Diffit Quality Rubric

A scoring rubric for evaluating AI-generated instructional materials, derived from the Diffit Quality Constitution and the WestEd four-category framework for instructional materials quality (Bugler et al., 2017).

How to use this rubric

For each criterion, score the artifact:

- **Meets** — clearly satisfies the commitment on this artifact, with evidence
- **Partial** — partly satisfies; minor gaps a teacher could correct in seconds
- **Fails** — does not satisfy; a teacher would have to substantially rework

Every score must cite specific evidence — a quote, a question number, a page reference, a screenshot moment. Unsourced scores don't belong in this comparison.

N/A is honest. If a criterion doesn't apply to the prompt (e.g., source fidelity when no source was provided), mark N/A rather than forcing a score.

Diffit will not score perfectly. The Constitution is forward-looking; some commitments are aspiration. We score Diffit honestly and let the gap to competitors speak.

---

1. Accuracy and Craft

> The facts are right, the layout is clean, and a teacher will not be embarrassed in front of the class.

1.1 Factual correctness

Are facts, numbers, dates, names, units, and scientific claims accurate at a level a student at the named grade can verify? Does math compute?

Score	Anchor
Meets	Spot-check finds no factual errors in passage, vocab definitions, questions, or answer key
Partial	One minor, non-load-bearing inaccuracy (e.g., a date is the year before the actual year)
Fails	A factual claim a student or family could catch — wrong process, wrong number, wrong scientific mechanism

1.2 Source fidelity

When a source is provided (URL, PDF, text), does the material work *from* it without invention or substitution? Are quotations and data accurate to the source?

Score	Anchor
Meets	Every substantive claim traces to the source; no facts smuggled in from elsewhere
Partial	Mostly source-grounded; one or two details added or implied beyond the source
Fails	Invents details not in the source, fabricates quotes, substitutes a different source, or rewrites the problem
N/A	No source provided in the prompt

1.3 Mechanical correctness

Spelling, grammar, punctuation, capitalization. The boring stuff.

Score	Anchor
Meets	Clean enough to hand a student as a final draft
Partial	1–2 minor errors a teacher would not flag in a colleague's draft
Fails	Visible errors a teacher must correct before assigning

1.4 Layout discipline (page density, graphics, ink)

Does the page earn its length? Density, whitespace, font sizing scaled to grade. Graphics only when they do pedagogical work, not for decoration.

Score	Anchor
Meets	Density matches grade band; no padding; every graphic is informative
Partial	Usable but visually inefficient (e.g., 3 pages of what fits on 1; one decorative graphic)
Fails	Wall of text, padded with filler, or decorative graphics that distract

1.5 Answer-key consistency

Does the answer key actually answer the questions in *this* artifact, in the form the questions imply?

Score	Anchor
Meets	Every answer matches a question in this artifact; form (fraction/decimal, full sentence, etc.) matches
Partial	Most match; one or two minor drifts in form or numbering
Fails	Missing answers, wrong questions answered, or no key produced when one was expected
N/A	No answer key requested or implied

---

2. Standards Alignment and Depth of Knowledge

> The material hits the standard the teacher named, at the depth that standard requires — not a watered-down version of it.

2.1 Standards alignment

Is a specific standard cited (NGSS, CCSS, state-specific), and does the content actually meet it (not just "cover the topic")?

Score	Anchor
Meets	Explicit standard cited, and content addresses the verbs in the standard (e.g., NGSS 5-ESS2-1 asks for a model — the artifact has students build one)
Partial	Topic-correct, but no standard cited, or weakly aligned
Fails	No standard cited and content drifts off-topic, or claims a standard but does not meet its cognitive demand

2.2 Cognitive depth (DOK spread)

Across the questions in the packet, is there a real spread of cognitive demand, or all recall?

Score	Anchor
Meets	At least one question at DOK 3+ (analyze, evaluate, construct argument from evidence) alongside foundational recall
Partial	Mostly recall with one stretch question
Fails	All recall / one cognitive level only

2.3 Genuine variety

Do different question *formats* (MC, short answer, open-ended, matching) ask for meaningfully different *thinking*, or just the same recall in different visual layouts?

Score	Anchor
Meets	The thinking varies across questions, not just the format
Partial	Format varies; thinking varies somewhat
Fails	Same shallow recall asked three different ways

2.4 Question quality

Are multiple-choice distractors plausible, related to the content, and educative (revealing common misconceptions)? Does the position of the correct answer vary?

Score	Anchor
Meets	Distractors are content-relevant and educative; correct-answer positions vary across the set
Partial	One of the two: educative distractors but predictable positions, or varied positions but throwaway distractors
Fails	Throwaway distractors ("none of the above," nonsense options), or "always B" patterns

2.5 Honesty about what the standard requires

When a standard asks for extended writing, modeling, or argument from evidence, does the artifact actually demand that work — or is it satisfied with a fill-in-the-blank?

Score	Anchor
Meets	Where the standard demands depth, the artifact provides it (essay prompts, modeling tasks, evidence-based argument prompts)
Partial	Some depth, but the hardest cognitive moves are routed around
Fails	A standard requiring extended writing or modeling is satisfied with recall-level tasks
N/A	No standard cited

---

3. Ease of Use and Completeness

> A teacher can walk into the classroom and use this. No missing pieces, no internal inconsistencies, no "the teacher should now explain..." placeholders.

3.1 Multi-activity coherence

When the packet contains a passage + quiz + vocab + key, do they reference the same content — same details, same vocabulary, same emphasis?

Score	Anchor
Meets	The packet reads as a single resource; quiz asks about this passage; vocab drawn from this passage
Partial	Mostly coherent; minor drift (e.g., vocab not all drawn from passage)
Fails	Questions ask about content not in the passage; vocab unrelated; pieces feel independently generated

3.2 Complete and ready to teach

Can a teacher use the artifact as-is without hunting for missing pieces or filling in placeholders?

Score	Anchor
Meets	No placeholders, no "[insert passage here]," no "the teacher should now explain..."
Partial	Minor gaps a teacher fixes in under a minute
Fails	Stage directions, placeholders, or instructions to "find a source and fill it in"

3.3 Considered defaults

Does a single, simple prompt produce a usable output? Or does the teacher need to know the right follow-up prompts to get there?

Score	Anchor
Meets	One prompt, usable output. Early-career teacher could ship it
Partial	Usable with 1–2 obvious follow-ups
Fails	Requires significant prompt engineering / iteration to get to "usable"

3.4 Honest layout for student work

Does the space for student writing match what they're being asked to produce? (Essay → real lines; one-word answer → one line; show-your-work math → room for steps.)

Score	Anchor
Meets	Writing space appropriate to each task
Partial	One task miscalibrated (e.g., 3 lines for an essay)
Fails	Multiple tasks miscalibrated, or no writing space provided where it's needed

3.5 Classroom-workflow format fidelity

Does the output land cleanly in the formats teachers actually use — print-ready PDF, Google Docs, Google Slides, Forms?

Score	Anchor
Meets	Exports / prints cleanly to a teacher-ready file in at least two formats
Partial	Usable in one format; breaks in others
Fails	Copy-paste artifact; formatting collapses; no export path

---

4. Engagement and Meeting Student Needs

> The materials meet students where they are — reading level, language, prior knowledge, cultural context — and engage them rather than droning at them.

4.1 Right reading level

Sentence length, clause complexity, vocabulary, and conceptual density match the named grade.

Score	Anchor
Meets	Reads like the named grade — short clear sentences for younger, compound sentences and academic vocabulary (introduced with context) for older
Partial	Slightly above or below the named grade
Fails	Clearly miscalibrated — either dumbed-down (deleting content) or unreadable for the grade

4.2 Engaging voice

Does the passage hold a student's attention, or read as generic "AI-written" textbook drone?

Score	Anchor
Meets	Concrete examples, varied sentence structure, a voice a student would not skim past
Partial	Serviceable but flat
Fails	Generic, repetitive, or filler ("In this passage, we will learn about...")

4.3 Differentiation in the workflow

Can the same lesson be produced at multiple reading levels / languages / scaffolds from the same brief, without rebuilding?

Score	Anchor
Meets	Multiple differentiated versions delivered or one click away (e.g., 3 reading levels, ELL scaffold, translation)
Partial	Differentiation possible but requires separate, manually-crafted prompts
Fails	One version only; no path to differentiation in the workflow

4.4 Pedagogically-useful visuals

For topics where a diagram carries conceptual load (water cycle, photosynthesis, geometry, data interpretation), is the visual present, accurate, and grade-appropriate?

Score	Anchor
Meets	Diagram present, accurately labeled, scientifically correct, clarifies the concept
Partial	Present but generic, or partly inaccurate
Fails	No diagram for a topic that needs one; or a diagram that misrepresents the science / has hallucinated text artifacts

4.5 Asset-based framing and safety

Do examples, names, and contexts draw from broad cultural backgrounds rather than assuming a default student? Is content age-appropriate?

Score	Anchor
Meets	Examples draw from broad contexts; no default-student framing; nothing age-inappropriate
Partial	Neutral but bland; minor default-student assumptions
Fails	Excluding / stereotyping defaults, or age-inappropriate content

---

Scoring summary template

A side-by-side comparison should end with a heatmap like this:

Criterion	Diffit	Competitor
1.1 Factual correctness	✓	✓
1.2 Source fidelity	✓	⚠
…	…	…

Followed by:

- **Net assessment** — one paragraph
- **3–5 "look here" moments** — the specific evidence callouts an admin should look at

- **What we don't claim** — criteria where Diffit underperforms or where the difference is small. Honesty is the differentiator.

Citation

This rubric operationalizes the Diffit Quality Constitution, itself drawing on:

Bugler, D., Marple, S., Burr, E., Chen-Gaddini, M., & Finkelstein, N. (2017). *How Teachers Judge the Quality of Instructional Materials: Selecting Instructional Materials, Brief 1 — Quality.* WestEd. https://www.wested.org/resource/selecting-instructional-materials-brief-1-quality/

Diffit vs. ChatGPT

A side-by-side comparison of Diffit and a general-purpose assistant (ChatGPT), each generating the same instructional packets across grade bands, scored against the Diffit Quality Rubric — which is derived from the Diffit Quality Constitution and WestEd's research on how teachers judge instructional-materials quality.

Audience: District and school leaders whose teachers already use ChatGPT and want to understand what Diffit adds for supplemental materials and differentiation.

Approach: Three identical prompts plus a differentiation task, each run once in each tool, scored on observable artifact qualities. In Diffit the topic, grade, standard, and source were entered as structured fields; in ChatGPT the equivalent was sent as a single chat message. The first/default output was scored.

What kind of contest this is. ChatGPT is a capable general-purpose assistant, and on the two straightforward worksheet prompts it shows: its math set is accurate and genuinely on-standard, and its Harriet Tubman packet is complete and genuinely RI.5.3-aligned. We credit that explicitly. The gap to Diffit opens on three fronts: **source fidelity** (working from the source the teacher provided, versus a generic summary that even cites Wikipedia), **the depth the named standard requires** (visual models and reasonableness for 5.NF.A.2), and — most sharply on the high-school prompt — **completeness** (a reading-analysis worksheet that never includes the reading).

How this maps to the summary page: the four criteria shown there — *Classroom-ready & standards-aligned*, *Complete & ready to teach*, *Content integrity*, and *Differentiation that holds up* — are presentation themes drawn from the 20-criterion rubric. This document is the full per-criterion scoring behind them.

TL;DR

When a teacher asks each tool for the same packet, **both can produce a clean, complete worksheet on the straightforward prompts — but Diffit works from the source the teacher provided, meets the depth the named standard requires, and (on the MLK prompt) includes the reading the tasks depend on, while ChatGPT delivers a competent-looking worksheet that flattens the source, skips the standard's modeling and reasonableness work, and — on the high-school prompt — omits the text entirely.**

	Diffit	ChatGPT
ELA packet (Harriet Tubman, NPS source, RI.5.3)	Reading passage built from the source + 3 distinct activities + 2-page answer key + 2 contextual images; interaction-analysis activity meets RI.5.3	ChatGPT's strongest output: complete and genuinely RI.5.3-aligned (cause-effect chart, relationship prompts). But the passage is a generic summary that cites "(Wikipedia)" despite the NPS source, and there is no image
Math packet (fractions, 5.NF.A.2)	Visual fraction models + unlike-denominator word problems + estimate-and-check / error-analysis activity + full answer key	Accurate, on-standard word problems with real unlike denominators and an error-analysis prompt — but no visual models (which the standard names) and no estimate-and-reasonableness step
HS packet (MLK Letter from Birmingham Jail, RH.9-10.2)	Built from the letter + multiple activity types + answer key that quotes the source	A reading-analysis worksheet with no reading: "use evidence from the text" with no text included. Question design is reasonable; the source it depends on is absent
Considered defaults (one-prompt-to-usable)	✓ across all three prompts	✓ on math and Tubman; on MLK the teacher must supply the missing letter first
Differentiation (2nd-grade relevel)	Keeps the full story and analysis at a Grade-2 reading level	Relevels by deleting — cut to a generic ~130-word summary, named detail and analysis dropped
Pattern overall	Wins on source fidelity, standards depth, completeness, and differentiation that holds up	Strong on the easy prompts; falls short on fidelity, the standard's depth, and (MLK) completeness

The single most important finding: ChatGPT is not a weak tool — on the math and Tubman prompts it produced accurate, standards-aligned, complete materials, and we score it that way. The divergence is about whether the worksheet works *from the source the teacher chose*, meets the *depth the standard names*, and — on the high-school prompt — includes the *text its own questions require*. That is where Diffit and ChatGPT separate.

We score Diffit honestly against its own (aspirational) Constitution and call out where Diffit falls short — a relevel oversimplification, a translation terminology slip, no images in the HS packet — so the wins it does post land with more weight.

What we tested

Three prompts spanning two grade bands, all CCSS-aligned, plus a differentiation task.

Prompt 1 — ELA

> **Topic:** Harriet Tubman **Grade:** 5 **Standard:** CCSS.ELA-LITERACY.RI.5.3 — *"Explain the relationships or interactions between two or more individuals, events, ideas, or concepts in a historical, scientific, or technical text based on specific information in the text."* **Source:** National Park Service biography of Harriet Tubman

Prompt 2 — Math

> **Topic:** Adding and subtracting fractions with unlike denominators **Grade:** 5 **Standard:** CCSS.MATH.CONTENT.5.NF.A.2 — *"Solve word problems involving addition and subtraction of fractions referring to the same whole, including cases of unlike denominators, e.g., by using visual fraction models or equations to represent the problem. Use benchmark fractions and number sense of fractions to estimate mentally and assess the reasonableness of answers."* **Source:** none (the standard is the brief)

Prompt 3 — HS ELA

> **Topic:** Martin Luther King Jr.'s "Letter from Birmingham Jail" **Grade:** 9–10 **Standard:** CCSS.ELA-LITERACY.RH.9-10.2 — *"Determine the central ideas or information of a primary or secondary source; provide an accurate summary of how key events or ideas develop over the course of the text."*
> **Source:** Stanford King Institute entry on the letter

The primary artifacts

Artifact	File	Pages
Diffit ELA — Harriet Tubman	diffit-ela-tubman.pdf	11
ChatGPT ELA — Harriet Tubman	chatgpt-tubman.pdf	5
Diffit Math — fractions	diffit-math-fractions.pdf	15
ChatGPT Math — fractions	chatgpt-math-fractions.pdf	5
Diffit HS — MLK Letter	diffit-ela-mlk.pdf	17
ChatGPT HS — MLK Letter	chatgpt-mlk.pdf	5
ChatGPT differentiation — Tubman releveled to Grade 2	chatgpt-tubman-2nd-grade.pdf	4

ELA scoring

Diffit ELA vs ChatGPT ELA (Harriet Tubman), against the rubric:

#	Criterion	Diffit	ChatGPT	Evidence
1.1	Factual correctness	✓ Meets	✓ Meets	Both are factually accurate against the NPS biography (born ~1822, escaped 1849, ~70 people led to freedom, Civil War nurse and scout). ChatGPT's facts check out
1.2	Source fidelity	✓ Meets	⚠ Partial	Diffit works from the NPS page in specific detail — Edward Brodess, the Bucktown two-pound-weight head injury, the Parson's Creek mariners, the Combahee River Raid. ChatGPT flattens the source to a generic summary and, in the student passage, cites "(Wikipedia)" for one paragraph despite being handed the NPS source — pulling in a source the teacher never chose
1.3	Mechanical correctness	✓ Meets	✓ Meets	Both are clean of spelling and grammar errors. (ChatGPT leaves raw inline "(National Park Service)" tags scattered through the student passage — a copy-edit nuisance a teacher would strip, not an error)
1.4	Layout discipline	✓ Meets	✓ Meets	Both paginate cleanly at an appropriate 5th-grade density
1.5	Answer-key consistency	✓ Meets	✓ Meets	Both include a complete answer key whose answers match the questions in the artifact
2.1	Standards alignment	✓ Meets	✓ Meets	Both name RI.5.3 and address its verb. ChatGPT includes a genuine cause-and-effect chain chart and an "explain the relationship between two of…" prompt — real relationship analysis, not just recall. Credit it
2.2	Cognitive depth (DOK)	✓ Meets	✓ Meets	Both span recall to analysis. ChatGPT mixes MC recall, short response, a cause-and-effect chart, and an open relationship-explanation challenge
2.3	Genuine variety	✓ Meets	✓ Meets	Both vary the thinking across question types, not just the format
2.4	Question quality	N/A	✓ Meets	Diffit's Tubman activities are open-ended (MC criterion N/A). ChatGPT's four MC items have content-relevant distractors and varied answer positions (B, C, C, A)
2.5	Honesty about what the standard requires	✓ Meets	✓ Meets	Both demand the relationship-explanation work RI.5.3 names — ChatGPT's Q8 asks students to "explain the relationship between two of: Harriet Tubman, the Underground Railroad, the Civil War, enslaved people… use evidence from the passage"
3.1	Multi-activity coherence	✓ Meets	✓ Meets	Both cohere around their own passage

#	Criterion	Diffit	ChatGPT	Evidence
3.2	Complete and ready to teach	✓ Meets	✓ Meets	Passage + questions + answer key, no placeholders, on both sides
3.3	Considered defaults	✓ Meets	✓ Meets	Both produce a usable worksheet from a single prompt
3.4	Honest layout for student work	✓ Meets	✓ Meets	Both give students writing space sized to each task
3.5	Classroom-workflow format fidelity	✓ Meets	✓ Meets	Both print cleanly to a teacher-ready PDF
4.1	Right reading level	✓ Meets	✓ Meets	Both read at roughly 5th-grade level
4.2	Engaging voice	✓ Meets	⚠ Partial	Diffit's passage is concrete and vivid (the weight to the head, the icy muskrat traps, the marshland skills, “never lost a passenger”). ChatGPT's is flat encyclopedic summary — “remembered as a hero whose determination and bravery changed many lives” — with inline citation tags interrupting the read
4.3	Differentiation in the workflow	✓ Meets	⚠ Partial	Both can produce differentiated versions — Diffit one-click, ChatGPT via a follow-up prompt. The relevel quality is the open question (see Differentiation in practice): ChatGPT's degrades
4.4	Pedagogically-useful visuals	✓ Meets	✕ Fails	Diffit embeds an authentic period photograph of Tubman plus a marshlands image that grounds the navigation-skills narrative the activities then analyze. ChatGPT includes no image — a missed scaffolding opportunity for a 5th-grade reading lesson
4.5	Asset-based framing	✓ Meets	✓ Meets	Both handle the subject respectfully and accurately

ELA scorecard: Diffit 17 Meets (out of 19 applicable; §2.4 N/A — Diffit's ELA activities are open-ended). ChatGPT 16 Meets, 3 Partials, 1 Fail (out of 20 applicable). This is ChatGPT's strongest output — complete, genuinely RI.5.3-aligned, with a real cause-and-effect chart. The gap to Diffit is source fidelity (the generic summary and the “(Wikipedia)” citation), the passage's flat voice, and the missing image.

Math scoring

Diffit Math vs ChatGPT Math (fractions), against the rubric:

#	Criterion	Diffit	ChatGPT	Evidence
1.1	Factual correctness	✓ Meets	✓ Meets	All computations check out in both artifacts. ChatGPT's answer key is arithmetically correct throughout (3/4 + 2/3 = 17/12, 7/8 + 1/3 = 29/24, 1 – 11/20 = 9/20, etc.)
1.2	Source fidelity	N/A	N/A	No source provided in either case
1.3	Mechanical correctness	✓ Meets	✓ Meets	Both clean
1.4	Layout discipline	✓ Meets	✓ Meets	Both clean and appropriately dense; ChatGPT's single worksheet is tidy and print-ready
1.5	Answer-key consistency	✓ Meets	✓ Meets	Both include a complete, correct answer key matching the questions
2.1	Standards alignment	✓ Meets	✓ Meets	5.NF.A.2 is about unlike denominators in word problems, and ChatGPT delivers exactly that — real unlike denominators (3/4 + 2/3, 7/8 + 1/3) framed as word problems with both addition and subtraction. Solidly on-standard
2.2	Cognitive depth (DOK)	✓ Meets	✓ Meets	Both reach beyond procedure. ChatGPT includes an error-analysis prompt (“a student added the denominators and got 2/5 — do you agree?”) and a create-your-own problem
2.3	Genuine variety	✓ Meets	⚠ Partial	Diffit's three activity types ask different thinking. ChatGPT's set is eight compute-the-sum word problems plus the two reasoning items — most of the sheet is the same thinking repeated
2.4	Question quality	✓ Meets	N/A	Diffit's “Fraction Freddy” names common misconceptions as educative content. ChatGPT's math worksheet uses no multiple choice (N/A)
2.5	Honesty about what the standard requires	✓ Meets	⚠ Partial	5.NF.A.2 also calls for visual fraction models and for students to use benchmark fractions to estimate and assess reasonableness. ChatGPT's directions say “use a visual model… if it helps” but provide none, and there is no estimation-or-reasonableness task — the standard's modeling and reasonableness moves are routed around
3.1	Multi-activity coherence	✓ Meets	✓ Meets	Both hang together around a consistent fractions context
3.2	Complete and ready to teach	✓ Meets	✓ Meets	Worksheet + key, no placeholders, on both sides
3.3	Considered defaults	✓ Meets	✓ Meets	Both produce a usable worksheet in one shot

#	Criterion	Diffit	ChatGPT	Evidence
3.4	Honest layout for student work	✓ Meets	✓ Meets	Both provide “show your work” space sized to the task
3.5	Classroom-workflow format fidelity	✓ Meets	✓ Meets	Both render and print cleanly
4.1	Right reading level	✓ Meets	✓ Meets	Word problems read at 5th-grade level in both
4.2	Engaging voice	✓ Meets	✓ Meets	Both use varied real-world contexts (Diffit: baking, hiking, pizza; ChatGPT: hiking, reading, ribbon, garden)
4.3	Differentiation in the workflow	✓ Meets	⚠ Partial	The workflow exists in both; ChatGPT's relevel quality is the open question (see Differentiation in practice)
4.4	Pedagogically-useful visuals	✓ Meets	✕ Fails	**5.NF.A.2 names visual fraction models in the standard text.** Diffit renders fraction bars and circles for students to shade. ChatGPT renders none — a teacher using it to fulfill 5.NF.A.2 has not delivered the modeling the standard requires
4.5	Asset-based framing	✓ Meets	✓ Meets	Both name a diverse set of students (Diffit: Maya, Leo, Sarah, Chloe; ChatGPT: Jamal, Maria, Emma, Sarah)

Math scorecard: Diffit 17 Meets (out of 17 applicable; §1.2 N/A). ChatGPT 14 Meets, 3 Partials, 1 Fail (out of 18 applicable; §1.2 and §2.4 N/A). ChatGPT's math is genuinely strong on the standard's core — real unlike denominators, real word problems, a real error-analysis prompt. The load-bearing gaps are §4.4 (the visual fraction models the standard names) and §2.5 (the estimate-and-check reasonableness the standard also names).

Differentiation in practice

Starting from the same 5th-grade Tubman packet, both tools were asked to produce a **2nd-grade relevel**. In Diffit this is a one-click workflow action; in ChatGPT it is a follow-up prompt (“relevel this to 2nd grade”). The question is what each path produces.

The headline finding

ChatGPT relevels by deleting. Both outputs reach a real Grade-2 reading level — ChatGPT's prose is clean and it adds an age-appropriate draw-and-write task, which we credit. But the relevel cut the lesson to

a generic ~130-word summary: Edward Brodess, the Bucktown injury, the marshland skills, the Combahee River Raid, and the entire relationship/cause-effect analysis are gone, replaced by recall multiple choice.

The Constitution's commitment is explicit: *"We do not 'dumb down' by deleting content. We find the simpler way to express the harder idea. A relevel returns roughly the same amount of content, organized in roughly the same way."*

Diffit's 2nd-grade relevel keeps the full story — Brodess, the dates, the marshland skills, the Combahee River Raid, the 750 freed — and the relationship-and-cause-effect activities, in simpler sentences. ChatGPT lowered the reading level by lowering the content.

Honest finding on Diffit's side

Against Diffit's 2nd-grade relevel: one Cause/Effect item flattens the Combahee River Raid to *"helps soldiers in a war. They go on a boat trip."* The 750 freed are still in the answer key, but a 2nd grader reading only that prompt gets a thin picture of what Tubman did. A single-item oversimplification, not a whole-relevel failure — but a real one, and we note it.

What this means for §4.3: ChatGPT scores **Partial**. The differentiation can be produced, but the relevel degrades the lesson rather than preserving it at a lower reading level.

Pattern at secondary: Letter from Birmingham Jail

A reasonable question: *"Does the pattern hold at high school, where the source is already at the target reading level?"* On this prompt the gap is at its widest — not because ChatGPT's questions are bad, but because the worksheet omits the reading its own questions require.

Scoring

#	Criterion	Diffit	ChatGPT	Evidence
1.1	Factual correctness	✓ Meets	✓ Meets	ChatGPT's Background blurb is accurate (April 1963, the Birmingham Campaign, King's arrest, the clergy's “unwise and untimely” charge). What is present is correct
1.2	Source fidelity	✓ Meets	✕ Fails	Diffit builds from the actual letter — the four-step campaign, the clergy's specific charges, King's own words (“injustice anywhere…”, “Wait” / “Never”, “wheels of inevitability”) — and its answer key quotes the source. ChatGPT reproduces none of the letter; it cites the King Institute but never works from it, and its questions stay generic enough to write before reading a word of the text
1.3	Mechanical correctness	✓ Meets	✓ Meets	Both clean
1.4	Layout discipline	✓ Meets	✓ Meets	Both paginate cleanly at a high-school density
1.5	Answer-key consistency	✓ Meets	⚠ Partial	Diffit's answer key quotes the letter (Q4: “King explicitly states: ‘This Wait has almost always meant Never’”). ChatGPT's “Teacher Answer Key (Brief)” is generic central-idea statements with no quotations — there is no source text for it to quote
2.1	Standards alignment	✓ Meets	⚠ Partial	RH.9-10.2 asks students to determine central ideas and trace how they develop over the course of the text. ChatGPT's questions are reasonably designed (central ideas, a development chart, a summary task) but operate on a text that isn't in the packet, so the alignment is nominal rather than real
2.2	Cognitive depth (DOK)	✓ Meets	✓ Meets	ChatGPT's question design reaches analysis — identify two central ideas and explain how each develops, then summarize the development — which is genuine DOK-3 design, even with the text missing
2.3	Genuine variety	✓ Meets	✓ Meets	ChatGPT fields varied formats: vocabulary-in-context, central-idea analysis, text-evidence prompts, summary writing, a constructed response, and an exit ticket
2.4	Question quality	✓ Meets	N/A	Diffit's MC distractors are educative. ChatGPT's MLK worksheet uses no multiple choice (N/A)
2.5	Honesty about what the standard requires	✓ Meets	⚠ Partial	Analyzing how King's argument *develops* requires the developing argument — the letter. ChatGPT asks for exactly that work but never supplies the text to do it on, so the

#	Criterion	Diffit	ChatGPT	Evidence
				standard's central demand cannot be met as the worksheet stands
3.1	Multi-activity coherence	✓ Meets	✕ Fails	Diffit's activities operate on content embedded on the page. ChatGPT's tasks repeatedly reference “the text” the packet never includes — the pieces don't cohere into a usable whole
3.2	Complete and ready to teach	✓ Meets	✕ Fails	**The load-bearing finding: ChatGPT shipped a reading-analysis worksheet with no reading.** “Use context clues from the text” (Part 1) and “find evidence from the text” (Part 4) point to a passage that is nowhere in the packet. A teacher must source and paste the letter before any of it is usable
3.3	Considered defaults	✓ Meets	⚠ Partial	Diffit is usable from one prompt. ChatGPT's output requires the teacher to supply the missing letter before it can be assigned
3.4	Honest layout for student work	✓ Meets	✓ Meets	Both provide writing lines for the constructed responses
3.5	Classroom-workflow format fidelity	✓ Meets	✓ Meets	Both render and print cleanly
4.1	Right reading level	✓ Meets	✓ Meets	Both pitched at a high-school academic register
4.2	Engaging voice	✓ Meets	⚠ Partial	Diffit uses primary-source quotation and narrative context. ChatGPT's Background is serviceable but generic, and the analysis tasks have no text to bring a voice to
4.3	Differentiation in the workflow	✓ Meets	⚠ Partial	Differentiation workflow exists in both; not exercised on this prompt
4.4	Pedagogically-useful visuals	⚠ Partial	⚠ Partial	Neither produced an image, defensible for a text source. But Diffit at least renders argument-structure organizers (a four-step nonviolence flow map and a consequences-of-inaction map). ChatGPT's “Beginning / Middle / End” development chart is an empty, unscaffolded list of labels
4.5	Asset-based framing	✓ Meets	✓ Meets	Both handle civil-rights content with appropriate seriousness

HS scorecard: Diffit 18 Meets, 1 Partial (out of 19 applicable). ChatGPT 9 Meets, 7 Partials, 3 Fails (out of 19 applicable; §2.4 N/A). The three Fails — §1.2 source fidelity, §3.1 coherence, §3.2 completeness — all trace to the same root: the letter the worksheet analyzes is never included.

The honest read

ChatGPT's MLK question *design* is reasonable — central ideas, idea development, a summary task, all plausibly aligned to RH.9-10.2 on paper. The failure is not the standard, it is completeness: a reading-analysis worksheet that ships without the reading, with tasks ("use context clues from the text," "find evidence from the text") pointing at a passage that is not in the packet. Diffit weaves King's actual words and structure into the activities and answer key, so a teacher can hand it out as-is.

The five "look here" moments

If a district admin reads nothing else, these are the moments to look at.

1. A reading-analysis worksheet with no reading

ChatGPT's MLK worksheet asks students to "use context clues from the text to define each term" (Part 1) and to "find evidence from the text" (Part 4) — but the Letter from Birmingham Jail is nowhere in the packet. The worksheet jumps from a short background blurb straight to the questions. As delivered, a student cannot complete it; a teacher has to find and paste the letter first.

2. "(Wikipedia)" cited in the student passage — despite the NPS source

ChatGPT's Tubman passage attributes a paragraph to "(Wikipedia)" and sprinkles "(National Park Service)" through the student text, even though the prompt named the National Park Service source. Diffit works from the NPS page in specific detail (Brodess, the Bucktown weight, the mariners, the Combahee River Raid); ChatGPT flattens it to a generic summary and pulls in a source the teacher never chose.

3. The math worksheet names visual models — and renders none

5.NF.A.2 names visual fraction models and asks students to estimate and check reasonableness with benchmark fractions. ChatGPT's directions even say "use a visual model… if it helps," then provide none, and there is no estimate-and-check step anywhere. Diffit renders fraction bars and circles and builds an estimate-then-solve activity. ChatGPT's problems are accurate — but the two things the standard names beyond computation are missing.

4. The 2nd-grade relevel deletes the lesson

Asked to relevel the Tubman packet to Grade 2, ChatGPT cut it to a generic ~130-word summary and replaced the relationship analysis with recall multiple choice. Brodess, the marshland skills, the Combahee

River Raid — all gone. Diffit keeps the same content and analysis at the lower reading level. A relevel should lower the reading level, not the content.

5. What ChatGPT does well — and where it still falls short

Credit where due: ChatGPT's math is accurate and genuinely on-standard (real unlike denominators, a real error-analysis prompt), and its Tubman packet is complete and RI.5.3-aligned, with a real cause-and-effect chart and a complete answer key. This is not a broken tool. The Diffit advantage is not “complete vs. incomplete” on the easy prompts — it is fidelity to the chosen source, the depth the standard names, and a relevel that holds up. On the high-school prompt, where the source matters most, it becomes completeness too.

What we don't claim

ChatGPT is a capable assistant. On the math and Tubman prompts it produced accurate, standards-aligned, complete materials, and we score it that way (16/20 on Tubman, 14/18 on math). The divergence is about source fidelity, standards depth, and — on MLK — completeness, not raw capability.

ChatGPT's MLK question design is reasonable. Its central-idea, development, and summary tasks are plausibly aligned to RH.9-10.2. The failure is that the worksheet omits the text those tasks require — an incompleteness problem, not a wrong-standard problem.

ChatGPT did not invent facts. Every artifact is factually accurate. The closest thing to a fidelity error is importing Wikipedia into a passage that was supposed to come from the NPS source.

Three prompts is not a complete evaluation. This covers ELA, math, and HS ELA with one prompt each, plus one differentiation task. A district considering both should run its own comparison on its own prompts.

This is enablement, not academic research. The rubric is grounded in WestEd's framework and the Diffit Constitution, but scoring was performed by a single evaluator, and the prompts were chosen to exercise the rubric well.

What this comparison shows about the underlying products

The difference is chat-box-versus-structured-tool. ChatGPT takes a single free-text prompt and returns a single best-guess artifact; it treats the standard code as a label and the source as a citation to mention, rather than as constraints the content must satisfy. On simple prompts that still yields a good worksheet. On a source-grounded reading lesson, it yields questions about a text it didn't include.

Diffit treats the standard and source as inputs that shape the content. The standard's verb drives the activity design (interaction analysis for RI.5.3; visual models for 5.NF.A.2), the source is worked *from* rather than merely cited, and the reading is part of the packet. That is why the same clean-worksheet baseline diverges into “meets the standard, from the source” versus “labels the standard, summarizes the source.”

And when a teacher needs the lesson at another reading level, Diffit relevels by keeping the content and simplifying the language — where ChatGPT relevels by deleting.

Methodology — reproducing this comparison

Prompts

ChatGPT: each artifact was generated from a single chat message naming the topic, grade, and standard, with the source URL pasted in for the ELA prompts (math: “Make me a worksheet for 5th grade, aligned to the standard CCSS.MATH.CONTENT.5.NF.A.2”; Tubman and MLK named the NPS and Stanford King Institute sources respectively; differentiation: “relevel this to 2nd grade”). The first response was captured.
Diffit: topic, grade, standard (from the Standards picker), and source URL entered as structured inputs, generated once.

Conditions

- Diffit and ChatGPT, June 2026.
- First responses captured before any follow-up prompting.
- All artifacts exported as PDF.
- Scoring against rubric.md, one evaluator pass with evidence cited per criterion. Diffit's per-criterion column reuses the same scoring as the Diffit vs. Gemini and Diffit vs. MagicSchool comparisons, since the Diffit artifacts are identical.

Artifacts archived

All artifacts are in artifacts/ (Diffit: diffit-ela-tubman, diffit-math-fractions, diffit-ela-mlk, plus the 2nd-grade variant; ChatGPT: chatgpt-tubman, chatgpt-math-fractions, chatgpt-mlk, chatgpt-tubman-2nd-grade).

Citation

This comparison applies the Diffit Quality Constitution (drawing on WestEd's research into how teachers judge instructional-materials quality) to a side-by-side artifact comparison between Diffit and ChatGPT.

Bugler, D., Marple, S., Burr, E., Chen-Gaddini, M., & Finkelstein, N. (2017). *How Teachers Judge the Quality of Instructional Materials: Selecting Instructional Materials, Brief 1 — Quality.* WestEd. https://www.wested.org/resource/selecting-instructional-materials-brief-1-quality/