📖 LitLab

Alignment Is Not a Prompt.

Why curriculum-aligned decodable text is a verification problem — and what that means for schools evaluating general-purpose AI assistants.

litlab.ai

August 2026

01 The right question

If a capable general-purpose model can do a job, you should use it for that job. That instinct is correct, and it is why this question deserves a real answer rather than a feature list.

So let us state it precisely. The question is not *can a general-purpose assistant write a short story for a first grader?* It plainly can. The question is narrower:

> **Can it guarantee that every word in that story is decodable using only the letter-sound correspondences this student has actually been taught — and can it show you the proof?**
>
> **That is the whole question.** Teacher trust, instructional fidelity, and whether a coach can hand the same material to twelve teachers all depend on the answer.

The answer is no, and the reason is structural rather than a matter of model quality. Wrapping a generic LLM in custom instructions changes what the model is *asked* to do. It does not supply the two things this task requires: a machine-readable representation of a program's cumulative scope and sequence, and a deterministic checker that verifies output against it before a teacher sees the page.

02 Decodability is a relation, not a property

Everything rests on this point, and it is routinely missed. **A text is not "decodable" in the abstract.** Decodability has two components: the proportion of phonically regular words, and — critically — the match between the letter-sound relationships in the text and those the reader *has already been taught*.[1] The literature formalizes the second as **lesson-to-text match**.[2]

The consequence is unavoidable: decodability is **reader-specific and curriculum-position-specific**. The passage on the next page is 57% decodable for a student at UFLI Foundations Lesson 45, and fully decodable by Lesson 116. Nothing about the text changes — only the inventory of correspondences the reader can bring to it.

So "write me a decodable passage" is an **underspecified request**. The well-formed version — *write a passage decodable at Lesson 45 of UFLI Foundations* — is no longer a writing task. It is a search under a hard whitelist of permitted graphemes and permitted irregular words, followed by verification.

> ℹ Readability formulas — Lexile, DRP, Spache — estimate difficulty from word frequency and sentence length; none of them represents grapheme-phoneme correspondence. A passage can score perfectly on every readability index and remain unreadable for a student who has not been taught VCe.[3]

03 What the failure actually looks like

The failure mode is not gibberish. It is **fluent, charming prose that looks entirely correct** to anyone not actively holding the cumulative inventory in mind. Below: a student at UFLI Foundations Lesson 45, where sh has just been introduced.

EXHIBIT A **A plausible AI response — and what a checker sees**

Prompt: *"Write a decodable passage for a first grader practicing the digraph sh."*

Shane[1] and Trish **rushed**[2] to the **shore**[3]. Trish **wished**[2] **for**[3] the **bright**[4] shells. Shane had **brought**[5] a **shovel**[6] in the **sunshine**[1].

1 VCe pattern — Lessons 54–62
2 Suffix **-ed** — Lesson 64
3 R-controlled vowel — Lessons 77–83
4 Vowel team **igh** — Lesson 87
5 Low-frequency **ough** — Lesson 116
6 Two-syllable schwa — Lesson 63+

9 of 21 words fall outside the taught inventory — 57% decodable. Eight words carry the requested sh, so the passage satisfies the literal instruction, and it reads beautifully. It also asks a six-year-old who knows sh, short vowels and single consonants to attempt *brought*, *shovel* and *sunshine* — words where decoding is not yet an available strategy, leaving guessing from picture or first letter as the only route through.[10] Lesson positions per the published UFLI sequence.[4]

The error is silent. Auditing decodability by eye means reconstructing the cumulative inventory for a specific position in a specific program and checking every word against it — expert work, and slow. Unverified output passes review by default.

And it is directionally predictable. Asked to hit a target grapheme, a model reaches for words containing it — and the words richest in any pattern tend to be longer, rarer and morphologically complex. The pressure runs *away* from controlled text.

The standards treat this as a defect, not a preference. The Reading League's *Curriculum Evaluation Guidelines* flags as a red flag early texts that "include phonic elements that have not been taught," and lists as aligned practice phonics applied "in decodable texts that match the phonics elements taught."[5]

04 This has been measured

A 2025 peer-reviewed evaluation in *Learning Disabilities Research & Practice* tested six AI tools — including **Google Bard/Gemini**, ChatGPT 3.5 and 4.0, two Llama 2 deployments and Diffit — across ten target letter combinations.[6] Passages requested as "highly decodable" averaged a **grade 6.4** reading level, some reaching ninth and tenth grade, by the same mechanism as above: tools hit their target by **selecting complex multisyllabic words**.

> **"At this time, we cannot recommend that AI tools be used to request decodable text, at least in the manner prompted in this study… when decodable text is desired, we recommend that educators continue to rely on professionally created materials."**

Shen, Kane-Cabello, Candelaria, Stratford & Clemens (2025) · *Learning Disabilities Research & Practice*, 40(4), 191–204 — peer-reviewed. The authors scope this carefully: they note that other prompting strategies, or **models tuned with access to a large corpus of decodable text**, might perform differently. We agree — that describes the pipeline in §05, not a wrapper around a general-purpose model.

05 Three reasons a wrapper cannot close this

None of this says the models are weak — they are extraordinary. It says this task sits across three of their structural seams at once.

1 The constraint is sub-lexical; the model is not.

Models operate over subword tokens that do not respect grapheme boundaries — `ship` may be one token, and the digraph `sh` has no guaranteed independent representation. On PhonologyBench, GPT-4 scored **23.3% on syllable counting against a 90.0% human baseline**.[7] These models can *spell*; what they do unreliably is *reason over* graphemes as a closed set to be checked against — exactly what decodability requires.

2 There is no single constraint set — there are hundreds.

A prompt encodes one instructional position, imprecisely. Alignment needs every position in every program a school uses, each defining a distinct cumulative allowed-grapheme set *and* allowed-irregular-word list.

PROGRAM	DISCRETE INSTRUCTIONAL POSITIONS
UFLI Foundations	128 numbered lessons, 121 with a decodable passage
Fundations (Wilson)	45 units across Levels 1–3, plus Level K
CKLA Skills	~337 lessons across K–1 alone, GPC-specified per lesson

Three programs, partial grade bands, already **500+ distinct valid constraint sets** — before Reading Horizons, Benchmark or Superkids, and before student-level variation. For a wrapper to be aligned, **the teacher must first encode the sequence themselves** — doing the analysis they adopted a tool to avoid.

3 Generation without verification cannot be trusted.

Even given a perfect constraint set, satisfying hard constraints is a measured weakness — on the COLLIE benchmark, GPT-4 satisfied roughly **50.9%** of constraints zero-shot.[8] More decisively, the model has **no mechanism to check its own output**: it cannot tell you it failed, because it cannot tell. Reliability comes from **a deterministic checker outside the model** — parse every word into graphemes, test each against the cumulative inventory for that position, score, regenerate below threshold. A database-and-verification pipeline, not a set of instructions.

06 Side by side

CAPABILITY	GENERIC LLM WITH CUSTOM INSTRUCTIONS	LITLAB
Machine-readable scope & sequence	✗ Hand-encoded by the teacher, per position, per program	✓ Six programs mapped to lesson level
Decodability verified, and a score reported	✗ No checker in the pipeline, and no ground truth to score against	✓ Every word grapheme-parsed and scored, per lesson position
Consistency across a grade team	✗ Varies by teacher, by prompt, by run; leaves when the teacher does	✓ Shared library that persists as an institutional asset
Print-ready books & activities	✗ Chat output requires manual layout	✓ Formatted, printable, with comprehension
Evidence for this specific task	✗ Independently evaluated and not recommended[6]	✓ Published, reproducible alignment method

07 What teachers actually get

Four capabilities sit on the alignment engine. Each exists because it requires knowing a student's position in a sequence.

1 · A library filterable to today's lesson

1,500+ decodables tagged to positions across UFLI, Fundations, CKLA, Reading Horizons, Benchmark and Superkids. Taught Unit 7 this morning? Filter to Unit 7 and get texts already verified against that inventory — the same one every teacher on the grade team gets.

2 · Constrained generation, checker in the loop

For what the library lacks — a student's name, a class pet, a topic mid-unit — teachers generate a custom decodable, bounded by the same verified inventory and scored before it returns.

3 · Vocabulary sorted by instructional role

Words pre-sorted into four categories *defined* relative to sequence position — the same word is a target at one lesson, review at a later one. A generic LLM can produce a plausible-looking version of the table below; not a *correct* one.

4 · Print-ready books and activities

Formatted to print and put in students' hands, with comprehension questions attached. A chat window returns text that still needs copying, reformatting and laying out — which is where a meaningful share of the claimed time saving goes.

TARGET SKILL	REVIEW	HIGH FREQUENCY	PRE-TEACH
ship · shop · fish · shell	crab · sled · hand · stop	the · to · they · was	water · ocean

08 What we measure

Because the checker exists, LitLab publishes decodability across the full UFLI sequence — a number a tool without a checker cannot produce.[9]

>90%	95%	<5%	96%
MEDIAN DECODABILITY, ALL 128 UFLI LESSONS	EXCEEDED BY MOST INDIVIDUAL SKILLS	UFLI-ALIGNED STORIES BELOW 86% DECODABLE	LIBRARY-WIDE MEDIAN DECODABILITY

A generic LLM is a better instruction-follower. It is not a scope-and-sequence database, and it is not a decodability checker.

The question worth asking any vendor here is not *can you generate a decodable story* — everything can now. It is: **what is your alignment source, what does your checker verify, and what decodability score can you report for this text at this lesson?** Those three answers separate a verified instructional material from a plausible one, and nothing about a generic LLM's trajectory changes that — the missing pieces are not model capabilities, they are infrastructure. We would rather be evaluated on those three questions than on a feature list.

SOURCES

1·2 Mesmer (2001), *Reading Research & Instruction* 40(2), 121–142; Pugh, Kearns & Hiebert (2023), *RRQ* — lesson-to-text match.

3 Hiebert & Pearson (2010), Reading Research Report 10.01, TextProject.

4 UFLI Foundations scope & sequence, Lessons 1–128, Univ. of Florida Literacy Institute. Exhibit A constructed for illustration.

5 The Reading League (2023), *Curriculum Evaluation Guidelines*, §1.24 (red flag) and §1.35 (aligned practice).

6 Shen, Kane-Cabello, Candelaria, Stratford & Clemens (2025), *Learning Disabilities Research & Practice* 40(4), 191–204.

7 Suvarna, Khandelwal & Peng (2024), PhonologyBench, KnowLLM Workshop, ACL 2024.

8 Yao, Chen, Hanjie, Yang & Narasimhan (2024), COLLIE, ICLR 2024.

9 LitLab's own analysis: decodability = % of words decodable using patterns taught to that point, against UFLI's phoneme-grapheme progression, over thousands of lesson-tagged stories. ESSA Tier IV ("demonstrates a rationale") is ESSA's entry tier — a research-grounded logic model plus ongoing study, not an efficacy finding. The tiers condition certain federal funds and bind their recipients.

10·11 Juel & Roper/Schneider (1985), *RRQ* 20(2), 134–152; Cheatham & Allor (2012), *Reading and Writing* 25(9), 2223–2246.

* Scope: teacher-facing surface only. Student oral-reading practice, fluency analysis and progress monitoring are documented separately.