

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. MagicSchool

A side-by-side comparison of two AI tools *built for educators*, 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 who already have MagicSchool and want to understand what Diffit adds.

Approach: Three identical prompts, each run once in each tool, scored on observable artifact qualities. For consistency, every MagicSchool artifact was produced with its **Worksheet Generator** tool, which yields one worksheet per prompt (passage + question sets + answer key).

Why this is a closer contest than a general-purpose chatbot. MagicSchool is purpose-built for teachers, and it shows: it clears the hygiene bar a general chatbot routinely fails. It paginates cleanly, it produces a real reading passage, and it includes a real, correct answer key on every worksheet. We credit all of that explicitly. The gap to Diffit opens one layer deeper — in **standards depth** (the cognitive demand the named standard actually requires), **source fidelity** (working from the real source versus a generic paraphrase), and **usability as a printed handout** (room for students to do the work the worksheet asks for).

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 produce a clean, complete, printable worksheet — but Diffit produces a packet that meets the named standard at depth, works from the actual source, and gives students room to do the work, while MagicSchool produces a tidy worksheet that labels the standard without reaching its cognitive demand and leaves no space to write.**

	Diffit	MagicSchool
ELA packet (Harriet Tubman, NPS source, RI.5.3)	Reading passage built from the source + 3 distinct activities + 2-page answer key + 3 contextual images; interaction-analysis activity meets RI.5.3	Clean 2-page worksheet: generic paraphrase + FIB/MC/open-ended + answer key. Recall-level; no writing space; no image
Math packet (fractions, 5.NF.A.2)	Visual fraction models + unlike-denominator word problems + error-analysis activity + full answer key	Clean worksheet + correct key, but mostly <i>like</i> denominators, no visual models, and 2 of 3 open-ended items are multiplication (5.NF.B.4). "Show your work" with no work space
HS packet (MLK Letter from Birmingham Jail, RH.9-10.2)	Built from the letter + 8 activity types + two answer keys (one quoting the source)	Strongest MagicSchool output: clean worksheet + varied MC + answer key. But the "passage" is a ~120-word paraphrase, so analysis of how the argument develops can't really happen
Considered defaults (one-prompt-to-usable)	✓ across all three prompts	✓ across all three prompts — MagicSchool's Worksheet Generator is genuinely one-shot
Differentiation (Spanish + 2nd-grade relevel)	Complete packets in both variants (passage, activities, key, images preserved)	Spanish is genuinely good (full content, native register). The 2nd-grade relevel simplifies by <i>deleting</i> content — cut to ~50 words
Pattern overall	Wins on standards depth, source fidelity, answer-key depth, and honest layout across all grade bands	Clears the completeness/hygiene bar a chatbot fails; falls short on depth, source fidelity, and room-to-write

The single most important finding: against a general-purpose chatbot, the story is "complete vs. incomplete." Against MagicSchool, both tools produce a complete, printable worksheet — so the comparison moves to a harder question: does the worksheet meet the standard the teacher named, at the depth that standard requires, with the source the teacher chose, in a form a student can actually write on? That is where Diffit and MagicSchool diverge.

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. And we credit MagicSchool everywhere it earns it: clean pagination, real passages, real answer keys, genuinely good translation.

What we tested

Three prompts spanning two grade bands, all CCSS-aligned. The two 5th-grade prompts (ELA + math) and the high-school MLK prompt (RH.9-10.2) are detailed in the scoring sections below.

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:** [NPS — Harriet Tubman Underground Railroad National Historical Park](#)

Prompt 2 — Math

Topic: Adding and subtracting fractions with unlike denominators (word problems) **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."* **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 encyclopedia entry on the letter](#)

The primary artifacts

Artifact	File	Pages
Diffit ELA — Harriet Tubman	diffit-ela-tubman.pdf	11
MagicSchool ELA — Harriet Tubman (Worksheet Generator)	magicschool-tubman-worksheet.pdf	2
Diffit Math — fractions	diffit-math-fractions.pdf	15
MagicSchool Math — fractions (Worksheet Generator)	magicschool-math-fractions.pdf	2
Diffit HS — MLK Letter	diffit-ela-mlk.pdf	17
MagicSchool HS — MLK Letter (Worksheet Generator)	magicschool-mlk-worksheet.pdf	3

Note: unlike a general-purpose chatbot, **MagicSchool's first response was usable in every case** — no follow-up prompting was required to get a student-facing worksheet. That is a real §3.3 (considered defaults) credit, and it holds across all three prompts.

ELA scoring

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

#	Criterion	Diffit	MagicSchool	Evidence
1.1	Factual correctness	✓ Meets	✓ Meets	Both artifacts' facts check out against the NPS source (born ~1822, escaped 1849, Underground Railroad, Civil War nurse/spy/scout, raid freed 700+, later women's-rights work). MagicSchool's answer key is internally correct
1.2	Source fidelity	✓ Meets	△ Partial	Diffit works from the NPS page in specific detail (Brodess, the two-pound-weight head injury, the Parson's Creek mariners, marshland navigation skills). MagicSchool names the NPS source but flattens it to a generic summary — none of those source specifics survive. Nothing is fabricated, but little is grounded in <i>this</i> source beyond what any encyclopedia would say
1.3	Mechanical correctness	✓ Meets	△ Partial	MagicSchool's opening sentence reads "worked on farms and farms' chores" — a garbled construction a teacher would fix before assigning. Otherwise clean. (Diffit's ELA packet is clean; the stray "\$" rendering artifact noted on Diffit's side appears only on the math error-analysis sheet, scored under Math §1.3)
1.4	Layout discipline	✓ Meets	✓ Meets	Both paginate cleanly and are appropriately dense for 5th grade. MagicSchool's 2-page worksheet is tidy — this is a bar a general-purpose chatbot often fails and MagicSchool clears
1.5	Answer-key consistency	✓ Meets	✓ Meets	Both include a complete answer key whose answers match the questions in the artifact. Credit MagicSchool here — the key is real, present, and correct
2.1	Standards alignment	✓ Meets	△ Partial	Both name RI.5.3. Diffit's activities operate on the included passage and target the standard's verb (relationships/interactions between individuals, events, ideas). MagicSchool labels the sheet "Note: RI.5.3" but the questions are recall and simple explanation — topic-correct, weakly aligned to the standard's actual demand
2.2	Cognitive depth (DOK)	✓ Meets	△ Partial	Diffit spans recall (cause/effect), application (interactions table), and evaluation (extended writing). MagicSchool is mostly recall (fill-in-the-blank + MC: "What year did Tubman escape?") with three "explain/describe" prompts that reach DOK 2 but no higher
2.3	Genuine variety	✓ Meets	△ Partial	Diffit's three activities each demand distinct cognitive work. MagicSchool varies the <i>format</i> (FIB, MC, open-ended) but the thinking is largely the same recall asked three ways
2.4	Question quality	N/A	△ Partial	Diffit's ELA activities are open-ended (MC criterion N/A). MagicSchool's 5 MC items have plausible distractors, but

#	Criterion	Diffit	MagicSchool	Evidence
				the answer key runs B, B, D, B, B — four of five land on B. Content-relevant distractors, predictable positions
2.5	Honesty about what the standard requires	✓ Meets	✗ Fails	RI.5.3 asks students to explain relationships and interactions <i>between</i> individuals, events, and ideas. Diffit builds an interactions activity that does exactly that. MagicSchool satisfies the labeled standard with recall and one-step explanation — the analytical move the standard names is routed around
3.1	Multi-activity coherence	✓ Meets	✓ Meets	Both read as a single coherent resource: every MagicSchool question references its own passage
3.2	Complete and ready to teach	✓ Meets	✓ Meets	Passage + questions + answer key, no placeholders, on both sides. MagicSchool clears this bar cleanly
3.3	Considered defaults	✓ Meets	✓ Meets	Both produce a usable worksheet from a single prompt with no follow-up. MagicSchool's Worksheet Generator is purpose-built for one-shot output — credit it
3.4	Honest layout for student work	✓ Meets	✗ Fails	Diffit gives table cells and ruled space sized to each task. MagicSchool tells students to "answer the following questions in complete sentences" and provides no writing space at all — the questions list straight into the answer key
3.5	Classroom-workflow format fidelity	✓ Meets	✓ Meets	Both export/print cleanly to a teacher-ready PDF
4.1	Right reading level	✓ Meets	✓ Meets	MagicSchool's passage reads at roughly 5th-grade level; Diffit's likewise
4.2	Engaging voice	✓ Meets	△ Partial	Diffit's passage uses concrete narrative detail. MagicSchool's is flat summary prose ("Her life shows how one person's actions can change many lives through bravery and organized effort") — serviceable but generic
4.3	Differentiation in the workflow	N/A	△ Partial	Tested in the Differentiation in practice section below (N/A here for Diffit). MagicSchool's Spanish translation is genuinely good; its 2nd-grade relevel, however, simplifies by deleting content — the variant is produced, but degraded
4.4	Pedagogically-useful visuals	N/A	✗ Fails	A biography reading does not require a conceptual diagram, so this is N/A for Diffit (which nonetheless includes three contextual images). MagicSchool's worksheet has none — a missed scaffolding opportunity for a 5th-grade reading lesson

#	Criterion	Diffit	MagicSchool	Evidence
4.5	Asset-based framing	✓ Meets	✓ Meets	Both handle the subject matter respectfully and accurately

ELA scorecard: Diffit 17 Meets, 0 Partial, 0 Fails (out of 17 applicable; §2.4, §4.3, §4.4 N/A for Diffit). MagicSchool 9 Meets, 8 Partial, 3 Fails (out of 20 applicable). MagicSchool clears completeness and hygiene (1.4, 1.5, 3.1–3.3, 3.5) but falls short on standards depth (2.1, 2.5), source fidelity (1.2), and room to write (3.4).

Math scoring

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

#	Criterion	Diffit	MagicSchool	Evidence
1.1	Factual correctness	✓ Meets	✓ Meets	All computations check out in both artifacts. MagicSchool's key is arithmetically correct throughout ($3/8 + 2/8 = 5/8$, $7/8 - 3/8 = 4/8$, $3 \times 3/4 = 9/4$, etc.)
1.2	Source fidelity	N/A	N/A	No source provided in either case
1.3	Mechanical correctness	△ Partial	✓ Meets	MagicSchool is clean. Diffit's error-analysis sheet carries a stray rendering artifact — " $\frac{1}{2} + \$14$ " where " $\frac{1}{2} + \frac{1}{4}$ " was meant — a minor, non-load-bearing slip, scored honestly
1.4	Layout discipline	✓ Meets	△ Partial	Diffit structures the packet into three activity types with visual scaffolding. MagicSchool is clean and paginated but minimal — no visuals, no scaffolding structure beyond a question list
1.5	Answer-key consistency	✓ Meets	✓ Meets	Both include a complete, correct answer key matching the questions. Credit MagicSchool
2.1	Standards alignment	✓ Meets	△ Partial	5.NF.A.2 is specifically about unlike denominators with visual models. MagicSchool titles the sheet "5.NF.A.2" but its own instructions say students will "add or subtract fractions with like denominators " — a grade-4 skill (4.NF.B.3). Topic-adjacent, mislabeled
2.2	Cognitive depth (DOK)	✓ Meets	△ Partial	Diffit includes a dedicated error-analysis activity (DOK 3). MagicSchool is procedural throughout — compute the sum/difference — with no reasoning-about-reasonableness or misconception work
2.3	Genuine variety	✓ Meets	△ Partial	Diffit's three activity types ask different thinking. MagicSchool varies format (FIB/MC/word problem) but every item is the same compute-the-fraction task
2.4	Question quality	N/A	△ Partial	MC quality is N/A for Diffit (its math is open-response — Fraction Freddy names misconceptions as educative content instead). MagicSchool's MC distractors are plausible fractions, but the key runs B, A, A, A, A — four of five on A
2.5	Honesty about what the standard requires	✓ Meets	× Fails	5.NF.A.2 explicitly calls for visual fraction models or equations and unlike denominators. MagicSchool provides no visual models, leans on like denominators, and — critically — two of its three open-ended problems are multiplication ($3 \times 3/4$; $5/8 \times 4$), which is 5.NF.B.4, a different standard
3.1	Multi-activity coherence	✓ Meets	✓ Meets	Both hang together around a consistent fractions context

#	Criterion	Diffit	MagicSchool	Evidence
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. Credit MagicSchool
3.4	Honest layout for student work	✓ Meets	✗ Fails	The worksheet instructs students to "show your work" and "write your answers in simplest form" — then provides no work space. There is nowhere to show the work it asks for
3.5	Classroom-workflow format fidelity	✓ Meets	✓ Meets	Both render and print cleanly — no broken LaTeX or layout collapse
4.1	Right reading level	✓ Meets	✓ Meets	Word problems read at 5th-grade level in both
4.2	Engaging voice	✓ Meets	✓ Meets	MagicSchool uses varied contexts (baking, ribbon, garden, lemonade); Diffit likewise. Both fine
4.3	Differentiation in the workflow	N/A	△ Partial	Tested in the Differentiation in practice section below (N/A here for Diffit). MagicSchool's relevance quality is the open question there
4.4	Pedagogically-useful visuals	✓ Meets	✗ Fails	5.NF.A.2 names visual fraction models in the standard text. Diffit includes circle and bar models on the first four pages. MagicSchool includes zero visuals — a teacher using it to fulfill 5.NF.A.2 has not delivered what the standard requires
4.5	Asset-based framing	✓ Meets	✓ Meets	Both name a diverse set of students (Diffit: Maya, Leo, Sarah; MagicSchool: Ava, Jamal, Maya, Lina)

Math scorecard: Diffit 16 Meets, 1 Partial, 0 Fails (out of 17 applicable; §1.2, §2.4, §4.3 N/A — the Partial is the stray-"\$" rendering artifact on the error-analysis sheet, §1.3). MagicSchool 10 Meets, 6 Partials, 3 Fails (out of 19 applicable; 1.2 N/A). The §2.5 and §4.4 Fails are the load-bearing ones: 5.NF.A.2 names visual fraction models and unlike denominators, and MagicSchool delivers neither — while routing two of three open-ended items into multiplication, a different standard.

Differentiation in practice

Starting from the same 5th-grade Tubman packet, both tools were asked to produce a **Spanish translation** and a **2nd-grade relevel**. In both tools this is a one-click / one-prompt workflow action. The question is what each path produces.

The headline finding

MagicSchool translates well and relevels by deleting. Its Spanish worksheet is genuinely strong — full content, native register, and it even adds detail (Tubman's birth name Araminta, her 1913 death, the suffrage movement). On translation, MagicSchool is comparable to Diffit; that row is a wash, and we say so.

The 2nd-grade relevel is where the tools diverge. 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 — Edward Brodess, the dates, the marshland skills, the Combahee River Raid, the 750 freed — in simpler sentences. MagicSchool's 2nd-grade relevel cuts the passage to roughly 50 words:

"Harriet Tubman was born into slavery. She escaped to freedom and helped many others find freedom too. Harriet guided people on the Underground Railroad at night. She was brave and worked to help people be free. Many remember her as a hero who cared for others."

The dates, the Civil War service, the raid — all gone. This is the specific failure mode §4.1 names: simpler by deleting, not by expressing the harder idea more simply.

Honest findings on each side

Against MagicSchool's Spanish translation: the worksheet is strong, but the English boilerplate line *"Review this closely for accuracy, especially in math, as AI may have limitations"* is appended **untranslated** at the bottom — on a Spanish reading worksheet that has nothing to do with math.

Against MagicSchool's 2nd-grade relevel: beyond the content cut, the teacher note mislabels the standard as a *"Verbatim Rhode Island Grade 5 Reading Standard (RI.5.3)."* RI is the Common Core code for "Reading: Informational," not Rhode Island.

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.

Against Diffit's Spanish translation: a mild terminology inconsistency — the passage uses "*pantanos*" (swamps) while one activity uses "*marismas*" (marshes) for the same feature.

What this means for §4.3: both tools score **Partial**. The differentiation workflow exists and runs cheaply in both. MagicSchool's translation is a genuine strength; its releve deletes content. Diffit's variants preserve the full packet, with the honest single-item finding noted above.

Pattern at secondary: Letter from Birmingham Jail

A reasonable objection: "*Fine for elementary — does the pattern hold at high school, where the source is already at the target reading level?*" This is MagicSchool's strongest output, so it is the fairest test.

Scoring

#	Criterion	Diffit	MagicSchool	Evidence
1.1	Factual correctness	✓ Meets	✓ Meets	Both artifacts are factually accurate. MagicSchool's paraphrase correctly represents the letter's key ideas (just vs. unjust laws, nonviolent direct action, "injustice anywhere is a threat to justice everywhere")
1.2	Source fidelity	✓ Meets	△ Partial	Diffit builds from the actual letter — the four-step campaign, the clergy's specific charges, King's own words — and its answer key quotes the source. MagicSchool's "passage" is a ~120-word paraphrase with a single embedded quotation; students analyze a summary, not King's text. Faithful to the ideas, but not grounded in the source itself
1.3	Mechanical correctness	✓ Meets	✓ Meets	Both clean
1.4	Layout discipline	✓ Meets	✓ Meets	Both paginate cleanly and are appropriately dense for high school
1.5	Answer-key consistency	✓ Meets	✓ Meets	Both include a complete answer key whose answers match the questions (MagicSchool's stray "review for accuracy, especially in math" boilerplate on this reading key is a craft slip, noted below — but the answers themselves are consistent)
2.1	Standards alignment	✓ Meets	△ Partial	Both name RH.9-10.2. The standard asks students to trace "how key events or ideas develop over the course of the text." Diffit's matching and synthesis activities probe development; MagicSchool's questions probe comprehension of a static paraphrase
2.2	Cognitive depth (DOK)	✓ Meets	△ Partial	MagicSchool reaches DOK 2–3 on its open-ended items ("explain in your own words why King believed nonviolent direct action was necessary") but the FIB and MC remain recall, and the analysis is bounded by the 120-word summary
2.3	Genuine variety	✓ Meets	△ Partial	Diffit fields eight distinct activity types. MagicSchool offers three (FIB, MC, open-ended) over the same comprehension target
2.4	Question quality	N/A	✓ Meets	MC quality is N/A for Diffit on this scoring pass. MagicSchool's MC is its strongest section: distractors are content-relevant and the answer positions vary (b, c, b, a, c)

#	Criterion	Diffit	MagicSchool	Evidence
2.5	Honesty about what the standard requires	✓ Meets	× Fails	Analyzing how King's argument <i>develops</i> requires the developing argument — the real letter. A 120-word paraphrase cannot show development, so the standard's central demand is routed around
3.1	Multi-activity coherence	✓ Meets	✓ Meets	Both reference common content throughout
3.2	Complete and ready to teach	✓ Meets	✓ Meets	Passage + questions + answer key on both sides, no placeholders
3.3	Considered defaults	✓ Meets	✓ Meets	Both produce a usable artifact from one prompt
3.4	Honest layout for student work	✓ Meets	× Fails	MagicSchool asks students to "answer each question in a complete sentence or two" and provides no writing space. Diffit's tables and prompts are sized to each task
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	✓ Meets	MagicSchool's paraphrase is readable and carries the famous quotation; Diffit uses narrative context plus primary-source quotation
4.3	Differentiation in the workflow	N/A	△ Partial	Tested in the Differentiation in practice section below (N/A here for Diffit)
4.4	Pedagogically-useful visuals	N/A	△ Partial	A text-only primary source does not require a conceptual diagram, so this is N/A for Diffit. MagicSchool likewise produced no images for the Birmingham campaign — a defensible-but-missed scaffolding opportunity
4.5	Asset-based framing	✓ Meets	✓ Meets	Both handle civil-rights content with appropriate seriousness

HS scorecard: Diffit 17 Meets, 0 Partials, 0 Fails (out of 17 applicable; §2.4, §4.3, §4.4 N/A for Diffit). MagicSchool 12 Meets, 6 Partials, 2 Fails (out of 20 applicable). This is MagicSchool's best showing — clean MC with varied answer positions, a real answer key, a readable paraphrase. The two Fails are the same two that recur across every prompt: the standard's depth is routed around (here, by paraphrasing a 7,000-word letter down to 120 words), and there is no room for students to write.

The honest read

At high school, **MagicSchool is closer to parity than at elementary, but Diffit still leads on source fidelity, answer-key depth, activity variety, and honest layout.** The pattern holds at narrower margins. A district admin who concludes "MagicSchool is a solid worksheet generator, and Diffit is a deeper standards-and-source tool" is reading the evidence correctly.

The five "look here" moments

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

1. "Show your work" — with nowhere to do it

Across all three MagicSchool worksheets, students are told to "answer in complete sentences" and "show your work" — and given no space. The questions list straight into the answer key. The math worksheet is the clearest proof: it asks students to "show your work and write your answer in simplest form," then prints the next question immediately below. This is a §3.4 (honest layout) Fail on every prompt.

2. The math worksheet labels a grade-4 skill as 5.NF.A.2 — and routes two problems into the wrong standard

MagicSchool's own instructions say students will "add or subtract fractions with **like denominators**." 5.NF.A.2 is specifically about **unlike** denominators, with visual models. Worse, two of the three open-ended problems are **multiplication** ($3 \times 3/4 = 9/4$; $5/8 \times 4 = 20/8$) — that's 5.NF.B.4, a different standard. A teacher using this sheet to cover 5.NF.A.2 has not covered it.

3. Recall labeled as RI.5.3

The Tubman worksheet is labeled "Note: RI.5.3" — a standard that asks students to explain relationships and interactions *between* individuals, events, and ideas. The questions are recall: "What year did Harriet Tubman escape?" Diffit's packet builds an interactions activity that asks how the Black mariners, the Maryland marshlands, and Governor Andrew each shaped Tubman's work — the analysis the standard names.

4. A 120-word paraphrase standing in for the Letter from Birmingham Jail

MagicSchool's HS "passage" is a ~120-word summary with one embedded quote. RH.9-10.2 asks students to analyze how King's argument *develops over the course of the text* — which requires the text. Diffit built

from the actual letter (the four-step campaign, the clergy's charges, King's own words) and its answer key quotes the source.

5. The craft tells: an English math disclaimer on a Spanish reading sheet, and "Rhode Island" RI.5.3

Two small, undeniable slips. The line *"Review this closely for accuracy, especially in math, as AI may have limitations"* appears on both ELA worksheets and, untranslated, on the Spanish one. And the 2nd-grade teacher note calls RI.5.3 a *"Verbatim Rhode Island Grade 5 Reading Standard"* — RI stands for "Reading: Informational." Neither is fatal; together they signal a tool generating *around* the standard rather than from a real understanding of it.

What we don't claim

MagicSchool is not a broken chatbot. It is purpose-built for teachers and it clears bars a general-purpose chatbot routinely fails: it paginates cleanly, produces a real reading passage, and includes a correct answer key on every worksheet. We credit all of that. The Diffit advantage here is not "complete vs. incomplete" — it is depth, source fidelity, and honest layout.

MagicSchool's Spanish translation is genuinely good. Full content, native register, even added biographical detail. The differentiation advantage Diffit holds is on the relevel, not the translation — and we say so.

MagicSchool did not fabricate facts. Every artifact is factually accurate. The divergence is about whether the worksheet meets the named standard at depth and works from the chosen source — not about accuracy.

Three prompts is not a complete evaluation. This covers ELA, math, and HS ELA with one prompt each. Other prompts will exercise other criteria. A district considering both tools should run their own comparison on their 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

Both tools are purpose-built for educators, so the difference is not chat-box-versus-structured-input (the story against a general chatbot). It is about how deep each tool goes once the worksheet is clean.

MagicSchool optimizes for a finished-looking worksheet. Diffit optimizes for a standard met with a real source. MagicSchool's Worksheet Generator reliably returns a tidy, complete, printable sheet in one shot — a genuine strength. But it treats the standard code as a label to print and the source as a citation to append, rather than as constraints the content must satisfy. The result clears every hygiene bar and then stops short of the standard's cognitive demand.

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), and the source is worked *from*, not merely cited. That is why the same clean-worksheet baseline diverges into "meets the standard" versus "labels the standard."

And Diffit leaves room for the work. The most teacher-visible difference is the simplest: when a worksheet says "show your work," Diffit gives space to show it. MagicSchool, on every prompt tested, does not.

Methodology — reproducing this comparison

Prompts

MagicSchool: each artifact was generated with MagicSchool's **Worksheet Generator** tool, entering the topic, grade, standard code, and (for ELA) the source URL into the tool's fields, then generating once with no follow-up. **Diffit:** topic, grade, standard (from the Standards picker), and source URL entered as structured inputs, generated once with optional follow-ups skipped.

Conditions

- Diffit and MagicSchool, May 2026.
- First responses captured before any follow-up prompting. (Both tools were one-shot usable; no follow-ups were needed.)
- 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 comparison, 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 Spanish and 2nd-grade variants; MagicSchool: magicsschool-tubman-worksheet, magicsschool-math-fractions, magicsschool-mlk-worksheet, magicsschool-tubman-spanish, magicsschool-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 MagicSchool.

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/>