

MISCONCEPTIONBENCH: Evaluating LLMs on Science Misconceptions and Student Response Distributions

Ying Zhang¹, Zhaoji Wang¹, Chenyu Zhang²
Yizhu Gao¹, Wenchao Ma³, Xiaoming Zhai^{1,*}

¹AI4STEM Education Center, University of Georgia, Athens, GA 30602, USA

²Harvard University, Cambridge, MA 02138, USA

³University of Minnesota, Minneapolis, MN 55455, USA

*Corresponding author: xiaoming.zhai@uga.edu

Abstract

Students' wrong answers in science assessments can reveal plausible misconceptions, yet standard LLM benchmarks usually evaluate only answer correctness. We introduce MISCONCEPTIONBENCH, a benchmark for evaluating whether LLMs can reason about misconception-linked distractors and empirical student response patterns. It contains 761 K–12 science items, 1,474 expert-annotated distractors, and response distributions from more than 150,000 students. We evaluate eight LLMs on three tasks: misconception matching, misconception generation, and student response distribution prediction. In matching, text-similarity baselines remain competitive in two- and three-candidate settings; clear LLM gains emerge only with four candidates, where GPT-4o and Llama-3.3-70B significantly outperform TF-IDF and MiniLM. In generation, the strongest models are tightly clustered, with Hit@1 of 38.5–39.6%, while retrieval-only baselines recover many targets without an LLM. Response distribution prediction remains difficult: no model exceeds F1=0.21 for identifying grade-band differences. Overall, misconception-centered reasoning requires evaluation beyond answer accuracy.

1 Introduction

Consider a student who answers “A powder dissolving in water to make lemonade” when asked, “Which of the following is an example of a chemical reaction?” (BSCS Science Learning, n.d.) The answer is incorrect, but *why*? The choice does not prove that this individual student holds a stable misconception: the student may have guessed, misread the question, or relied on partial knowledge. However, the selected distractor is still pedagogically informative because it is consistent with a common misconception that dissolving is a chemical

reaction. An experienced science teacher would therefore treat the wrong answer as evidence for a plausible misconception to check and address, rather than as a mere incorrect label. Different wrong answers can call for different feedback. A useful educational model should therefore do more than mark answers as correct or incorrect. It should identify the expert-linked misconception that a distractor is designed to target, explain that misconception, and predict how students at different grade levels distribute their answers across options.

Large language models (LLMs) are increasingly used to support such educational assessment, where they can reduce the expert labor needed to interpret student responses at scale (Liu et al., 2025). Still, one especially hard task is misconception inference: not just deciding whether a student answered correctly, but identifying *which* misunderstanding led them to a specific wrong option. Educational researchers have shown that wrong options (distractors) in well-designed multiple-choice items carry useful information about how students think (Treagust, 1988), and that knowing the specific misconception behind a wrong answer is the first step toward targeted feedback (Posner et al., 1982). In practice, building such annotations requires experts to write distractors, link each one to a likely misconception, and check these links against real student data. This process is costly and hard to scale. Recent Natural Language Processing (NLP) work has started to explore whether LLMs can reduce this cost: generating distractors that look like real student errors (Scarlatos et al., 2024; Fernandez et al., 2024), predicting how students choose among options (Liu et al., 2025), and retrieving misconceptions from a labeled list for domains like mathematics (King et al., 2024). A critical unaddressed question underlies this line of work: whether LLMs possess the reasoning capabilities that this work

demands, linking distractors to specific misconceptions, articulating the student thinking underlying each distractor, and accurately predicting how real students distribute their responses across choices, has yet to be tested, particularly within the domain of science education.

Existing benchmarks are ill-equipped to answer it. Widely used science question-answering benchmarks only check whether a model picks the correct answer (Section 2). They treat all wrong options as a single “incorrect” category and throw away the information that distractors are designed to carry. The Eedi benchmark (King et al., 2024) does evaluate misconception retrieval, but only in mathematics, and only as a single retrieval task. To our knowledge, no English-language science benchmark combines expert-checked misconception labels on each distractor with large-scale student response data, and no prior work evaluates LLMs across the full range of misconception tasks: from closed-set matching to open-ended generation to response distribution prediction. Figure 1 shows this difference: standard evaluation reduces an item to correct/incorrect, while misconception-centered evaluation treats each distractor as a diagnostic option linked to a specific misconception and to the share of students at each grade band who pick it.

To fill this gap, we introduce MISCONCEPTIONBENCH, a benchmark built from publicly available AAAS Project 2061 science items¹. We test misconception reasoning through three tasks: misconception matching (pick the right misconception from the misconception list), misconception generation (write the student thinking behind a distractor in open text), and response distribution prediction (estimate the share of students who pick each option). We evaluate 8 LLMs and find that model rankings differ across the three tasks: no single model performs best on all of them, and cross-task Spearman correlations are non-significant across 8 LLMs.

This dissociation suggests that misconception reasoning is not a unitary ability, and that answer accuracy alone is insufficient to characterize it. Code, prompts, model outputs, and analysis scripts are available at Github repository.²

¹<https://assess.bsccs.org/science/>

²<https://github.com/wkeaner/MisconceptionBench>

2 Related Work

Science QA Benchmarks Primarily Measure Correct-Answer Selection. Science and academic QA benchmarks such as ARC (Clark et al., 2018), SciQ (Welbl et al., 2017), MMLU (Hendrycks et al., 2021), and AGIEval (Zhong et al., 2024) primarily evaluate whether models select or generate correct answers. They neither examine expert-constructed distractors grounded in student misconceptions nor incorporate empirical data on how students distribute their choices across options. Yet in real classrooms, what teachers need to know is rarely whether the LLM itself can solve a problem; instead, it is whether the LLM can help them understand what and why their students make mistakes. MISCONCEPTIONBENCH addresses this gap by evaluating both distractor-level misconception labels and population-level response patterns.

Misconception-based assessment in Science Education. Decades of science education research establish that student errors are diagnostic: distractors in expert-designed tests target specific, persistent misconceptions (Sadler, 1998). The Force Concept Inventory (Hestenes et al., 1992) formalize this through field-tested item banks in which each distractor is linked to a named misconception drawn from an empirically-grounded taxonomy. Using these taxonomies as evaluation targets for LLMs (rather than only as instructional tools) is, to our knowledge, a novel bridge between NLP and science education. Critically, the misconception labels in AAAS Project 2061 are not authored ad hoc but published in peer-reviewed studies and empirically grounded in large-scale student response distributions (BSCS Science Learning, n.d.). This gives our MISCONCEPTIONBENCH a ground truth whose validity is independently established by the science education community, rather than labels we constructed by ourselves.

LLMs for distractor generation, student simulation, and misconception retrieval. Recent work deploys LLMs for distractor generation (Lee et al., 2025), individual student response simulation (He-Yueya et al., 2024), and misconception identification at the individual level (Liu et al., 2025). Yan et al. (2024) and Kasneci et al. (2023) survey broader opportunities and risks. The Eedi benchmark (King et al., 2024) evaluates misconception retrieval in mathematics, where many labels

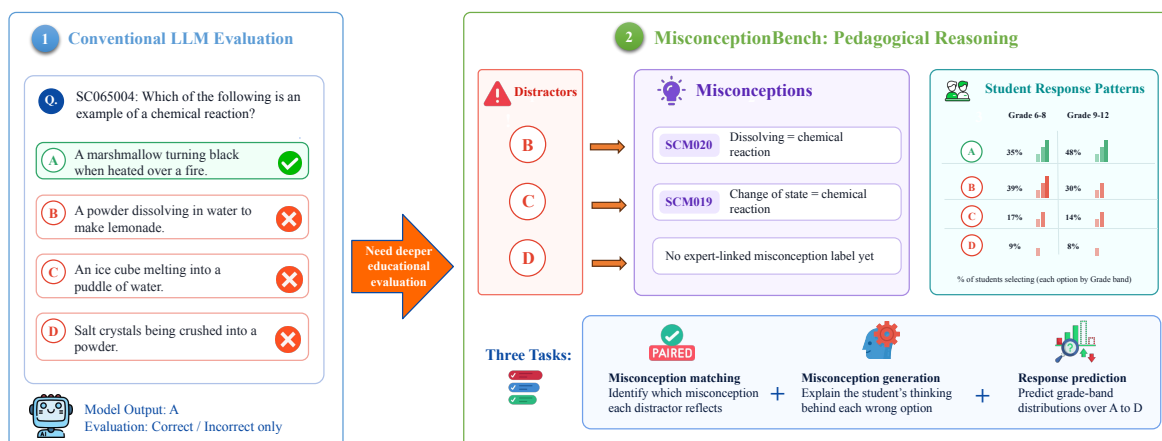


Figure 1: Overview of MISCONCEPTIONBENCH with an AAAS Project 2061 science item (BSCS Science Learning, n.d.). Conventional LLM evaluation scores only correct-answer selection, while we evaluate whether models can link wrong options to expert-defined misconceptions and predict student response patterns across grade bands.

correspond to procedural error patterns or incorrect solution strategies. MISCONCEPTIONBENCH extends misconception-centered evaluation to K–12 science, a domain in which misconceptions are often framed as alternative conceptions about underlying scientific phenomena. This distinction is important because science distractors are commonly designed to diagnose conceptual understanding rather than only answer accuracy or procedural correctness. In addition, MISCONCEPTIONBENCH links these misconception annotations to empirical grade-band response distributions, enabling evaluation of both misconception identification and population-level response prediction.

3 MisconceptionBench Dataset

Items derive from AAAS Project 2061 science misconception assessments (BSCS Science Learning, n.d.), developed through iterative validation by science education researchers, Grades 6–12 teachers, and student piloting. Each item includes a question stem, four answer options, expert-curated misconception code(s) for each distractor; a large subset of empirical student response distributions from over 150,000 U.S. classroom administrations. We processed 761 items spanning 16 science topics across Physical Science (371 items), Life Science (274 items), Earth Science (61 items), and Nature of Science (55 items). Of these, 681 items are eligible for misconception-identification tasks (containing 1,474 expert-annotated distractors in total), 750

items contain complete empirical response distributions for both grade bands 6–8 and 9–12, and the misconception taxonomy contains 694 unique expert-authored codes. Full per-topic counts and filtering details are in Appendix A.

4 Three Evaluation Tasks

We derive three tasks from the same item set: A) selecting a misconception from a candidate set, B) generating a misconception of a given distractor, and C) predicting student response distributions.

4.1 Task A: Misconception Matching

Given a science item and one incorrect answer option (distractor), the model selects the expert-linked misconception code for that distractor. The candidate set consists of the misconception codes annotated to the same item. Thus, the candidate-set size K is an item property rather than a manually controlled experimental variable in the main evaluation, ranging over $K \in \{2, 3, 4, 5\}$ in our data.

All candidate misconceptions are drawn from the same question item, so they are plausible labels for the same question. The task therefore requires fine-grained misconception discrimination among related misconceptions, rather than excluding obviously unrelated labels. Figure 2 illustrates an easy case with two candidate misconceptions ($K = 2$) and a harder case with four candidate misconceptions ($K = 4$), in which the candidates describe closely related accounts of the same phenomenon and half of the evaluated models select a code that

the source annotation assigns to a different distractor of the same item. Because a distractor may carry more than one expert-assigned code, a prediction is scored correct if it matches any code assigned to the queried distractor.

We report accuracy separately for each candidate-set size K . Because K is an item property, items evaluated at larger K may also be intrinsically harder. Thus, a drop from $K = 3$ to $K = 4$ could reflect two factors: the model sees more candidate codes, and the items with larger K may also be harder. To separate these factors, Section 6.1 includes a controlled analysis in which we keep the same item fixed and vary only the number of candidate codes shown to the model.

4.2 Task B: Misconception Generation

Given a science item and a distractor, the model generates a one-sentence description of the misconception that explains why a student might select that distractor. Because the ground-truth label is a misconception code from our 694-code taxonomy rather than a free-form sentence, we cannot directly compare the model’s sentence to the gold answer. Instead, we map the generated sentence to taxonomy codes through fixed retrieval: we encode the sentence and each of the 694 taxonomy entries with the same encoder all-MiniLM-L6-v2 (Reimers and Gurevych, 2019), and rank the taxonomy codes by cosine similarity to the generated sentence.

A prediction counts as Hit@ r if the gold misconception code appears among the top r retrieved taxonomy entries. We report Hit@ r at $r \in \{1, 3, 5, 10\}$. Here, r is a retrieval cutoff over one ranked list, not the candidate-set size K used in Task A. Hit@1 measures whether the first retrieved code is correct, Hit@10 measures whether the correct code appears within a broader top-10 list, and Hit@3 and Hit@5 provide intermediate points. We omit adjacent cutoffs such as Hit@2 and Hit@4 to keep the table compact, since they do not change the interpretation of the ranked-retrieval results.

This design has two consequences. First, the Hit@ r score reflects a *joint* property of the model, which writes the sentence, and the retriever, which maps the sentence to a code; a strong generator paired with a weak retriever may underperform. To assess the robustness of model rankings to this choice, we re-run all evaluations with two alternative encoders, MPNet and BGE-small, and report cross-retriever rank correlations in Appendix C.2. Second, because the benchmark distractors were

authored from the misconception taxonomy, the distractor text itself can be encoded and retrieved without involving any LLM; we report this no-LLM retrieval baseline in Section 6.2 to calibrate how much of the apparent performance comes from generation versus surface-text similarity between distractors and their target misconceptions.

4.3 Task C: Response Distribution Prediction

The first two tasks focus on *identifying* the misconception behind one wrong answer. Task C asks a complementary question: can the model predict how an entire student population distributes across the four answer options, and how that distribution changes with grade level? Science education research documents that many misconceptions are developmental: a misconception common among grade 6–8 students often becomes less prevalent in grades 9–12 as formal instruction accumulates (Treagust, 1988). A model that has internalized misconception-centered reasoning should not only know *which* distractor reflects which misconception, but also predict *how often* students at different grade levels choose each option.

For each item with empirical response distributions in both grade bands, we query the model twice: once specifying the grade 6–8 band, once specifying grade 9–12; and ask it to output a probability distribution over the four answer options $\{A, B, C, D\}$ that sums to 1. The ground truth is the empirical option-selection frequency observed in the AAAS administration data (mean $n = 2,769$ students per item-band).

We report four metrics that capture both distributional fit and the ability to predict grade-band shifts. Mean Jensen–Shannon Divergence (JSD) measures distributional fit between predicted and empirical option-choice distributions. Three shift-oriented metrics evaluate grade-band differences: differentiation rate measures how often a model predicts a meaningful grade-band shift; direction-match measures whether predicted shifts move in the empirically observed direction; and item-level F_1 measures whether the model identifies which items differentiate between grade bands. Formal definitions and thresholds are provided in Appendix B.4.

5 Experimental Setup

Models. We evaluate 8 LLMs across four vendor families: OpenAI, Meta, Alibaba, and DeepSeek. Four are reasoning-oriented variants: GPT-5, GPT-5.4, Qwen3-32B (Yang et al., 2025; Qwen Team,

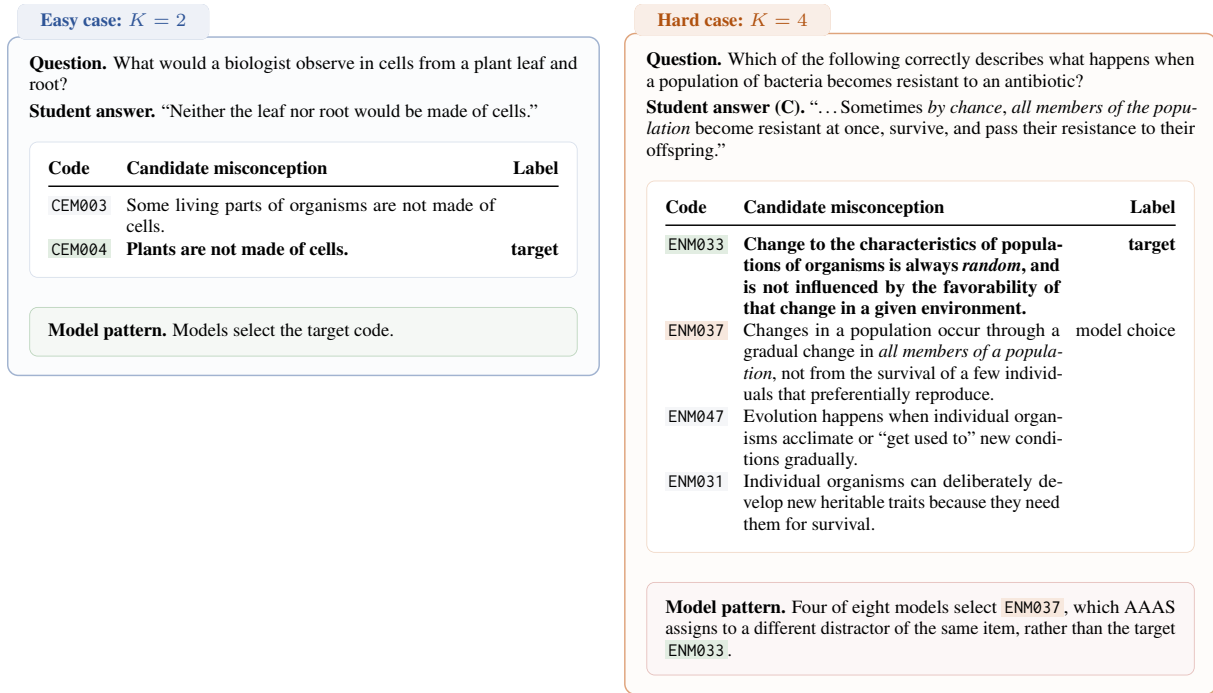


Figure 2: Worked examples for Task A. Candidates are the misconception codes annotated to the same item, so a candidate that is not the target for the queried distractor is typically the expert-assigned code for one of its sibling distractors rather than an implausible option. In the $K = 4$ case, all three non-target candidates are assigned to the item’s other two distractors, and half of the evaluated models select ENM037 instead of the target ENM033.

2025), and DeepSeek-R1-Distill-Llama-70B (Guo et al., 2025; DeepSeek-AI, 2025). Four are instruction-tuned variants: GPT-4o, Llama-3.1-8B (Meta, 2024a), Llama-3.2-3B (Meta, 2024b), and Llama-3.3-70B (Meta, 2024c). GPT models are accessed via the OpenAI API with model aliases gpt-4o-2024-11-20, gpt-5, and gpt-5.4.³ Open-weight models are served via vLLM on a single NVIDIA L40S 48 GB GPU, with W4A16 quantization for Llama-3.3-70B and DeepSeek-R1-Distill-Llama-70B and BF16 for the remaining open-weight models.

Decoding. We use low-temperature decoding where supported. GPT-4o and the four Llama models are run with temperature=0. For GPT-5 and GPT-5.4, the API rejects explicit temperature settings for these reasoning endpoints; we therefore omit the temperature field and the API applies its default sampling, and no seed parameter is exposed. Qwen3-32B and DeepSeek-R1 use their model-card-recommended sampling settings rather than greedy decoding: Qwen3-32B uses temperature=0.6, top_p=0.95, and top_k=20, following the Qwen3-32B model card’s

³For gpt-5 and gpt-5.4, we report the exact API aliases used; run timestamps span 2026-04-27 to 2026-05-04.

warning against greedy decoding (Qwen Team, 2025); DeepSeek-R1 uses temperature=0.6, within DeepSeek-R1’s recommended 0.5–0.7 range (Guo et al., 2025). GPT-5 and GPT-5.4 are run with reasoning_effort=low in the main evaluation due to compute cost; a high-effort rerun on Task B and repeated-run variance checks are reported in Appendix C.

Baselines. Each task includes non-LLM baselines that probe whether performance is recoverable without LLM inference: random, TF-IDF, MiniLM-cosine for Task A; BM25 and SBERT retrieval over the question and distractor text for Task B; uniform and topic-marginal predictors for Task C.

Statistical analyses. We report 95% confidence intervals via item-clustered bootstrap (1,000 resamples, seed 42). In each resample, we draw item IDs with replacement so all (item, distractor) records from the same item stay together; the metric is recomputed on the resampled set, and the 2.5th and 97.5th percentiles define the interval. We compute these intervals for the main metrics: Task A accuracy at each K and the $K = 3 \rightarrow K = 4$ drop, Task B Hit@ r , and Task C F_1 , mean JSD, and direction-match.

For Task A, we test whether each model’s ac-

Task B: Misconception generation and retrieval

Input. A student selects the distractor: “Neither the root nor the leaf would be made up of cells because plants are not made up of cells.” The model is asked to describe the misconception in one sentence.

GPT-5.4 output. “The student mistakenly believes that plants are not living things made of cells, and that only animals or other organisms have cellular structure.”

Interpretation. The model writes a free-text misconception description without seeing the taxonomy code or candidate list.

Fixed retrieval. The generated sentence and all 694 taxonomy entries are encoded with all-MiniLM-L6-v2; entries are ranked by cosine similarity.

Rank	Code	Retrieved misconception	Sim.
1	CEM004	Plants are not made of cells.	.73
2	CEM003	Some living parts are not made of cells.	.64
3	CEM047	Only certain kinds of cells are alive.	.61

Score. The ground-truth code CEM004 is ranked first, so this example receives Hit@1.

Figure 3: Compact example of Task B. A generated misconception sentence is mapped back to the expert taxonomy by fixed embedding retrieval.

curacy decreases monotonically as the candidate-set size K grows from 2 to 5, using a one-sided Cochran–Armitage trend test on per-record correctness, run separately for each model. We then compare models at $K = 4$ using McNemar’s exact test on per-(item, distractor) agreement, treating each (item, distractor) pair as one paired observation between the two models being compared, with Bonferroni correction over $\binom{8}{2} = 28$ pairs at $\alpha = 0.05$. We restrict pairwise testing to $K = 4$ because at $K = 2$ and $K = 3$ the six larger models cluster within roughly 5 percentage points (Table 1), leaving little resolvable signal, while $K = 4$ shows the largest cross-model spread (39 percentage points between the strongest and weakest model) and the model-ranking reversal discussed in Section 6.1.

We compute Spearman rank correlations across tasks over $N = 8$ models and report them descriptively only: at this sample size, $|\rho| > 0.71$ is required for $p < 0.05$, so non-significant correlations do not constitute evidence that the tasks are independent. Additional checks, including a 50-item per-model verbatim contamination probe and Task B retriever-robustness analyses (MiniLM, MPNet, BGE-small), are reported in Appendix C.

6 Results

We report results by task: misconception matching (Task A), misconception generation (Task B), and response distribution prediction (Task C). Tables 1, 2, and 3 report per-model numbers; we summarize the main patterns and close with a cross-task summary in Section 6.4.

6.1 Task A: Misconception Matching

$K = 4$ is where Task A begins to separate LLMs from text-similarity baselines. Across the eight LLMs, accuracy shows a significant negative trend

as the candidate pool grows from $K = 2$ to $K = 5$ (all models $p < 0.001$). At $K = 2$ and $K = 3$, the strongest LLMs and the text-similarity baselines are nearly tied: top LLM accuracy reaches 90–93% at $K = 2$ and 89–90% at $K = 3$, while TF-IDF and MiniLM reach 90.8%/90.3% at $K = 2$ and approximately 91% at $K = 3$. At $K = 4$, the pattern changes sharply: TF-IDF and MiniLM fall to 40.2% and 42.7%, while the best LLMs, GPT-4o and Llama-3.3-70B, remain at 65.9%. We therefore treat $K \leq 3$ as a regime in which text similarity already explains much of the performance, and $K = 4$ as the first setting where Task A clearly separates LLMs from retrieval-style baselines.

Only the two strongest LLMs clearly outperform the baselines at $K = 4$. We compare each LLM against each text-similarity baseline at $K = 4$ using McNemar’s exact test on per-(item, distractor) agreement, with Bonferroni correction over 16 LLM-baseline comparisons. Only GPT-4o and Llama-3.3-70B significantly outperform both TF-IDF and MiniLM after correction (all $p_{\text{adj}} < 0.02$). Other LLMs do not show reliable baseline-beating gains after correction.

Reasoning-oriented models do not show a measurable advantage at $K = 4$. At $K = 2$ and $K = 3$, GPT-5, GPT-5.4, and Qwen3-32B occupy the top three positions among LLMs. At $K = 4$, however, the two highest-accuracy models are instruction-tuned variants: GPT-4o and Llama-3.3-70B, both at 65.9%. All four reasoning-oriented models fall below them: Qwen3-32B at 57.3%, GPT-5 at 52.4%, GPT-5.4 at 50.0%, and DeepSeek-R1 at 49.4%. Pairwise McNemar tests at $K = 4$ show that the significant differences mainly separate the two smallest Llama models from stronger models; no pair among the six larger

Table 1: Task A accuracy (%) by candidate-set size.

Model	K=2	K=3	K=4	K=5	$\Delta_{3 \rightarrow 4}$
Random	50.0	33.3	25.0	20.0	8.3
TF-IDF	90.8	90.8	40.2	52.8	50.6
MiniLM	90.3	90.6	42.7	63.9	47.9
GPT-4o	90.5	88.0	65.9	50.0	22.2
GPT-5 ^R	90.2	90.1	52.4	44.4	37.7
GPT-5.4 ^R	91.0	90.1	50.0	38.9	40.1
Llama-70B	86.0	87.6	65.9	44.4	21.7
Llama-8B	64.1	77.2	42.7	36.1	34.5
Llama-3B	61.0	63.9	26.8	25.0	37.1
Qwen-32B ^R	93.2	89.4	57.3	47.2	32.1
DSeek-R1 ^R	89.7	88.4	49.4	27.8	39.0

Note. $\Delta_{3 \rightarrow 4}$ is the accuracy drop from $K = 3$ to $K = 4$, in percentage points. Blue cells mark the best LLM accuracy in each column; green marks the smallest LLM drop. ^R marks reasoning-oriented models. $K = 1$ is omitted because it is trivial.

models is significantly distinguishable after correction. Within this model set, reasoning-oriented models therefore do not show a measurable advantage over instruction-tuned models at $K = 4$. This comparison is descriptive because the groups also differ in vendor and decoding settings.

The $K = 3 \rightarrow K = 4$ drop reflects both larger candidate pools and item selection. Because K is an item property in the main evaluation, items at larger candidate-set sizes may also be intrinsically harder. To check whether the $K = 3 \rightarrow K = 4$ drop is only an item-selection artifact, we ran a within-item controlled analysis on the 12 items with at least five annotated misconceptions (Appendix C.1). Six of eight models still show statistically reliable $K = 3 \rightarrow K = 4$ drops when the item set is held fixed, ranging from 13.2 to 21.3 percentage points. These controlled drops are smaller than the corresponding main-evaluation drops, suggesting that the observed decrease reflects both candidate-set size and item-set composition. An approximate decomposition suggests that item-set composition accounts for a median of 42.9% of the main-evaluation drop among the six models with significant controlled drops. Thus, $K = 4$ remains harder than $K = 3$ even after holding items fixed, but roughly half of the observed drop is associated with which items enter the larger-pool condition.

6.2 Task B: Misconception Generation

Top models are tightly clustered at Hit@1. Table 2 reports Hit@ r for each model and three retrieval-only baselines. At Hit@1, the four strongest models are separated by only 1.1 percentage points: GPT-5.4 reaches 39.6%, followed by DeepSeek-R1 (39.4%), Qwen3-32B (38.7%),

and GPT-4o (38.5%). GPT-5 is close behind at 37.8%, while the three Llama models trail further (Llama-3.3-70B: 34.3%; Llama-3.1-8B: 31.3%; Llama-3.2-3B: 30.0%). Because these gaps are small and we do not run pairwise significance tests for Hit@1, we interpret the top group as a cluster rather than a reliable ordering among individual models. At Hit@10, all models recover the target code for most examples, with scores ranging from 80.5% to 90.2%. The spread at Hit@10 (9.7pp) is essentially the same as at Hit@1 (9.6pp), so larger retrieval cutoffs do not eliminate between-model differences.

Retrieval-only baselines recover many target misconceptions. The no-LLM baselines are strong: SBERT retrieval using only the distractor text reaches 37.8% Hit@1, essentially matching GPT-5 and exceeding all three Llama models. SBERT retrieval using question plus distractor text reaches 85.1% Hit@10; six of eight LLMs exceed this Hit@10 baseline, but the margins are modest, ranging from 1.9 to 5.1 percentage points, and the two smaller Llama models remain below it. Because the benchmark distractors are expert-linked to misconception descriptions within the same assessment design, lexical and embedding-based retrieval can recover a substantial fraction of correct misconceptions without any LLM inference. We therefore read Task B as a generate-then-retrieve pipeline: the model must produce a useful misconception description, and the fixed retriever must map that description back to the correct taxonomy code. The common retriever makes model comparisons consistent under the same evaluation pipeline, and retriever-robustness checks show that rankings are broadly stable across MiniLM, MPNet, and BGE-small encoders (Appendix C.2).

6.3 Task C: Response Distribution Prediction

No model reliably identifies grade-band shifts. Table 3 reports four Task C metrics: predicted differentiation rate (%Diff), binary F_1 for identifying empirically differentiating items, mean JSD to the empirical response distributions, and direction-match. We report %Diff, F_1 , mean JSD, and direction-match; definitions and thresholds are summarized in Table 3 and Appendix B.4.

Models show different frequency-direction tradeoffs. DeepSeek-R1 and Qwen3-32B predict grade-band differences most often (%Diff = 29.1% and 19.6%), well above the empirical rate of

Table 2: Task B misconception generation evaluated by fixed retrieval against the misconception taxonomy.

Model	H@1	H@3	H@5	H@10
<i>BM25 Q+D</i>	30.8	51.3	58.2	67.5
<i>SBERT D</i>	37.8	59.2	69.6	80.4
<i>SBERT Q+D</i>	36.4	63.7	73.9	85.1
GPT-4o	38.5	63.0	75.2	87.9
GPT-5 ^R	37.8	66.7	76.9	90.2
GPT-5.4 ^R	39.6	65.5	76.7	89.5
Llama-70B	34.3	61.1	73.2	87.0
Llama-8B	31.3	58.7	70.0	83.2
Llama-3B	30.0	56.2	68.5	80.5
Qwen-32B ^R	38.7	66.4	77.3	88.9
DSeek-R1 ^R	39.4	67.4	77.5	87.7

Note. H@K = Hit@K: the percentage of examples where the gold misconception code appears in the top K retrieved taxonomy entries. Q+D denotes question plus distractor text; D denotes distractor-only text. Blue cells mark the best LLM value in each column. ^R marks reasoning-oriented models.

11.3%. GPT-5.4 over-predicts shifts to a smaller degree (%Diff = 14.7%), while GPT-5 predicts shifts slightly less often than observed (%Diff = 8.8%). Direction-match shows the opposite pattern: GPT-5.4 and GPT-5 have higher directional agreement among the items they flag, while DeepSeek-R1 and Qwen-32B flag more items but with lower directional agreement. Because direction-match denominators vary widely, we report it descriptively rather than as a model-ranking metric.

LLMs improve over simple baselines, but the gains are bounded. The topic-marginal baseline achieves $F_1=0.056$, mean JSD = 0.062, and differentiates 3.1% of items; the uniform baseline has $F_1=0$ and mean JSD = 0.069. The best distribution-fit models, GPT-5.4 and GPT-4o, roughly halve mean JSD relative to the topic-marginal baseline. The best shift-detection models, DeepSeek-R1 and GPT-5, also exceed the topic-marginal F_1 . However, the absolute F_1 values remain low, showing that response-distribution prediction is still difficult. Llama-3.2-3B is a clear distribution-fit failure case, with mean JSD more than twice the uniform baseline. Full confidence intervals are reported in Appendix C.3.

6.4 Cross-Task Summary

The three tasks produce different model rankings. GPT-4o and Llama-3.3-70B are strongest on Task A at $K = 4$, GPT-5.4 has the highest numerical Task B Hit@1, and DeepSeek-R1 has the highest Task C F_1 . No model is best across all three tasks. Pairwise Spearman correlations between task rankings are positive but not statistically significant with $N = 8$ models; we therefore treat them as descriptive rather than as evidence that the tasks

Table 3: Task C response-distribution prediction results.

Model	%Diff	F_1	JSD↓	Dir. (n)
<i>Empirical</i>	11.3	—	—	—
<i>Uniform</i>	0.0	0.000	0.069	—
<i>Topic-marg.</i>	3.1	0.056	0.062	—
GPT-4o	4.1	0.052	0.033	<i>INSUFF. (1)</i>
GPT-5 ^R	8.8	0.199	0.040	82.6 (23)
GPT-5.4 ^R	14.7	0.185	0.031	90.9 (22)
Llama-70B	5.5	0.095	0.038	90.0 (10)
Llama-8B	4.6	0.034	0.044	81.8 (11)
Llama-3B	6.9	0.123	<i>0.141</i>	47.7 (44)
Qwen-32B ^R	19.6	0.181	0.038	64.6 (65)
DSeek-R1 ^R	29.1	0.205	0.049	54.8 (166)

Note. %Diff is the predicted grade-band differentiation rate, defined as either a top-choice change or pairwise JSD > 0.05 between predicted distributions for grade bands 6–8 and 9–12; empirical %Diff is 11.3%. F_1 is binary classification F_1 for empirically differentiating items, over 750 evaluated items. JSD is mean Jensen–Shannon divergence in bits to empirical response distributions; lower is better. Dir. (n) is direction-match among items where the model predicts a JSD-based shift and the empirical dominant-option shift exceeds 0.02; n is the denominator. INSUFF. indicates fewer than 15 eligible items. Bootstrap 95% CIs for F_1 and JSD are reported in Appendix C.3. Blue marks best LLM values; green marks the lowest JSD; yellow marks caveated values. ^R marks reasoning-oriented models. Parse-failure exclusions: Llama-8B (5 items), Llama-3B (40 items), Qwen-32B (1 item).

are independent. These results suggest that misconception matching, misconception generation, and response-distribution prediction constitute distinct evaluation targets, each requiring its own benchmark.

7 Conclusion

We introduced MISCONCEPTIONBENCH, a diagnostic benchmark that links 761 AAAS Project 2061 science assessment items to 1,474 expert-annotated misconception labels and empirical response distributions from over 150,000 students. Across three tasks—misconception matching, misconception generation, and response distribution prediction—our results show that misconception-centered evaluation captures capabilities not reflected by correctness-oriented science QA alone. No single model performs best across all tasks, and cross-task Spearman correlations are non-significant, suggesting that misconception matching, explanation, and response-distribution prediction should be evaluated as distinct abilities.

Our findings suggest three directions for future work. First, MISCONCEPTIONBENCH covers only English-language K–12 science; extending to other languages, domains, and grade levels would test the generality of the observed patterns. Second, the competitive retrieval baseline in Task B motivates hybrid architectures that combine LLM generation with structured misconception retrieval, rather than treating generation and retrieval as alternatives. Third, the poor Task C performance across

all models points to a fundamental gap: LLMs may lack the internalized student models needed to predict population-level response distributions, and future work might explore whether fine-tuning on large-scale student response data can close this gap.

More broadly, MISCONCEPTIONBENCH operationalizes a principle that has long been central to science education research — that wrong answers are diagnostic, not merely incorrect — and translates it into a concrete NLP evaluation framework. We hope this benchmark encourages the NLP community to move beyond answer correctness as the primary lens for educational AI evaluation, and toward richer, student-model-aware assessments that better reflect what teachers actually need from AI-assisted tools in the classroom.

Limitations

Single-run measurements. Several evaluated models use stochastic or API-default sampling, and the main evaluation therefore relies primarily on single-run measurements. Although additional robustness analyses are provided in Appendix C, comprehensive repeated-run evaluation across all models and tasks remains beyond the scope of this work.

Reasoning effort. GPT-5 and GPT-5.4 use low reasoning effort in the main evaluation because of computational cost. While Appendix C includes additional analyses under alternative reasoning settings, we do not exhaustively evaluate performance across different reasoning budgets.

Task C metric sensitivity. Our differentiation metric combines changes in the most frequently selected answer with a fixed Jensen–Shannon divergence threshold. The usefulness of this threshold may depend on the magnitude of grade-band differences in a particular dataset. Results based on this metric should therefore be interpreted with caution when applied to populations or datasets with different response patterns.

Cross-task statistical power. The cross-task analysis is based on a limited set of evaluated models. As a result, rank correlations across tasks have limited statistical power and should be interpreted descriptively rather than as evidence that the evaluated capabilities are statistically independent.

Snapshot validity. All models are evaluated using specific API snapshots or local checkpoints.

Because LLMs evolve rapidly, the reported findings may not reflect later versions or the current capabilities of the corresponding model families.

Model coverage. Our evaluation covers a limited set of model families and does not include all major proprietary and open-weight models. In particular, some models were excluded because of computational and budget constraints. Evaluating a broader range of model families would strengthen the generalizability of cross-model comparisons.

Public-source contamination. AAAS Project 2061 materials are publicly accessible, so exposure during model pretraining cannot be fully ruled out. Our contamination analysis focuses on near-verbatim memorization and cannot exclude paraphrased exposure, memorization of related misconception descriptions, or exposure to aggregate response statistics.

Ethics Statement

All evaluation items are derived from publicly available AAAS Project 2061 educational materials ([BSCS Science Learning, n.d.](#)); No individual-level student data are collected, retained, or released in this work. The empirical response distributions used in MISCONCEPTIONBENCH are population-level aggregate statistics.

MISCONCEPTIONBENCH is intended to evaluate LLM capabilities for research purposes and to improve the quality of AI-assisted science education tools, not to replace educator judgment. We flag that reasoning models’ higher grade-band differentiation rates represent a potential calibration risk: if these models are deployed for student response prediction without human oversight or empirical recalibration, they may produce overconfident grade-level distinctions that do not reflect actual classroom variation. Practitioners should treat MISCONCEPTIONBENCH results as research findings to inform model selection, not as deployment certification.

Acknowledgments

This work was supported by the Institute of Education Sciences (GENIUS, #R305C240010) and the National Science Foundation (TALENT, #2507128). The opinions expressed are those of the authors and do not represent views of the funding agencies.

References

- BSCS Science Learning. n.d. AAAS Science Assessment: Topics. <https://assess.bsos.org/science/topics>. Accessed May 05, 2026.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#). *arXiv preprint arXiv:1803.05457*.
- DeepSeek-AI. 2025. DeepSeek-R1-Distill-Llama-70B model card. <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B>. Hugging Face model card; accessed May 05, 2026.
- Nigel Fernandez, Alexander Scarlatos, Wanyong Feng, Simon Woodhead, and Andrew Lan. 2024. [DiVERT: Distractor generation with variational errors represented as text for math multiple-choice questions](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9063–9081, Miami, Florida, USA. Association for Computational Linguistics.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shitong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Joy He-Yueya, Wanjin Anya Ma, Kanishk Gandhi, Benjamin W Domingue, Emma Brunskill, and Noah D Goodman. 2024. [Psychometric alignment: Capturing human knowledge distributions via language models](#). *arXiv preprint arXiv:2407.15645*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *9th International Conference on Learning Representations (ICLR 2021)*, Virtual Event, Austria. OpenReview.net.
- David Hestenes, Malcolm Wells, and Gregg Swackhamer. 1992. [Force concept inventory](#). *The Physics Teacher*, 30(3):141–158.
- Enkelejda Kasneci, Kathrin Sessler, Stefan Küchermann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, and 4 others. 2023. [ChatGPT for good? on opportunities and challenges of large language models for education](#). *Learning and Individual Differences*, 103:102274.
- Jules King, L. Burleigh, Simon Woodhead, Panagiotis Kon, Perpetual Baffour, Scott Crossley, Walter Reade, and Maggie Demkin. 2024. [Eedi - Mining Misconceptions in Mathematics](#). <https://www.kaggle.com/competitions/eedi-mining-misconceptions-in-mathematics>. Kaggle competition; Accessed May 25, 2026.
- Yooseop Lee, Suin Kim, and Yohan Jo. 2025. [Generating plausible distractors for multiple-choice questions via student choice prediction](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23669–23692, Vienna, Austria. Association for Computational Linguistics.
- Naiming Liu, Shashank Sonkar, and Richard G. Baraniuk. 2025. [Do LLMs make mistakes like students? exploring natural alignments between language models and human error patterns](#). In *Artificial Intelligence in Education: 26th International Conference, AIED 2025, Palermo, Italy, July 22–26, 2025, Proceedings, Part IV*, volume 15880 of *Lecture Notes in Computer Science*, pages 364–377. Springer Nature Switzerland.
- Meta. 2024a. Llama 3.1 model card. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/. Accessed May 05, 2026.
- Meta. 2024b. Llama 3.2 model card. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/. Accessed May 05, 2026.
- Meta. 2024c. Llama 3.3 model card. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/. Accessed May 05, 2026.
- George J. Posner, Kenneth A. Strike, Peter W. Hewson, and William A. Gertzog. 1982. [Accommodation of a scientific conception: Toward a theory of conceptual change](#). *Science Education*, 66(2):211–227.
- Qwen Team. 2025. [Qwen3-32B model card](#). Hugging Face model card; accessed May 05, 2026.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Philip M. Sadler. 1998. [Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments](#). *Journal of Research in Science Teaching*, 35(3):265–296.
- Alexander Scarlatos, Wanyong Feng, Digory Smith, Simon Woodhead, and Andrew Lan. 2024. [Improving automated distractor generation for math multiple-choice questions with overgenerate-and-rank](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 222–231, Mexico City, Mexico. Association for Computational Linguistics.
- David F. Treagust. 1988. [Development and use of diagnostic tests to evaluate students' misconceptions in](#)

science. *International Journal of Science Education*, 10(2):159–169.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. [Crowdsourcing multiple choice science questions](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. [Practical and ethical challenges of large language models in education: A systematic scoping review](#). *British Journal of Educational Technology*, 55(1):90–112.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. [AGIEval: A human-centric benchmark for evaluating foundation models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

A Data Details

A.1 Item Sourcing and Processing

Items were exported from the AAAS Project 2061 Science Assessment database. Of 780 items processed, 761 were successfully built into the canonical format; 19 items failed due to non-standard table markup or missing required fields. Items requiring question-critical visuals are excluded from tasks that require text-only evaluation; a small number of items with non-standard option sets (e.g., true/false, more than four choices) are also excluded.

A.2 Eligibility Criteria

Task A eligibility. An item is Task A eligible if: (1) it has exactly four answer options (A–D), (2) at least one distractor has at least one annotated misconception code, (3) the item is text-only (no image dependency in question or options), and (4) at least 2 candidate misconceptions are available for pool construction. 681 items meet these criteria.

Task B eligibility. An item–distractor pair is Task B eligible if the item is text-only, has exactly four answer options, and the distractor has an expert-linked misconception code. We evaluate Task B on 1,474 unique item–distractor pairs. Because Task

Table 4: Item counts by topic in MISCONCEPTION-BENCH. Total = 761 items across 16 topics.

Topic	Items
Atoms, Molecules, and States of Matter	44
Cells	48
Control of Variables	21
Energy: Forms, Transformation, Transfer, and Conservation	102
Evolution and Natural Selection	42
Force and Motion	26
Human Body Systems	38
Interdependence in Ecosystems	45
Matter and Energy in Living Systems	21
Models	34
Plate Tectonics	27
Reproduction, Genes, and Heredity	80
Substances, Chemical Reactions, and Conservation of Matter	35
Weather and Climate I: Basic Elements	60
Weather and Climate II: Seasonal Differences	104
Weathering, Erosion, and Deposition	34
Total	761

B is evaluated by retrieval against the misconception taxonomy, each eligible pair must have a gold taxonomy code.

Task C eligibility. An item is Task C eligible if it has empirical student response distributions for both grade bands 6–8 and 9–12 from real classroom administrations. 753 items meet this criterion; 750 were evaluated after final processing.

A.3 Per-Topic Item Counts

Table 4 shows the number of canonical items per topic and domain.

A.4 Sample Item Schema

To illustrate the canonical item structure, we describe the schema without reproducing AAAS item text. Each canonical item in MISCONCEPTION-BENCH contains:

- `item_id`: unique identifier (e.g., AM059002)
- `topic`: one of 16 science topics
- `question_text`: item stem (not redistributed)
- `options`: dictionary with keys A, B, C, D
- `correct_answer`: letter (A–D)
- `distractor_misconception_map`: mapping from each distractor letter to a list of misconception code(s) from the 694-code expert taxonomy

- `empirical_distribution`: dictionary mapping grade band (“6–8” or “9–12”) to an option-level probability distribution over A, B, C, D (where available)

To reconstruct full item text, use `item_id` to query the AAAS Project 2061 platform at <https://assess.bscs.org/science/>.

Illustrative synthetic example. A hypothetical item in the Atoms, Molecules, and States of Matter topic might ask: “Which of the following describes how the volume of a gas changes when its temperature increases at constant pressure?” with four options including the correct answer (volume increases proportionally) and three distractors targeting common misconceptions about gas behavior. Each distractor would be annotated with a misconception code (e.g., a misconception that gas molecules grow in size when heated, rather than spreading apart).

B Exact Prompts

All prompts are reproduced verbatim from `src/experiments/prompts.py`. Each task uses a two-part (system, user) structure sent via the chat completions API. Placeholders in angle brackets are filled with item-specific content at runtime.

B.1 Task A: Closed-Set Discrimination

System prompt:

You are an expert in science education research.
 ↳ A student answered a multiple-choice science question incorrectly.
 ↳ Your task is to identify which misconception from a provided list best explains why the student chose that wrong answer. Select from the candidate misconceptions only.

User prompt:

Question: `<item["question_text"]>`

Answer options:

- A. `<item["options"]["A"]>` [or [CORRECT] if ↳ correct]
- B. `<item["options"]["B"]>` [or [CORRECT] if ↳ correct]
- C. `<item["options"]["C"]>` [or [CORRECT] if ↳ correct]
- D. `<item["options"]["D"]>` [or [CORRECT] if ↳ correct]

The student chose option `<distractor_letter>`:
 ↳ `"<item["options"]>[<distractor>]"`

Candidate misconceptions:

1. `<code_1>`: `<description_1>`
2. `<code_2>`: `<description_2>`
- ...

Output ONLY the misconception code (e.g., ↳ `AMM127`). No reasoning, no extra text.

B.2 Task B: Misconception Generation

System prompt:

You are an expert in science education research.
 ↳ A student answered a multiple-choice science question incorrectly.
 ↳ Your task is to articulate, in one sentence, the underlying misconception ↳ that would lead a student to choose that wrong answer.

User prompt:

Question: `<item["question_text"]>`

Answer options:

- A. `<item["options"]["A"]>` [or [CORRECT] if ↳ correct]
- B. `<item["options"]["B"]>` [or [CORRECT] if ↳ correct]
- C. `<item["options"]["C"]>` [or [CORRECT] if ↳ correct]
- D. `<item["options"]["D"]>` [or [CORRECT] if ↳ correct]

The student chose option `<distractor_letter>`:
 ↳ `"<item["options"]>[<distractor>]"`

Output ONE sentence describing the misconception ↳ that would lead a student to choose this answer. No preamble, no ↳ JSON.

B.3 Task C: Grade-Band Distribution Prediction

System prompt:

You are an expert in science education research.
 ↳ Your task is to predict how a population of students would respond to a ↳ multiple-choice science question, based on known patterns in student ↳ thinking. Produce a probability distribution over the four answer ↳ choices.

User prompt:

Question: `<item["question_text"]>`

Answer options:

- A. `<item["options"]["A"]>` [or [CORRECT] if ↳ correct]
- B. `<item["options"]["B"]>` [or [CORRECT] if ↳ correct]
- C. `<item["options"]["C"]>` [or [CORRECT] if ↳ correct]
- D. `<item["options"]["D"]>` [or [CORRECT] if ↳ correct]

Predict the probability that a typical student in ↳ grade `<grade_band>` in the United States would choose each option.

Output ONLY a JSON object with the form `{"A": ↳ float, "B": float,`

"C": float, "D": float} where probabilities sum $\hookrightarrow$ to 1.0. No markdown fences, no extra text.

B.4 Task C Metric Definitions

Mean Jensen–Shannon divergence (JSD). Average JSD between the predicted and empirical distribution, computed separately for each grade band and then averaged. Measures overall distributional fit; lower is better.

Differentiation rate (%Diff). The proportion of items for which the model predicts a meaningful shift between grade bands, defined as either a change in the top-choice option or a pairwise JSD between the two predicted distributions exceeding 0.05 bits. Compared against the empirical differentiation rate of 11.3%, this measures whether the model produces grade-band differences at the right frequency.

Direction-match. Direction-match is computed on a narrower denominator than %Diff and item-level F_1 . It is evaluated only for items where the model predicts a JSD-based grade-band shift, defined as $\text{JSD}(\hat{p}_{i,6-8}, \hat{p}_{i,9-12}) > 0.05$, and the empirical dominant-option shift exceeds 0.02. For each eligible item, we compare the sign of the predicted change with the sign of the empirical change for the empirically dominant shifting option. A direction-match occurs when the model predicts the same direction of change as the empirical data, such as both predicting that a distractor becomes less popular in grades 9–12. Because the denominator depends on how often each model predicts a JSD-based shift, we report direction-match descriptively rather than as a primary ranking metric.

Item-level F_1 . Treating each item as a binary classification of “empirically differentiates between grade bands or not,” we compute precision, recall, and F_1 over the model’s predicted-differentiation labels. Measures whether the model correctly identifies *which specific items* show developmental shifts.

C Additional Results

C.1 Task A Controlled Candidate-Set Analysis

The main Task A evaluation is observational: an item appears at candidate-set size K only if it has at least K annotated misconception codes. Therefore, lower accuracy at larger K can reflect two factors at once: the model must choose among more candidate misconceptions, and the items that appear at larger K may also be harder. To hold the item

set fixed, we ran a within-item controlled experiment on the 12 items with at least five annotated misconceptions.

For each eligible item–distractor target and each pool size $K \in \{2, 3, 4, 5\}$, we drew five random candidate sets containing the target code plus $K - 1$ distractor codes from the same item’s misconception pool. All eight models were evaluated on the same seed-fixed candidate sets. We then computed the within-item $K = 3 \rightarrow K = 4$ accuracy drop and compared it with the corresponding drop in the main evaluation.

Table 5 reports the results. Six of eight models show statistically reliable within-item drops, with 95% paired bootstrap confidence intervals excluding zero: GPT-4o, GPT-5, GPT-5.4, Llama-3.3-70B, DeepSeek-R1, and Qwen3-32B. The two smaller Llama models do not show reliable controlled drops: Llama-3.1-8B shows a smaller non-significant decrease, while Llama-3.2-3B shows a non-significant positive change, consistent with a floor effect in the controlled pool.

For the six models with reliable within-item drops, the controlled drops are smaller than the main-evaluation drops. We compute an approximate item-selection share as

$$1 - \frac{|\Delta_{\text{ctrl}}|}{|\Delta_{\text{main}}|},$$

where Δ_{main} is the observed $K = 3 \rightarrow K = 4$ drop in the main evaluation and Δ_{ctrl} is the corresponding within-item drop. The median item-selection share is 42.9%, suggesting that roughly half of the main-evaluation drop is associated with which items appear at $K = 4$, while the remaining drop reflects the harder candidate-set discrimination itself.

C.2 Task B Retriever Robustness

To assess whether Task B Hit@ r depends on the choice of retriever, we re-scored all 8 models’ generated sentences using two alternative retrievers in addition to the primary all-MiniLM-L6-v2 (384-dim) used in Section 6.2: all-mpnet-base-v2 (MPNet, 768-dim) and bge-small-en-v1.5 (BGE-small, 384-dim). All three retrievers score the same generated sentences against the same 694-code taxonomy on the same 1,474 unique (item, distractor) pairs.

Cross-retriever Spearman rank correlations on model rankings at Hit@1 are high across all retriever pairs: $\rho = 0.970$ ($p < 0.001$, MiniLM

Table 5: Within-item control for the $K = 3 \rightarrow K = 4$ accuracy drop.

Model	Main			Ctrl.			Att.
	K3	K4	Drop	K3	K4	Drop	%
GPT-4o	88.0	65.9	22.2	64.9	50.5	14.5	34.8
GPT-5 ^R	90.1	52.4	37.7	61.2	44.0	17.2	54.2
GPT-5.4 ^R	90.1	50.0	40.1	64.2	42.9	21.3	46.8
Llama-70B	87.6	65.9	21.7	60.6	47.4	13.2	39.0
Llama-8B	77.2	42.7	34.5	51.1	44.9	6.2 [†]	82.2
Llama-3B	63.9	26.8	37.1	36.4	40.9	−4.5 [‡]	112.2
DSeek-R1 ^R	88.4	49.4	39.0	57.5	38.8	18.8	51.9
Qwen-32B ^R	89.4	57.3	32.1	67.4	47.1	20.3	36.7

Note. Main is the full evaluation; Ctrl. is the within-item controlled experiment. Drop is $K = 3$ accuracy minus $K = 4$ accuracy, in percentage points. Att. = $1 - \text{Ctrl.Drop}/\text{Main.Drop}$, reported as an approximate item-selection share. Blue marks highest accuracy within each K column; green marks the smallest reliable drop. Yellow marks caveated values. ^R marks reasoning-oriented models. [†]95% CI includes 0. [‡]Controlled change is positive, consistent with a floor effect.

Table 6: Task B Hit@1 across three retrieval models (all using the same 1,474 dedup (item, distractor) pairs). R = reasoning-oriented model. Spearman rank correlations (Hit@1): MiniLM vs. MPNet $\rho = 0.9701$ ($p = 0.0001$); MiniLM vs. BGE-small $\rho = 0.9286$ ($p = 0.0009$); MPNet vs. BGE-small $\rho = 0.8982$ ($p = 0.0024$).

Model	MiniLM	MPNet	BGE-small
GPT-4o	0.3853	0.3602	0.2822
GPT-5 ^R	0.3779	0.3589	0.3338
GPT-5.4 ^R	0.3962	0.3765	0.344
Llama-3.3-70B	0.3433	0.3541	0.272
Llama-3.1-8B	0.3128	0.2999	0.2551
Llama-3.2-3B	0.2999	0.3114	0.2388
Qwen3-32B ^R	0.3874	0.3731	0.2951
DeepSeek-R1 ^R	0.3948	0.3765	0.3433

vs. MPNet), $\rho = 0.929$ ($p = 0.001$, MiniLM vs. BGE-small), and $\rho = 0.898$ ($p = 0.002$, MPNet vs. BGE-small). These results suggest that the main Task B ranking patterns are not specific to the MiniLM retriever, although the absolute Hit@ K values should still be interpreted as properties of the generate-then-retrieve pipeline rather than pure generation quality.

C.3 Task C Additional Results

F₁ and JSD confidence intervals. Table 7 reports item-clustered bootstrap 95% confidence intervals (1,000 resamples, seed 42) for Task C F₁ and mean JSD. These intervals support two main readings from Section 6.3: first, the best distribution-fit models have JSD values well below the uniform and topic-marginal baselines; second, GPT-4o and Llama-3.1-8B have F₁ intervals that include zero, indicating unstable shift-detection performance under resampling.

Table 7: Task C F₁ and mean JSD with 95% confidence intervals.

Model	F ₁ [95% CI]	JSD [95% CI]
GPT-4o	0.052 [0.000, 0.113]	0.033 [0.032, 0.035]
GPT-5	0.199 [0.114, 0.286]	0.040 [0.038, 0.042]
GPT-5.4	0.185 [0.118, 0.259]	0.031 [0.030, 0.033]
Llama-3.3-70B	0.095 [0.029, 0.165]	0.038 [0.036, 0.040]
Llama-3.1-8B	0.034 [0.000, 0.092]	0.044 [0.042, 0.046]
Llama-3.2-3B	0.123 [0.054, 0.200]	0.141 [0.134, 0.147]
Qwen3-32B	0.181 [0.109, 0.250]	0.038 [0.036, 0.039]
DeepSeek-R1-70B	0.205 [0.143, 0.263]	0.049 [0.047, 0.052]

Note. CIs are item-clustered bootstrap intervals with 1,000 resamples. JSD is Jensen–Shannon divergence in bits; lower is better.

Table 8: Contamination probe results.

Model	N	Lev>0.7	Max Lev	Mean SBERT	Flag
GPT-4o	50	0%	0.125	0.059	GREEN
GPT-5	50	0%	0.125	0.051	GREEN
GPT-5.4	50	0%	0.221	0.060	GREEN
Llama-70B	50	0%	0.280	0.326	GREEN
Llama-8B	50	0%	0.378	0.213	GREEN
Llama-3B	50	0%	0.486	0.130	GREEN
Qwen-32B	50	0%	0.123	0.479	GREEN
DSeek-R1	50	0%	0.231	0.017	GREEN

Note. Lev>0.7 is the percentage of probed items whose completion exceeds a near-verbatim Levenshtein-ratio threshold against the held-out question text. Mean SBERT is cosine similarity between the model response and the full question text. GREEN indicates no item exceeded Lev>0.7 or Lev>0.9.

Direction-match denominators. Direction-match is reported only descriptively because its denominator depends on how often each model predicts a JSD-based shift. The denominator ranges from $n = 1$ for GPT-4o to $n = 166$ for DeepSeek-R1. As a result, models with high point estimates but small denominators, such as Llama-3.3-70B (90.0% on $n = 10$) and Llama-3.1-8B (81.8% on $n = 11$), should not be interpreted as outperforming models with larger denominators. In Section 6.3, we therefore use direction-match only to describe qualitative failure modes, not to rank models.

C.4 Contamination Probe

Because AAAS Project 2061 items are publicly accessible, we audit possible verbatim item-text memorization. For each model, we sample the same 50 task-A-eligible items uniformly at random (seed = 42). The probe gives the model the first 60% of the question text, truncated at a word boundary, and asks it to complete the remaining text exactly as previously seen, or output UNKNOWN if it has not seen the item. We compare the model output to the held-out remaining 40% of the question text using Levenshtein ratio and compute SBERT similarity to the full question text with all-MiniLM-L6-v2.

No model exceeds the Lev>0.7 threshold on any probed item, and no model exceeds Lev>0.9.

Qwen-32B and Llama-70B show higher mean SBERT similarity than the other models, but their $\text{Lev} > 0.7$ rate remains 0%; we interpret these cases as topically coherent science responses rather than near-verbatim recall. This probe provides evidence against verbatim item-text memorization, but it cannot rule out paraphrased exposure, exposure to misconception descriptions, or exposure to aggregate response statistics.

C.5 Sampling-Variance Audit

We ran a variance experiment on GPT-5.4 to assess the stability of single-run estimates. We sampled 100 items stratified by topic (seed=54) and ran 3 replicate calls per (item, distractor) pair for Task A and per item for Task C, all with `reasoning_effort=low`.

Task A. Of 231 (item, distractor) evaluations with all three replicates complete, 96.1% had all three replicates agree, meaning that all three were either correct or incorrect. Majority-vote accuracy across replicates was 90.9%, supporting the stability of the single-run main estimate. The all-agree rate of 96.1% indicates that disagreement across replicates is rare even for a reasoning model run with `reasoning_effort=low`.

Task C. The mean across-replicate standard deviation of predicted distributions was 0.046 (noise/signal ratio 13.9%; signal is the mean L_1 distance between the model prediction and the uniform distribution). No items showed deviation ≥ 2 standard deviations from the replicate mean. These stability results support the use of single-run estimates throughout the paper.

GPT-5 Task A stability (single-rep check). To verify that GPT-5 main-run Task A estimates are similarly stable to GPT-5.4, we ran a single additional inference pass on the same 100-item variance sample using `reasoning_effort=low`. On the 231 matched (item, distractor) pairs, Hit@1 changed from 91.3% (main run) to 92.2% (rep 1), a delta of +0.9pp. The all-agree rate between the two runs is 96.5% (223/231), comparable to GPT-5.4’s 96.1% across three reps. We treat this as evidence that GPT-5 main-run Task A estimates have variance of the same order as GPT-5.4’s, despite the budget constraint that prevented a full three-rep experiment for GPT-5.

Task B variance check. The main variance experiment covered Task A and Task C; Task B

Table 9: Task B Hit@ r at `reasoning_effort=low` vs. high (1,474 pairs each). Δ is high – low; 95% paired bootstrap CI.

Model	Metric	Low	High	Δ	95% CI
GPT-5	Hit@1	37.8	38.3	+0.5	−1.8, +2.7
	Hit@3	66.7	65.4	−1.3	−3.5, +1.0
	Hit@5	76.9	76.7	−0.2	−2.2, +1.8
	Hit@10	90.2	88.2	−2.0	−3.3, −0.5
GPT-5.4	Hit@1	39.6	39.6	0.0	−2.3, +2.2
	Hit@3	65.5	65.9	+0.4	−1.8, +2.5
	Hit@5	76.7	76.5	−0.3	−2.0, +1.5
	Hit@10	89.5	89.2	−0.3	−1.8, +1.3

Hit@ r was not part of the original design. We ran one additional GPT-5.4 rep on the same 100-item variance sample. On the 231 matched (item, distractor) pairs, Hit@1 and Hit@5 are stable within 1.5pp (rep 1 vs main: −0.4pp and +1.3pp respectively); Hit@10 is stable within 1.7pp; Hit@3 shows a larger swing of +9.5pp on this subset. The Hit@3 swing is consistent with sampling noise at $n = 231$: the full-set Hit@3 comparison from the reasoning-effort robustness check (Section C.6) shows Hit@3 deltas within ± 1.3 pp when evaluated on all 1,474 pairs.

C.6 Reasoning-Effort Robustness Check (Task B)

To verify that the choice of `reasoning_effort=low` for the main evaluation does not materially affect our findings, we re-ran GPT-5 and GPT-5.4 on the full 1,474 (item, distractor) pair set with `reasoning_effort=high`. All other parameters match the main run. Table 9 reports Hit@ r under both settings with paired bootstrap confidence intervals (5,000 resamples) on the deltas.

For GPT-5.4, all four metric deltas are within 0.5 percentage points with CIs containing zero, indicating that high reasoning effort does not improve Task B performance. For GPT-5, Hit@1, Hit@3, and Hit@5 show statistically indistinguishable deltas; Hit@10 shows a small but reliable decrease under high effort (−2.0pp, CI [−3.3, −0.5]), consistent with reduced top- k diversity under extended deliberation. We conclude that our main-run low-effort estimates are not artifacts of the chosen reasoning budget.