Toward Grade-Aligned Educational Text Generation with Training-Free Readability Steering

Maya Bialik [1,2] and Will Cummings[2]

[1] Boston University Wheelock College of Education and Human Development
mbialik@bu.edu
ORCID: 0009-0002-9860-0508
[2] QuestionWell
will@questionwell.ai

Abstract. Teachers and curriculum teams regularly need the same content at different reading levels, yet the AI tools they already use miss requested grade bands by wide margins. We introduce FK-steered decoding, a readability-aware logits processor that monitors the running Flesch–Kincaid grade level during generation and nudges token scores toward a target grade, requiring no retraining and no model-specific adaptation. On a 30-prompt benchmark spanning grades 3–12 with both from-scratch writing and source-faithful rewriting tasks, the tools teachers are most likely to use—GPT-5.4, Gemini 3 Pro, and MagicSchool—miss their targets by 1.8–2.6 mean absolute grades. Applying FK-steered decoding to four open-weight models (3B–24B parameters) reduces mean absolute grade error from 1.945 to 0.649, with no observed degradation in faithfulness or coherence. These results suggest that training-free decoding-time control can make grade-level targeting substantially more reliable for drafting and re-leveling educational texts.

Keywords: readability control · text simplification · constrained decoding · differentiated materials · Flesch-Kincaid

1 Introduction

In educational settings, the same content often needs to be presented at multiple reading levels, while preserving the content ([3]). This reflects a well-established principle in learning science: comprehension depends on the match between text difficulty and reader ability, and mismatches in either direction—whether too hard or too easy—limit learning ([7]). Some have suggested that Large Language Models (LLMs) could support teachers in adapting the reading level of text ([6]). Teachers are already using AI for this purpose with some frequency ([2]). However, the accuracy of LLMs at hitting requested grade levels is not well understood. This paper asks: 1) how accurately do LLMs, particularly those used by teachers, hit requested grade levels, and 2) can a readability-aware decoder reduce that error without retraining or model-specific adaptation?

To ground these questions, we construct a 30-prompt benchmark with two task types. In the *write* task, the model produces a passage from scratch given a topic, a target grade level, and a target word count. In the *rewrite* task, the model adapts an existing source passage to a lower grade level while preserving its facts and sequence of ideas.

We also include three external baselines on the same 30 prompts, chosen for ecological validity: GPT-5.4, the most widely used AI tool among educators ([1]); Gemini 3 Pro, whose LearnLM variant claims pedagogical grounding ([4]); and MagicSchool, the most adopted K–12 AI platform with over five million educators ([5]). On this benchmark, these tools miss their targets by wide margins: GPT-5.4 scores 1.828 mean absolute grade error, Gemini 3 Pro scores 1.832, and MagicSchool scores 2.552.

We therefore introduce FK-steered decoding, a readability-aware logits processor that monitors the running grade level during generation and intervenes only when the text drifts beyond a set tolerance of the target. The base model weights remain unchanged, and the same control logic transfers across model families without retraining. Applied to four open-weight models (3B–24B parameters) under three temperatures, FK-steered decoding improves grade matching in every condition tested, with the best tolerance reducing mean absolute error to 0.649.

2 Method

Educational text adaptation is rarely a from-scratch writing problem. More often, educators need to revise existing material while preserving the core concepts, vocabulary targets, and/or factual content. In that setting, there are two important failure types:

- Missing the requested reading level
- Changing the meaning of the text, such as oversimplifying complexity

That framing shapes both the method and the evaluation. We call this method FK-steered decoding. The core idea is this: at each step of text generation, the model assigns a score to every possible next word. Our method monitors the reading level of the text produced so far. If the text is drifting away from the requested grade level beyond a set "tolerance cutoff", the method adjusts those scores to favor words that would bring the reading level back on track. If the text is already close enough, no adjustment is made and the model generates normally. More formally:

$$\tilde{\ell}_t(v) = \ell_t(v) + \lambda \,\mathrm{clip}(s_t(v), -c, c)\, I(|e_t| > \tau)$$

where $\ell_t(v)$ is the base-model logit for token v, $e_t = \hat{g}(x_{1:t}) - g^\star$ is the current grade error, $s_t(v)$ is a projected reduction in error after a short rollout, and τ is a tolerance threshold. The indicator $I(|e_t| > \tau)$ activates steering only when the current prefix deviates from the target grade beyond this tolerance.

Three properties matter for educational use.

- The method requires no retraining and no supervision for each new target grade.
- The same settings work across all tested models without adjustment, meaning the method is not tied to any single AI system.
- The tolerance makes intervention adjustable: lower values intervene more often and improve grade control more aggressively, while higher values are faster and usually safer for rewrite faithfulness.

For reproducibility, we report the fixed settings used in all runs: three lookahead tokens, a candidate pool of eight tokens, guidance scale 4.0, and score clip 1.5. The only control parameter varied across runs is the tolerance, tested at 0.0, 0.5, and 1.0.

3 Study Design

The 20 write prompts and 10 rewrite prompts span target grades 3 through 12, with rewrite sources at grades 7–12, across three subject areas: science (e.g. the water cycle, photosynthesis), social studies (e.g. the American Revolution, the Roman Republic), and math (e.g. fractions and decimals). The primary metric throughout is mean absolute Flesch-Kincaid grade-level error. As a basic safeguard, we also collect automated faithfulness and hallucination judgments for rewrite tasks using an LLM judge. These scores are not validated against human ratings and are not treated as a primary metric. Their purpose is narrower: to verify that decoding-time grade control does not degrade output coherence or introduce gross factual distortion, i.e., that the logits processor is not 'gaming' the readability target at the expense of meaningful text. Grade alignment remains the main focus.

We test four instruction-tuned base models at three temperatures (0.0, 0.3, 0.7): Llama 3.1 3B Instruct[3], Llama 3.1 8B Instruct, Mistral 7B Instruct, and Mistral Small 24B Instruct. For each model-temperature condition, we run one baseline and three FK-steered variants (tolerances 0.0, 0.5, 1.0). The open-weight baselines, GPT-5.4, Gemini 3 Pro, and MagicSchool (collected through its product UI) are all evaluated with the same readability metric. This is an engineering-style evaluation rather than a classroom trial; Section 5 returns to what these metrics do and do not establish.

4 Results

Table 1 reports the best FK-steered result for each model across all temperature and tolerance combinations, alongside the three external baselines. FK-steered decoding improves grade control in every condition tested (12 of 12 model-temperature pairs). Across those 12 conditions, mean absolute grade error falls from 1.945 to 0.649, an absolute reduction of 1.296 grades.

[3] Resolves to `Llama-3.2-3B-Instruct`; Meta does not publish a 3.1 3B checkpoint.

Table 1. Best FK-steered result per model across all temperatures. Each open-weight row is the lowest-error run from 9 configurations (3 temperatures × 3 tolerances). $\Delta|e|$: change in mean absolute grade error vs. unsteered baseline (more negative is better). Len.: output length as % of target. Runtime: latency multiple of unsteered baseline (open-weight) or absolute (external).

System	Temp.	Tolerance	$\lvert e\rvert$	$\Delta\lvert e\rvert$	Len.	Runtime
L3.1-3B	0.3	0.5	0.673	-1.403	111%	2.92×
L3.1-8B	0.7	0.0	0.473	-0.841	101%	4.96×
Mistral-7B	0.0	0.0	0.978	-1.620	173%	5.62×
Mistral-S24B	0.7	0.0	**0.233**	-1.396	119%	5.97×
GPT-5.4	n/a	–	1.828	–	95%	3.8s
Gemini-3P	1.0	–	1.832	–	100%	23.0s
MagicSchool	–	–	2.552	–	96%	n/a

Comparing against the external baselines clarifies the practical gap for educators. GPT-5.4 reaches 1.828 mean absolute error on the same prompts, Gemini 3 Pro reaches 1.832, and MagicSchool reaches 2.552. All three remain far worse than the selected FK-steered runs, which average 0.649. In short, the current tools teachers are likely to reach for first still miss requested grade levels by large margins.

Length adherence does not separate conditions as clearly: external baselines average 95–100% of target length, while open-weight models range from 100% (Llama 8B) to 173% (Mistral 7B).

As a secondary check, automated faithfulness and hallucination judgments on the rewrite tasks showed no degradation under FK-steered decoding: hallucination rates remained at or near zero and mean faithfulness scores stayed between 4.0 and 4.3 on a 5-point scale across all conditions, comparable to the external baselines.

The best overall result is Mistral Small 24B at temperature 0.7 with tolerance 0.0, reaching 0.233 mean absolute grade error. Lower tolerances dominate: tolerance 0.0 wins 8 of the 12 model-temperature conditions and 0.5 wins the remaining 4.

Figure 1 shows the cumulative distribution of errors across individual prompts: for any given error threshold on the x-axis, the curve shows what fraction of prompts had errors at or below that level. The best unsteered condition places 20.0% of prompts within 0.5 grade of the target and 40.0% within 1.0 grade; the best FK-steered condition raises those rates to 86.7% and 100.0%. GPT-5.4, Gemini 3 Pro, and MagicSchool remain farther to the right at 13.3/23.3, 10.0/33.3, and 10.0/20.0%.

The main operational choice is how much latency to spend for tighter control. If a single tolerance is fixed across all runs, tolerance 0.0 gives the best mean grade control (0.671 mean absolute error) but costs 5.16× baseline runtime. Tolerance 1.0 is cheaper at 2.51× baseline runtime, but mean grade error rises

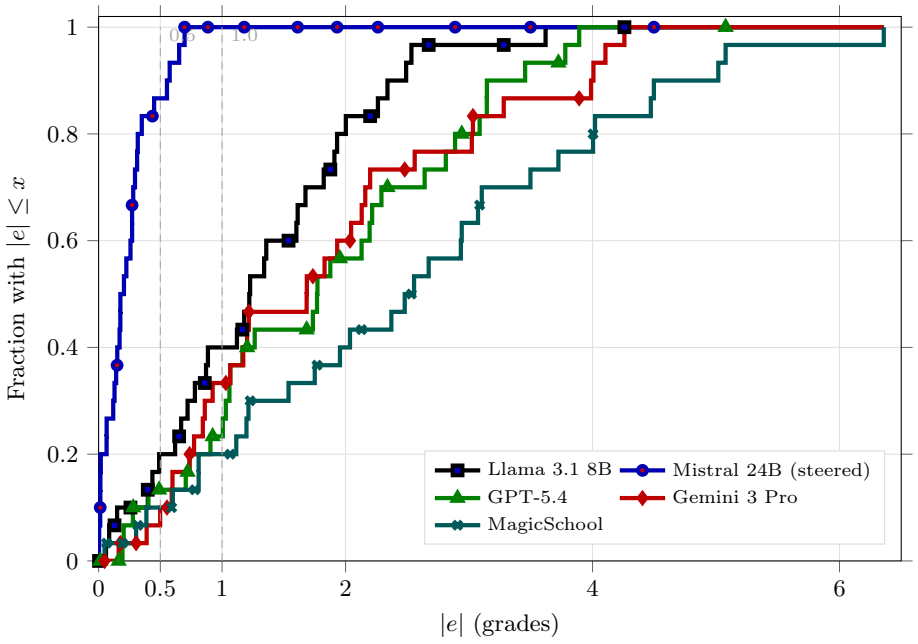

Fig. 1. ECDF of prompt-level absolute grade error. Internal curves show the best unsteered baseline (Llama 3.1 8B, T=0.7) and best FK-steered setting (Mistral Small 24B, T=0.7, tol=0.0); external baselines are GPT-5.4, Gemini 3 Pro, and MagicSchool. Higher and further-left is better.

to 0.930. For educational deployment, this is a useful property of the method: the tradeoff is explicit rather than hidden inside a prompt or a product workflow.

5 Discussion and Conclusion

FK-steered decoding narrows the gap between requested and produced reading level across every tested condition, and the rewrite results suggest that existing materials can be adapted rather than only generated from scratch.

The study has clear limitations. The benchmark is small, the main readability signal is Flesch-Kincaid, and grade alignment is only one part of instructional quality. A passage can hit a target grade and still be pedagogically weak, culturally mismatched, or poorly aligned to a lesson objective. These metrics do not tell us whether students comprehend the texts better, whether teachers prefer the outputs, or whether the adapted materials are instructionally sound without human review. Human review therefore remains necessary, especially for classroom use. The automated faithfulness and hallucination metrics serve as guardrails rather than validated quality measures; their uniformity across conditions suggests FK-steered decoding does not compromise basic output integrity, but validating against human judgment (e.g., propositional alignment or

structured rubric annotation) is necessary before recommending unsupervised classroom use.

Additionally, mean absolute error treats overshooting and undershooting symmetrically, but pedagogically the direction matters: text above the target may be inaccessible, while text below it may underestimate readers. Future work should also include a more systematic sampling design, extend coverage to grades 1–2 where readability control matters most, and evaluate upward re-leveling.

Even with those limits, the core result is useful. FK-steered decoding, with no retraining and no model-specific adaptation, improves reading-level accuracy across multiple model families and sizes, and outperforms GPT-5.4, Gemini 3 Pro, and MagicSchool on grade alignment, making training-free decoding-time control a plausible drafting aid and motivating larger studies with teacher evaluation and classroom-centered quality measures.

References

1. Carnegie Learning: The state of AI in education 2025: Key findings from a national survey. Tech. rep., Carnegie Learning (2025), https://discover.carnegielearning.com/hubfs/PDFs/Whitepaper%20and%20Guide%20PDFs/2025-AI-in-Ed-Report.pdf
2. Diliberti, M.K., Schwartz, H.L., Doan, S., Shapiro, A., Rainey, L., Lake, R.: Using artificial intelligence tools in K–12 classrooms. Tech. Rep. RR-A956-21, RAND Corporation (2024), https://www.rand.org/pubs/research_reports/RRA956-21.html
3. Firmender, J.M., Reis, S.M., Sweeny, S.M.: Reading comprehension and fluency levels ranges across diverse classrooms: The need for differentiated reading instruction and content. Gifted Child Quarterly **57**(1), 3–14 (2013)
4. Google DeepMind: Learnlm: Improving Gemini for learning. arXiv preprint arXiv:2412.16429 (2024), https://arxiv.org/abs/2412.16429
5. MagicSchool AI: Magicschool: AI platform for school districts (2024), https://www.magicschool.ai, accessed 2026
6. Smirnova, P., Chun, B.W.: Text simplification for children: Evaluating LLMs vis-à-vis human experts. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25). ACM (2025). https://doi.org/10.1145/3706599.3719889
7. Wolfe, M.B.W., Schreiner, M.E., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W., Landauer, T.K.: Learning from text: Matching readers and texts by latent semantic analysis. Discourse Processes **25**(2–3), 309–336 (1998)