Online Resource 1

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

European Journal of Pediatrics

Conal Cunningham, Colm Breatnach

Corresponding author: Conal Cunningham, Paediatric Intensive Care Unit, Children's Health Ireland at Crumlin, Dublin, Ireland, conal.cunningham2@childrenshealthireland.ie

Contents

- Methods S1: Contamination pipeline
- Methods S2: Source-level composition refinements
- Methods S3: Sentence-boundary normalisation
- Methods S4: Clustering specification
- Table S1: All seven readability metrics, by group
- Fig. S1: Attainment of recommended reading-level standards, by group
- Fig. S2: Effect sizes across all seven readability metrics

Section numbers (§) refer to the main manuscript. Names in monospace refer to scripts and data files in the accompanying data and code repository.

Methods S1: Contamination pipeline

Referenced from §2.2. Pilot review of the initially scraped guideline corpus found widespread non-article content: navigation chrome and legal boilerplate mixed into the extracted text, and passages duplicated as unpunctuated run-on blocks (the same class of formatting the sentence- boundary normalisation rule addresses at the scoring stage, §2.6, but present here in the raw extraction itself, before any readability scoring).

Two-stage design

1. **Automated structural denoising** — rule-based removal of navigation menus, boilerplate, and detectable duplicated blocks from the raw scraped HTML/text.
2. **LLM-based cleaning pass** (gpt-oss-20b) — reproduces genuine article text verbatim, without summarisation, from whatever automated denoising left behind. This is a *reproduction* task, not a rewriting task: the model's output is checked against the input for verbatim fidelity, not graded on quality.

Quantified outcome

This lowered the primary outcome as expected: median FKGL fell from 10.03 to 9.33 across the 17,846 pages the two-stage pipeline cleaned (Wilcoxon $p < 0.001$), consistent with removing artificially-inflating boilerplate and duplication rather than any change to genuine article content.

The 893/817-page residual pipeline-failure subset

A separate, more specific failure mode within that same population: extraction calls that complete normally (`finish_reason="stop"`, not a token-limit truncation) but return anomalously little content relative to what was fed in — long, heavily duplicated source documents where the model's verbatim-reproduction task silently drops content rather than reproducing all of it. Identified via the `task_a5_truncation_bias.py` analysis (source: `task_a5_truncation_bias.csv`), using a retention ratio (post-extraction word count ÷ pre-extraction, structurally-denoised word count), on the population of pages that were flagged contaminated, successfully LLM-cleaned, and present in the final corpus.

The anomaly threshold is empirical, not asserted: the bottom 5% of this population's own ratio distribution.

This subset was originally identified on the corpus as it stood before the §2.2 composition refinements (20,658 pages) — the original computation gave 893 pages (4.32% of that corpus), median 1.32 grades easier after extraction (Wilcoxon $p = 4.9\times10^{-59}$), and an estimated corpus-level bias of −0.057 grades. **Recomputed here on the refined 18,392-page corpus** (identical rule, empirical threshold re-derived fresh on the refined population, not reusing the original cut value): **817 pages (4.44% of the refined corpus)**, median **1.55 grades** easier after extraction (Wilcoxon **$p = 5.7\times10^{-77}$**), estimated corpus-level bias **−0.069 grades**. Checked for stability across a range of identification thresholds rather than reported at a single arbitrary cut — bottom 3%/5%/8% of the ratio distribution give corpus-level bias estimates of −0.050, −0.069, and −0.085 grades respectively, all the same direction (extraction scores easier than source) and all significant at $p < 10^{-45}$. The finding is not sensitive to the exact percentile chosen.

On recoverability: the original 893-page identification was not lost — it was found (`task_a5_truncation_bias.py`, `bottom_5pct_ratio_whole_population` row of its output CSV) and reproduces the four originally published values exactly. It predates the §2.2 refinements, though, so the count, the median shift, and the corpus-level bias all needed recomputing against the refined corpus rather than simply re-reported — reported here as the honest record of that recomputation, not as if the refined-corpus numbers had been the ones originally computed.

Residual contamination after cleaning: failure mechanisms

Even after both stages, a residual-contamination audit of the pipeline's output (`task1g_residual_contamination_classified.csv`, 8,154 pages audited) classified remaining contamination into two named mechanisms, distinguishing pages where both apply from pages where neither could be positively classified:

Mechanism	Pages	Mean prefix-contamination fraction
Cross-page near-duplicate (content repeated verbatim across multiple pages of the same source, not caught by the single-page denoising pass)	3,241	0.189
Internal duplication (content repeated within a single page)	177	0.141
Both mechanisms present	118	0.077
Unclassified (some contamination signal present, mechanism not confidently attributable to either category above)	4,618	0.183

"Prefix-contamination fraction" is the proportion of the page's leading text still attributable to a contamination signal after the two-stage clean, not a document-wide measure — it characterises how much of what remains is concentrated near the start of the extracted text (consistent with navigation/boilerplate remnants, which typically precede genuine article content rather than being interleaved with it).

Validation

The pipeline's detection/cleaning accuracy was validated against a manually labelled ground-truth set of 30 contaminated pages, scored against two finalist configurations head-to-head on the identical pairs (`task1p_results.csv`, confirmed as the citable round): DeepSeek-v4-flash in JSON-response mode (90.0% agreement, 27/30) and gpt-oss-20b in plain-text line-based mode (93.3% agreement, 28/30). gpt-oss-20b/plaintext was the better-performing and is the configuration actually deployed in the published pipeline (§2.2) — this is the source of the "validated against 60 sampled pairs" figure (30 pages x 2 configurations = 60 scored predictions). Two earlier candidate files (`task1n_groq_comparison_results.csv`, `task1o_fireworks_comparison_results.csv`) also tested gpt-oss-20b but in a JSON-response-mode configuration that scored 3-7% agreement; those were earlier, abandoned configurations, not the validation of the deployed pipeline, and are not cited here.

Methods S2: Source-level composition refinements

Referenced from §2.2. Both post-hoc refinements to the guideline corpus (20,658 → 18,392 pages, 93 → 86 organisations) were specified on structural grounds, before their effect on any comparison was known: (1) removal of sources whose scraped content is genuinely undifferentiated between family and clinician audiences, from the audience-stratified refined corpus entirely; (2) exclusion of one source that is a structurally distinct document class (a reference database, not narrative guidance).

These are two distinct treatments, not one. The six "reassigned" sources below are **not** duplicated into both the family and clinician groups — that would double-count 440 pages inside an audience-stratified analysis whose entire point is a clean split. Instead they are held out of the refined corpus as their own separate, unstratified group, for the same practical reason LactMed is excluded: this study's audience comparisons require a page to have one defensible label, and these sources don't have one. "Reassigned" describes what changed about their status (no longer forced into a single label), not where they ended up (in the refined corpus under two labels) — despite what an earlier draft of this section implied.

Every decision below was made at the level of the specific content actually scraped for this corpus, not the organisation's website as a whole. Several organisations host a distinct clinician-facing section elsewhere on their site (continuing professional development material, clinical practice guidelines) while the specific pages retrieved for this corpus are shared, undifferentiated content — those cases are reassigned on that basis, not because the organisation lacks any clinician-facing material at all.

Removed from the refined corpus: six sources (family ↔ clinician undifferentiated)

Source	Family pages	Clinician pages	Basis
BfN_Drugs	70	5	Single "Drugs in Breastmilk" factsheet series; no separate professional portal exists at all, confirmed on both the About page and the specific factsheet page.
Child_Welfare_US	2	0	Audience-specific landing

			pages route to the same underlying topic content; no separate topic hierarchy for caseworkers vs. families.
ENT_UK	77	0	Org-level segmentation exists (a genuine Health Professionals section, with separate clinical guidelines/journals/CPD), but the e-leaflet content type actually scraped for this corpus is the *same* document set the professional section links back to, not a distinct version. Reassigned on content-level (page-as-unit-of-analysis) grounds, not because the organisation has no clinician content anywhere.
GIKids	177	0	No provider-specific content areas anywhere in navigation.
Oxford_Vaccine	108	0	Unified navigation (Vaccines / Infectious Diseases / Vaccines Explained); no professional-specific path.
Young_Womens_Health	1	0	Same content-level-vs-org-level pattern as ENT_UK: `/clinicians` exists as a nav item, but the Health Guides content actually scraped is explicitly described as unified, shared content for all audiences. Minimal impact (1 page).

Total removed on this basis: 440 pages (435 family-side + 5 clinician-side from BfN_Drugs) taken out of the audience-stratified refined corpus entirely — no longer forced into a single label, but not counted under either label going forward, and not duplicated into both.

Excluded: one source (structurally distinct document class)

Source	Pages excluded	Basis
LactMed	1,826 (clinician-facing)	A fixed drug-monograph schema (NCBI Bookshelf / NLM technical bulletin format), indexed by substance rather than narrative guidance indexed by clinical presentation — the same structural distinction the guideline-inclusion criteria are built around for every other source in the corpus. Checked explicitly for dual-audience wording before deciding on the reference-database exclusion rather than reassignment: the only dual-audience language found was in an App Store listing, which is not primary NIH

		documentation and was not used as grounds for reassignment.

LactMed was also, independently, a substantial leverage point in the clustering analysis (1,826 pages from one organisation against a corpus median cluster size of 38 pages — see Methods S4), which is consistent with, but not the basis for, its exclusion.

What this changes and doesn't

These are composition decisions about which pages carry which audience label and which pages belong in the guideline corpus at all — they are not adjustments to any computed reading-level score. Every retained page's FKGL (and the other six metrics) is scored identically before and after refinement; refinement only changes which group a page is counted in, or whether it is counted at all. The original-versus-refined comparison of the summary statistics is provided in the data repository (`table_S12_original_vs_refined.csv`).

Source data: `task_a1_9_audience_decisions.csv` (93 rows, one per source organisation, with the full evidence note for every decision — not only the six reassigned and one excluded, which are the two decision types shown above).

Methods S3: Sentence-boundary normalisation

Referenced from §2.6. Full rationale, the line-unit rule, heading handling, and corpus-scale verification for the sentence-boundary normalisation step applied identically to all three corpora before scoring.

Rationale

Classical readability formulas (including FKGL) depend on words-per-sentence, and every general-purpose sentence splitter detects sentence boundaries from terminal punctuation. List- and bullet-formatted text without trailing punctuation on each item — pervasive in both clinical guidance and LLM output — is merged by such splitters into one artificially long "sentence," inflating the apparent grade level for reasons that have nothing to do with linguistic complexity. A five-item bulleted list with no punctuation reads, to a sentence splitter, as a single 40-word sentence; the same five items each ending in a period read as five ordinary sentences.

The line-unit rule

`readability_common.normalize_for_scoring()` applies one rule, identically, to every one of the three corpora (guideline pages, patient/parent questions, AI responses):

3. Split the raw text on literal newline characters (`\n`) into lines. Boundary detection is newline-based, not list-marker-based, and happens **before** any marker stripping.
4. Within each line, strip a leading list marker if present (numbers, bullets, dashes) — stripping never affects boundary detection, since boundaries were already fixed in step 1.
5. If the (marker-stripped) line does not already end in terminal punctuation, append a period.

Because line-splitting happens strictly before marker-stripping, the failure mode "stripping the marker first destroys the evidence needed to detect the boundary" cannot occur in this implementation — markers are never used to detect boundaries in the first place.

Heading handling

The rule has no special case for "is this line a heading" — every non-empty newline-delimited line is treated identically, whether it is a genuine list item or a bare heading with no list marker at all (e.g. a line reading only `Warning Signs to Watch For`). This is deliberate: the underlying problem (a heading with no terminal punctuation, immediately followed by list-formatted body text with no terminal punctuation) is the same defeat-the-splitter pattern either way, and confirmed in practice to produce the correct effect — a heading line becomes its own short, correctly-terminated "sentence" rather than being fused onto the content that follows it.

Corpus-scale verification

A corpus-wide audit (6,500 responses: 5,000 main-corpus + 1,500 A12) after normalisation found:

- Proportion of texts with words-per-sentence over 40: **0.0% in every one of the 20 model×corpus groups**, including the model (Claude) whose low normalisation shift might otherwise be mistaken for undetected list usage.
- Proportion with zero detected newlines, or collapsed to a single sentence: effectively 0.0% (3 isolated responses out of 6,500 total; not a systematic pattern).
- Median words-per-sentence across groups: 10.65-17.77 — within the range expected for ordinary prose, with no group showing the multi-hundred-word "mega-sentence" pattern normalisation is designed to prevent.

The single code path (`readability_common.compute_metrics()`) is imported and called identically by the main-corpus scoring script and the A12 scoring script — there are not two implementations that could have diverged.

A note on how this was actually verified

An earlier audit of this exact normalisation step was triggered by a displayed passage that appeared, on manual review, to be one long unpunctuated block. Tracing it to source (the raw response store, not a rendered view of it) confirmed the raw text does contain the expected line breaks, and that `normalize_for_scoring()` correctly split it into 22 properly terminated sentences (mean sentence length 11.73 words) — not one inflated sentence. The apparent problem was a **display artefact**: the audit-sample markdown export embeds response text via an f-string with no hard line break, and standard Markdown rendering collapses a single `\n` between two plain-text lines into a space unless a paragraph break or explicit hard break is present. The underlying `.md` file's bytes contained the real newlines throughout; only a rendered view of that file lost them. This was confirmed decisively before concluding the scoring pipeline itself required no correction — see the disposition in the normalisation-audit verdict document for the full trace (`task_norm_audit_verdict.md`).

Net effect on the pipeline: none. No re-scoring was required. One minor, separable correction was identified during the same audit: an older, narrower list-detector undercounted DeepSeek's list usage in some contexts (main corpus 41.3% -> 44.7% under a detector that also catches bullets and bare headings). This affects only a descriptive list-prevalence percentage, not any readability score.

Methods S4: Clustering specification

Referenced from §2.7. Guideline pages are nested within source organisations (a single organisation may contribute anywhere from 1 to 3,550 pages), so the family-vs-clinician audience comparison was refitted with organisation explicitly modelled as a clustering variable, using three independent estimators, plus a permutation test as an assumption-free check.

1. Linear mixed model

`fkgl ~ audience + (1 | source_org)`, fit on the refined corpus (18,392 pages, 86 organisations). Reports the audience fixed effect on the raw FKGL scale with a Satterthwaite/Wald confidence interval, alongside the intraclass correlation (ICC) for source organisation and the design effect (and resulting effective sample size) this implies.

2. Cluster bootstrap

Resamples whole organisations with replacement (not individual pages), so the resampling unit matches the actual clustering structure rather than treating pages as independent. The audience effect is recomputed on each resampled dataset; the reported interval is the percentile (2.5-97.5%) of the resulting distribution. Because this resamples clusters rather than observations, its precision is governed by the number of organisations (86), not the number of pages (18,392) -- consistent with it producing a visibly wider interval than the mixed model's parametric CI.

3. Rank-transformed refit

The same mixed-model specification refit on rank-transformed FKGL, as a sensitivity check against the mixed model's normality assumption on residuals -- residuals from the raw-scale model are not normally distributed (Shapiro-Wilk $W=0.957$, $p<0.001$ in every corpus scenario tested; full diagnostics in the data repository), so this refit checks whether the fixed-effect conclusion survives relaxing that assumption. Reported on the rank scale, not comparable in magnitude to the other two estimators, but comparable in direction and significance.

4. Permutation test

An assumption-free check independent of all three model-based estimators above: audience labels are permuted within organisation 10,000 times (so the permutation respects clustering rather than treating pages as exchangeable across organisations), and the observed family-vs-clinician median difference is compared against the resulting null distribution.

5. Organisation-weighted medians

Alongside the page-weighted medians reported as the primary descriptive statistic throughout, organisation-weighted medians (compute each organisation's own median FKGL, then take the median of those 86 per-organisation medians) are provided in the data repository (`table_S10_per_organisation.csv`), so a reader can see the audience comparison is not solely an artefact of a few very large organisations dominating the page-weighted count.

Results across all three corpus scenarios

Run on three scenarios to isolate the effect of each post-hoc refinement -- the original corpus, LactMed excluded only, and the full refined corpus (both refinements, the one used throughout the main text). Headline figures (full per-scenario detail in `table_S7_clustering_diagnostics.csv`, data repository):

Scenario	n pages	n orgs	ICC	Effective n	Mixed-model effect (95% CI)	Bootstrap median (95% CI)	Rank-transformed	Permutation p
Original	20,658	93	0.379	31.9	2.26 (2.05-2.46)	2.84 (1.69-4.89)	4155.9 (3759.4-4552.4)	<0.0001
LactMed excluded	18,832	92	0.388	28.5	2.26 (2.06-2.45)	3.54 (1.96-5.36)	3729.4 (3363.7-	<0.0001

only							4095.0)	
Refined (both refinements)	**18,392**	**86**	**0.393**	**26.9**	**2.27 (2.07-2.47)**	**3.53 (1.85-5.37)**	**3652.1 (3293.0-4011.2)**	**<0.0001**

All three model-based estimators agree in direction and significance in every scenario. The refined-corpus row corresponds to the main-text analysis (§2.7, §3.5: ICC 0.39, effective n roughly 27); the other two rows are reported here to show the sensitivity of these figures to each refinement individually, not as alternative headline results.

Source: `task_a1_mixed_model.json`, `task_a1_permutation.json` and `table_S7_clustering_diagnostics.csv` (data repository).

Table S1: All seven readability metrics, by group

Median [interquartile range], refined corpus (18,392 guideline pages). Flesch Reading Ease is scored inversely to the other six metrics (higher = easier); Dale-Chall is a vocabulary-familiarity score, the other six are grade-level formulas driven mainly by sentence length.

Panel A. Distribution by group

Group	n	FKGL	Flesch Reading Ease	SMOG	Gunning Fog	Coleman-Liau	ARI	Dale-Chall
Whole guideline corpus	18,392	9.19 [7.61–11.22]	57.93 [44.49–67.13]	11.66 [10.29–13.21]	11.61 [9.77–14.02]	10.56 [8.91–12.90]	10.27 [8.54–12.57]	9.87 [8.96–11.14]
Family-facing guidance	17,516	9.08 [7.55–10.97]	58.67 [46.32–67.58]	11.57 [10.21–13.05]	11.47 [9.68–13.74]	10.43 [8.84–12.56]	10.13 [8.46–12.32]	9.79 [8.91–10.94]
Clinician-facing guidance	876	12.51 [10.74–14.25]	32.11 [21.70–42.99]	13.98 [12.58–15.30]	15.73 [13.62–17.65]	14.98 [13.11–16.84]	13.65 [11.74–15.70]	12.68 [11.63–13.58]
AI responses (all models)	5,000	8.37 [7.02–9.83]	59.69 [51.57–67.26]	10.92 [9.89–12.03]	10.56 [9.10–12.22]	10.24 [8.86–11.91]	9.49 [7.85–11.28]	9.55 [8.85–10.35]
ChatGPT 5.5	1,000	8.12 [7.23–9.21]	58.77 [52.34–64.22]	10.71 [9.97–11.54]	10.23 [9.14–11.53]	10.57 [9.51–11.73]	9.24 [8.24–10.43]	9.55 [9.01–10.20]
Claude Sonnet 4.6	1,000	7.36 [6.29–8.59]	64.78 [57.39–70.48]	10.28 [9.43–11.21]	9.63 [8.41–11.05]	9.41 [8.25–10.70]	7.83 [6.68–9.26]	9.02 [8.40–9.68]
DeepSeek V4	1,000	7.13 [5.87–8.51]	66.78 [59.00–73.66]	10.01 [9.09–11.08]	9.25 [7.92–10.75]	8.96 [7.76–10.28]	7.98 [6.76–9.47]	9.40 [8.78–10.10]
Gemini 3.5 Flash	1,000	9.59 [8.39–10.85]	58.30 [50.89–64.80]	11.91 [10.86–13.02]	12.07 [10.74–13.48]	9.97 [8.88–11.21]	10.77 [9.35–12.11]	9.35 [8.78–9.99]
Grok 4.3 Fast	1,000	9.61 [8.37–11.11]	50.33 [41.02–58.09]	11.74 [10.78–12.77]	11.70 [10.32–13.28]	12.94 [11.49–14.66]	11.94 [10.46–13.56]	10.73 [9.95–11.47]
Patient/parent questions	1,000	6.15 [4.87–7.45]	78.04 [72.19–83.17]	8.84 [7.71–9.80]	8.23 [6.83–9.70]	5.60 [4.63–6.68]	5.91 [4.44–7.57]	8.09 [7.49–8.72]

Panel B. Three-group comparison by metric (family-facing guidance n = 17,516, clinician-facing guidance n = 876, AI responses n = 5,000; Kruskal–Wallis, $\varepsilon^2 = (H - k + 1)/(n - k)$)

Metric	H	df	ε^2	p
FKGL	1,423.6	2	0.061	< 0.001
Flesch Reading Ease	1,270.6	2	0.054	< 0.001
SMOG	1,345.8	2	0.057	< 0.001
Gunning Fog	1,557.3	2	0.066	< 0.001
Coleman-Liau	1,178.0	2	0.050	< 0.001
ARI	1,170.8	2	0.050	< 0.001
Dale-Chall	1,543.8	2	0.066	< 0.001

Supplementary figures

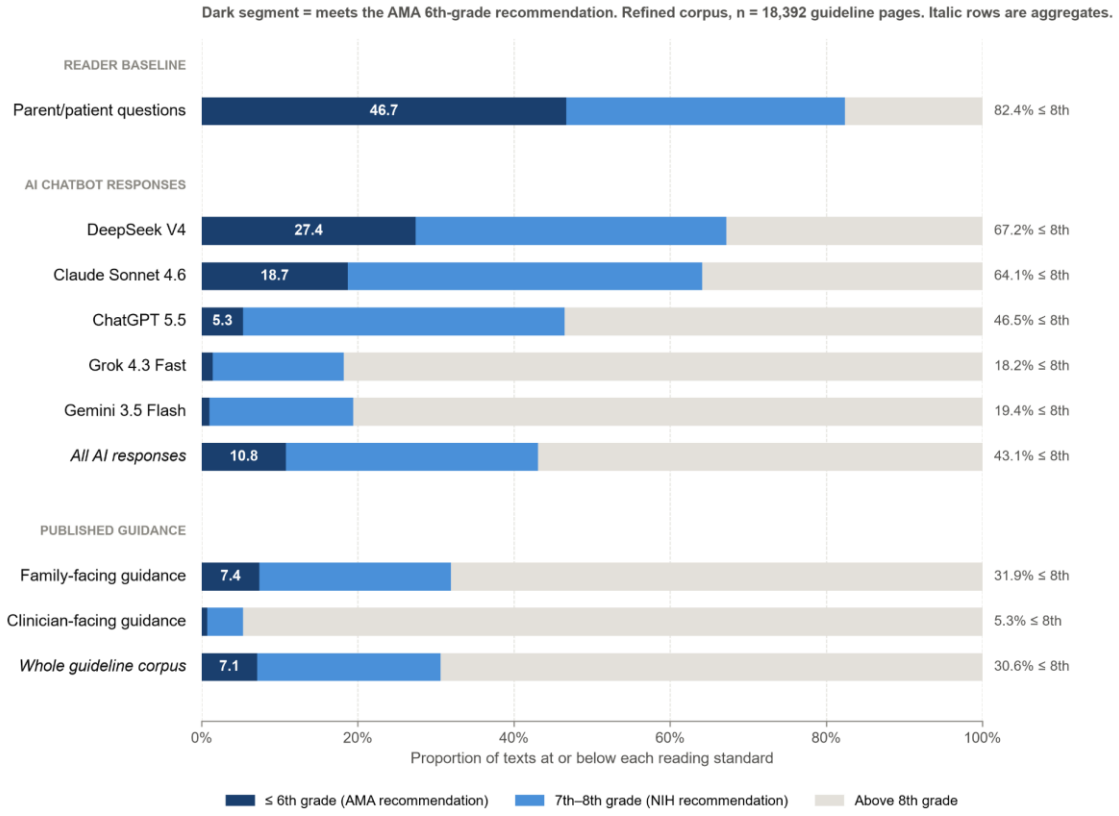


Fig. S1 Proportion of each group at or below the American Medical Association sixth-grade and National Institutes of Health eighth-grade Flesch–Kincaid Grade Level standards

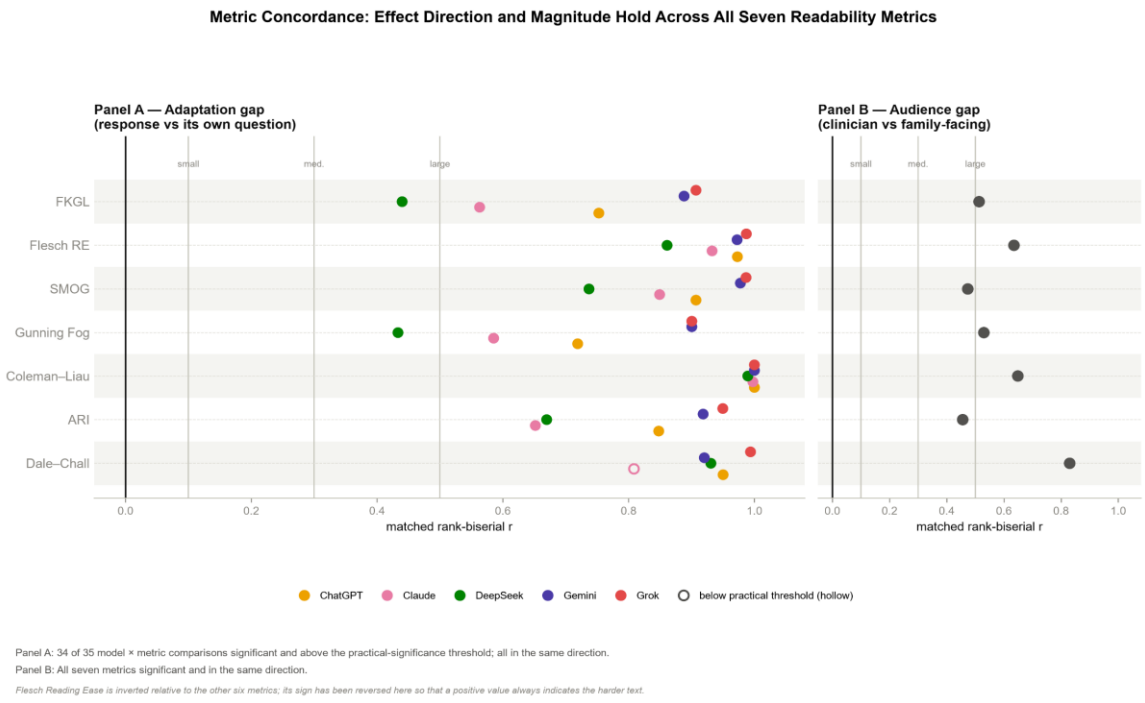


Fig. S2 Matched rank-biserial effect sizes across all seven readability metrics: a, the adaptation gap (each model's responses versus the questions that prompted them); b, the audience gap (clinician-facing versus family-facing guidance). Hollow markers fall below the practical-significance threshold