Smith EA, Senter RJ (1967) Automated readability index. AMRL-TR-66-22. Aerospace Medical Research Laboratories, Wright-Patterson AFB, OH

28. Dale E, Chall JS (1948) A formula for predicting readability. Educ Res Bull 27:11–20, 28

29. Chall JS, Dale E (1995) Readability revisited: the new Dale–Chall readability formula. Brookline Books, Cambridge, MA

30. Bansal S, Aggarwal C (n.d.) textstat: Python package to calculate readability statistics of a text object, version 0.7.13. https://pypi.org/project/textstat/

31. Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. J Am Stat Assoc 47:583–621. https://doi.org/10.1080/01621459.1952.10483441

32. Dunn OJ (1964) Multiple comparisons using rank sums. Technometrics 6:241–252. https://doi.org/10.1080/00401706.1964.10490181

384 33. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70

385 34. Wilcoxon F (1945) Individual comparisons by ranking methods. Biom Bull 1:80–83.
386 https://doi.org/10.2307/3001968

387 35. Tomczak M, Tomczak E (2014) The need to report effect size estimates revisited: an
388 overview of some recommended measures of effect size. Trends Sport Sci 21:19–25

389 36. Kerby DS (2014) The simple difference formula: an approach to teaching nonparametric
390 correlation. Compr Psychol 3:11.IT.3.1. https://doi.org/10.2466/11.IT.3.1

391 37. Moons P, Van Bulck L (2024) Using ChatGPT and Google Bard to improve the readability
392 of written patient information: a proof of concept. Eur J Cardiovasc Nurs 23:122–126.
393 https://doi.org/10.1093/eurjcn/zvad087

394 38. Swisher AR, Wu AW, Liu GC, Lee MK, Carle TR, Tang DM (2024) Enhancing health
395 literacy: evaluating the readability of patient handouts revised by ChatGPT's large language
396 model. Otolaryngol Head Neck Surg 171:1751–1757. https://doi.org/10.1002/ohn.927

397 39. Tiller NB, Marcon AR, Zenone M, Kidd KE, Jeukendrup AE, Master Z et al (2026)
398 Generative artificial intelligence-driven chatbots and medical misinformation: an accuracy,
399 referencing and readability audit. BMJ Open 16:e112695. https://doi.org/10.1136/bmjopen-2025-
400 112695

401 40. Carlson JA, Cheng RZ, Lange A, Nagalakshmi N, Rabets J, Shah T et al (2024) Accuracy
402 and readability of artificial intelligence chatbot responses to vasectomy-related questions: public
403 beware. Cureus 16:e67996. https://doi.org/10.7759/cureus.67996

41. Agency for Healthcare Research and Quality (n.d.) Tip 6: use caution with readability formulas for quality reports. TalkingQuality. AHRQ, Rockville, MD. https://www.ahrq.gov/talkingquality/resources/writing/tip6.html

42. Zheng J, Yu H (2017) Readability formulas and user perceptions of electronic health records difficulty: a corpus study. J Med Internet Res 19:e59. https://doi.org/10.2196/jmir.6962

43. Wang LW, Miller MJ, Schmitt MR, Wen FK (2013) Assessing readability formula differences with written health information materials: application, results, and recommendations. Res Social Adm Pharm 9:503–516. https://doi.org/10.1016/j.sapharm.2012.05.009

44. Shoemaker SJ, Wolf MS, Brach C (2014) Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 96:395–403. https://doi.org/10.1016/j.pec.2014.05.027

Statements and Declarations

Ethics approval. The Children's Health Ireland Research Ethics Committee determined that this study did not require research ethics review, as it involves no human participants and no active recruitment. See §2.8 for data-privacy safeguards.

Data availability. The derived readability dataset (25,892 scored texts), all 6,500 model responses, the guideline corpus URL index and the statistical output tables are deposited at Zenodo (DOI: 10.5281/zenodo.**22846367**) under CC BY 4.0. Guideline page text is not redistributed, as it is third-party copyright; every source URL is released so the corpus can be re-retrieved and re-scored. Verbatim parent-question text is not released: the questions were written by pseudonymous authors who did not consent to research use and who retain the ability to delete their posts, which permanent archiving would remove. All derived metrics for the query corpus are released in full, and query identifiers are internal sequential indices that cannot be resolved to source records.

Code availability. All analysis code is deposited in the same repository under the MIT licence. Readability scores were computed with textstat 0.7.13; other versions of that package alter the syllable counter and will not reproduce these values exactly.

Funding. No funding was received for this work.

Competing interests. None of the organisations whose systems were evaluated had any role in the design of the study, the collection or analysis of data, the preparation of the manuscript, or the decision to publish. No author has any financial, advisory or employment relationship with any of them.

Use of AI in this work. Large language models were used as (i) an object of study, (ii) a component of the corpus-cleaning pipeline (gpt-oss-20b, §2.2). No AI system was used to generate, interpret or select results. All statistical analyses were performed with conventional software and verified by the authors.

Author contributions. CC: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing: original draft, Writing: review & editing. CB: Supervision, Writing – review & editing

Consent to participate. Not applicable: no participants were recruited; the study analysed publicly available text.

Consent to publish. Not applicable: the article contains no individual person's data, and no verbatim post text is reproduced.

Table 1 Flesch–Kincaid Grade Level and attainment of recommended standards, by group.

Group	n	Median FKGL	IQR	% ≤ 6th grade (AMA)	% ≤ 8th grade (NIH)
Whole guideline corpus	18,392	9.19	7.61–11.22	7.1	30.6
Family-facing guidance	17,516	9.08	7.55–10.97	7.4	31.9
Clinician-facing guidance	876	12.51	10.74–14.25	0.7	5.3
AI responses (all models)	5,000	8.37	7.02–9.83	10.8	43.1
- ChatGPT 5.5	1,000	8.12	7.23–9.21	5.3	46.5
- Claude Sonnet 4.6	1,000	7.36	6.29–8.59	18.7	64.1
- DeepSeek V4	1,000	7.13	5.87–8.51	27.4	67.2
- Gemini 3.5 Flash	1,000	9.59	8.39–10.85	1.0	19.4
- Grok 4.3 Fast	1,000	9.61	8.37–11.11	1.4	18.2
Patient/parent questions	1,000	6.15	4.87–7.45	46.7	82.4

Table 2 Adaptation sub-study: post-instruction gap and content loss, by model and arm.

Model	Arm A (control)	Arm B (implicit)	Arm C (explicit)	>40% shorter than arm A, arm C
ChatGPT 5.5	+4.34	+3.65	+2.18	2%
Claude Sonnet 4.6	+2.82	+1.24	−0.30	2%
DeepSeek V4	+1.33	−1.02	−1.57	12%
Gemini 3.5 Flash	+4.40	+2.73	+0.84	6%
Grok 4.3 Fast	+2.53	+0.17	−2.30	30%

Table 3 Median Flesch–Kincaid Grade Level with and without sentence-boundary normalisation, against pooled published estimates for the same model families.

Group	**Normalised**	**Unnormalised**	**Shift**	**Pooled published** [16]
Whole guideline corpus	9.19	9.36	+0.16	-
Family-facing guidance	9.08	9.22	+0.14	-
Clinician-facing guidance	12.51	12.90	+0.39	-
AI responses (all models)	8.37	9.73	+1.36	-
- ChatGPT 5.5	8.12	14.26	+6.13	13.85
- Claude Sonnet 4.6	7.36	8.15	+0.79	11.22
- DeepSeek V4	7.13	7.79	+0.67	8.45
- Gemini 3.5 Flash	9.59	10.43	+0.84	13.14
- Grok 4.3 Fast	9.61	10.03	+0.42	-
Patient/parent questions	6.15	6.56	+0.40	-

Figure legends

Fig. 1 Flow diagram of the guideline, parent question and AI response corpora

Fig. 2 Flesch–Kincaid Grade Level by group and model. Dotted lines mark the American Medical Association sixth-grade and National Institutes of Health eighth-grade standards

Fig. 3 Post-instruction gap (response minus question Flesch–Kincaid Grade Level) by model and arm in the adaptation sub-study, n = 100 queries per model per arm. Arm A, question alone; arm B, instruction to match the reader's level; arm C, instruction to answer at sixth-grade level. Box plots suppress outliers beyond the whiskers (1.5 × IQR). ">40% shorter than arm A" is the proportion of that model and arm's responses losing more than 40% of their matched arm-A word count (pre-specified threshold)

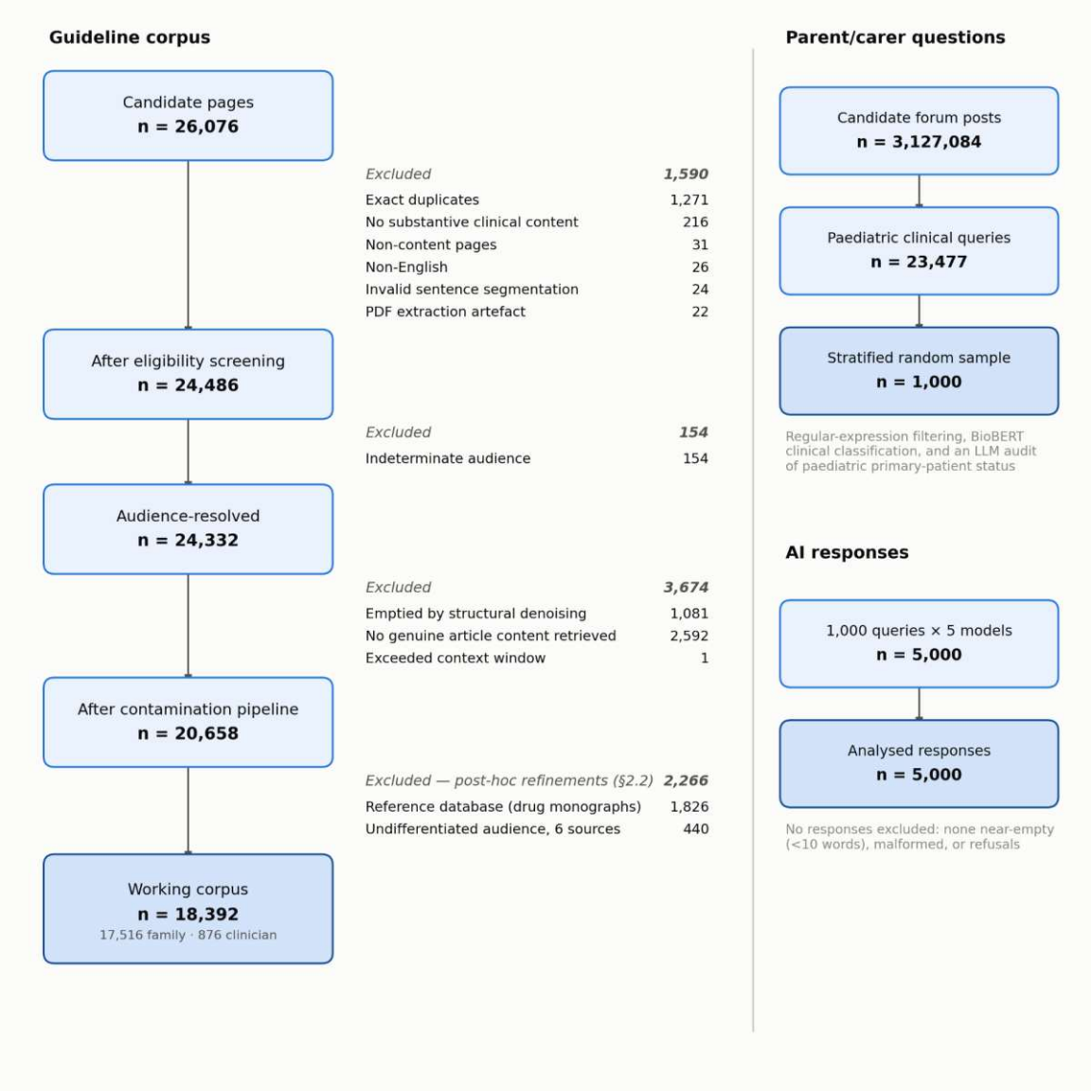
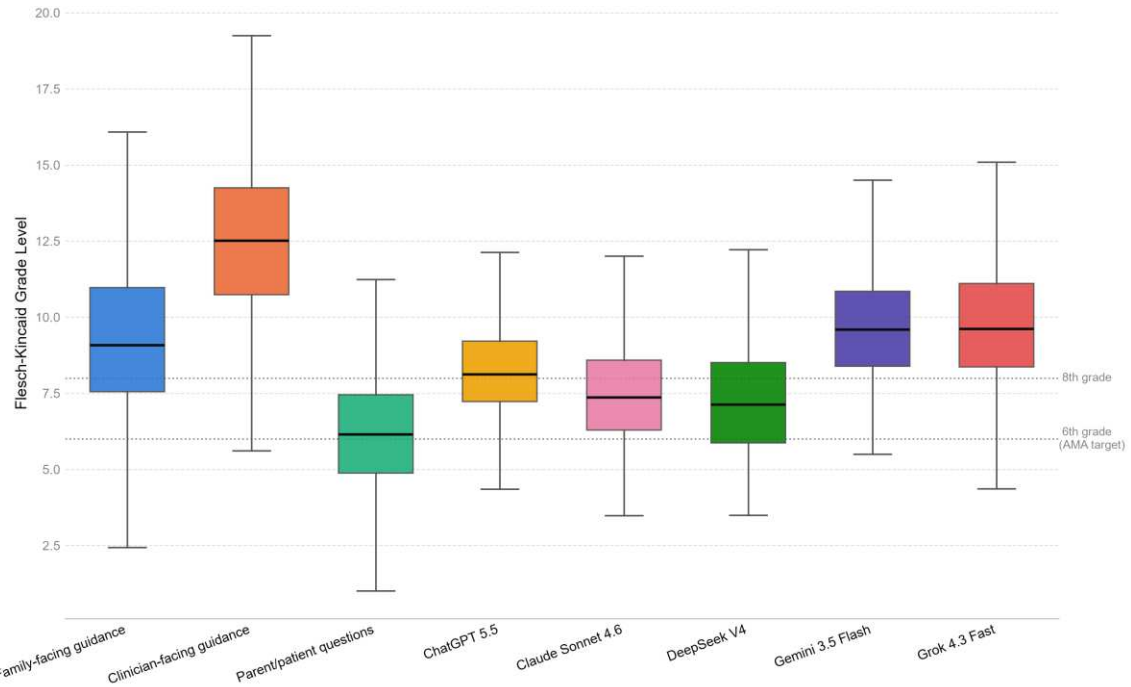


Fig. 1 Flow diagram of the guideline, parent question and AI response corpora



471

472 **Fig. 2** Flesch–Kincaid Grade Level by group and model. Dotted lines mark the American
473 Medical Association sixth-grade and National Institutes of Health eighth-grade standards

474

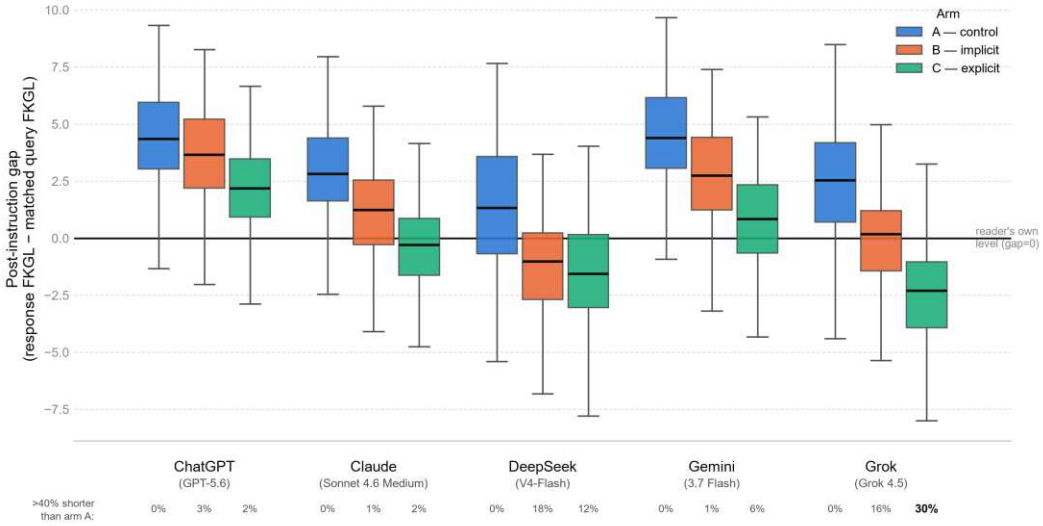


Fig. 3 Post-instruction gap (response minus question Flesch–Kincaid Grade Level) by model and arm in the adaptation sub-study, n = 100 queries per model per arm. Arm A, question alone; arm B, instruction to match the reader's level; arm C, instruction to answer at sixth-grade level. Box plots suppress outliers beyond the whiskers (1.5 × IQR). ">40% shorter than arm A" is the proportion of that model and arm's responses losing more than 40% of their matched arm-A word count (pre-specified threshold)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource1.pdf
- OnlineResource2CHARTchecklist.pdf