Conservative Reference-Based Judging of AI-Generated STEM Analogies Using an Open Language Model

Lauren Mercer[1], Maya Kalyani Reed[2], Hoang Justin Tran[3], Alexander J. Garnier[4*]

[1]Simon Fraser University, Canada.
[2]University of Missouri, Columbia, United States.
[3]Université Grenoble Alpes, France.
[4*]Concordia University, Canada.

*Corresponding author(s). E-mail(s): alexander.j.garnier@concordia.ca;
Contributing authors: lauren.mercer@sfu.ca;
maya.kalyani.reed@umsystem.edu;
hoang.justin.tran@etu.univ-grenoble-alpes.fr;

Abstract

Generative AI can rapidly produce analogies for STEM teaching, but fluent comparisons may encode false relations or obscure boundary conditions. This study evaluates whether a locally served open LLM, Qwen3-14B, can act as a conservative reference-grounded screen for school-level analogies. The judge assessed 376 public English and Hindi items under three decision policies, yielding 1,128 parseable calls. It rejected every planted wrong probe, although the probe set represented only five unique misconception structures and the Hindi probes were direct translations. Across tiers, all good and decent probes were accepted, 29 of 30 vague observations were routed away from acceptance, and all 30 wrong observations were rejected. For 256 ordinary English generations, source-domain scores differed overall, but corrected pairwise tests supported only limited contrasts, mainly lower scores for cooking analogies. Against original-panel silver labels, an English TF–IDF and metadata model achieved macro-F1 of 0.706 with item-level cross-validation and 0.661 with concept-grouped folds, while the best English Qwen policy reached 0.632. The evidence supports open-LLM screening for explicit reference contradictions, but broader claims about instructional quality remain constrained by policy choice, language scope, translated probes, surface-predictable labels, and the absence of teacher ratings or learner outcomes.

Keywords: AI education, assessment, scholarly integrity

Corresponding author ORCID: https://orcid.org/0009-0007-5211-5722

1 Introduction

AI-generated teaching materials create a practical assessment problem for educational technology: candidate explanations can be produced cheaply, but checking whether they preserve the intended conceptual relation is slower and more expert-dependent. Analogies are a high-risk case because an appealing comparison may make an abstract STEM idea memorable while also strengthening a misconception if a boundary condition or relation is wrong. This study contributes a diagnostic evaluation of reference-grounded screening rather than evidence of classroom effectiveness. Using public English and Hindi analogy items, three decision policies, and a locally served Qwen3-14B judge, it examines what an open model reliably accepts, rejects, or routes away from automatic use. It also tests whether original-panel silver labels contain surface-predictable signal by comparing LLM decisions with lexical and metadata baselines.

Teaching analogies connect an unfamiliar idea to a familiar situation. In science and mathematics classrooms, this connection can make an abstract mechanism easier to discuss, visualize, and remember. The same mechanism can also cause harm: if a fluent analogy maps the wrong relation, treats a partial similarity as an equivalence, or omits the point at which the comparison breaks down, it may give a learner a durable misconception rather than a useful explanation.

Large language models (LLMs) make this problem more consequential because they can generate many candidate analogies for lessons, tutoring systems, and educational content libraries. Generation is inexpensive; evaluating the conceptual mapping is not. An automatic screen could help by accepting clearly suitable candidates and routing uncertain or contradictory candidates away from automatic use. Such a screen, however, is useful only within the limits of the evidence used to evaluate it. Format compliance is not conceptual validity, a score is not a validated probability, and agreement with another model is not agreement with teachers or learners.

Meduri et al. [1] recently reported a small-model pipeline for generating and judging science, technology, engineering, and mathematics (STEM) teaching analogies, including planted probes designed to test whether automatic judges detect deliberately wrong examples. Their study motivates a focused question for educational content assessment: what does a stronger open judge reliably distinguish when it is given both a candidate analogy and a standard explanation of the target concept? The present study addresses that question using the public concepts, generated analogies, probes, and panel scores released with the prior work. The contribution is not a new analogy generator. It is a diagnostic evaluation of a reference-grounded screening decision.

Task-specific evaluation has likewise shaped recent machine-learning work on segmentation, novel-object captioning, referring-expression grounding, model reprogramming, and generative reconstruction [2–7].

The evaluation centers on four distinctions. First, planted wrong probes are separated from good, decent, and deliberately vague probes instead of being summarized only through a pooled "safe" category. Second, the three prompts are treated as different decision policies, because they impose different instructions and thresholds rather than expressing the same rule in interchangeable words. Third, translated Hindi probes are not counted as new misconception structures independent of their English sources. Fourth, scores from the original automatic panel are treated as silver labels whose properties must be tested against simple lexical and metadata baselines.

These distinctions lead to three research questions:

1. How does a single open LLM judge behave across good, decent, vague, and wrong reference-grounded probes?
2. How do decision policy, language subset, and source-domain strategy affect its acceptance decisions and scores?
3. How much of the original-panel silver-label signal can be predicted from non-LLM surface features?

The study uses one locally served Qwen3-14B judge, one deterministic response for each item-policy pair, and public qwen2.5:3b generations. It does not collect teacher ratings or learner outcomes. Its evidential target is therefore narrow: behavior on this fixed reference-grounded test, not general educational effectiveness.

The contributions are:

- A tier-wise evaluation of an open LLM screen for AI-generated STEM analogies, distinguishing acceptable, vague, and explicitly contradictory probes.
- A comparison of three stated decision policies across English and Hindi items, including policy-switch rates and an explicit separation between translated-probe behavior and multilingual validity.
- A paired analysis of source-domain scores that reports both the omnibus test and corrected post-hoc contrasts, avoiding a general strategy ordering unsupported by the data.
- Lexical and metadata baselines against automatic-panel silver labels, including a concept-grouped recheck that reveals substantial surface-predictable signal.

2 Related work

2.1 Analogy, teaching, and misconception risk

Structure-mapping theory characterizes analogy as a mapping of relations from a familiar base domain to a target domain rather than a match based mainly on surface resemblance [8]. This relational view is especially important in education. Reviews of instructional analogies report that analogies can aid learning, while emphasizing the need to identify the intended mapping and its limits [9–11]. Clement's bridging analogies in physics illustrate how intermediate comparisons can be used to address learners'

existing conceptions rather than merely decorate an explanation [12]. Research on conceptual-change strategies provides further evidence that instruction must engage with the structure of misconceptions, not only with presentation quality [13].

These findings imply that a useful analogy screen must distinguish relational correctness from fluency. An analogy can be vivid, grammatical, and memorable while still teaching a false causal or structural relation. Conversely, an analogy can be conceptually safe but too vague to support instruction. Generic analogy-solving and semantic-network benchmarks measure relevant reasoning abilities, but they do not directly validate whether a generated analogy is suitable for a classroom decision [14–16].

2.2 LLM-generated educational analogies

Prompted LLMs can generate analogical explanations, and early work with InstructGPT demonstrated that analogy generation can be elicited without task-specific training [17]. Creativity-oriented studies similarly show that LLMs can support cross-domain ideation [18]. Educational use requires additional evidence because the relevant outcome is not novelty alone; it is whether a comparison helps a particular learner understand the intended relations without acquiring a misconception.

Shao et al. [19] studied LLM-generated scientific analogies with students and teachers, providing human-centered evidence about understanding and classroom practice. Barakat and Kochmar [20] developed a modular educational analogy-generation pipeline, evaluated multiple model families on established resources, and validated automatic scoring against human annotators. These studies establish a broader, human-validated setting than the present experiment.

The closest data source for the current question is the public small-model STEM analogy study by Meduri et al. [1]. The present work takes from that study the concept set, released qwen2.5:3b generations, planted probes, and automatic-panel scores. What is new here is the measurement focus: one separately served open judge is evaluated under explicit acceptance policies, the probe tiers are analyzed individually, source-domain contrasts are tested pairwise, and the silver labels are compared with non-LLM baselines.

2.3 Automatic judges in educational assessment

LLM-as-judge methods can scale evaluation, but judge outputs are not neutral measurements. Work associated with MT-Bench and Chatbot Arena documented position, verbosity, and related biases in model-based judging [21]. Other studies report useful aggregate correspondence with human evaluations while also showing that performance depends on task design and instructions [22–24].

Education adds domain-specific concerns. Reviews of LLM use in education identify validity, privacy, inequity, feedback quality, and teacher oversight as central issues [25, 26]. Automated question generation, short-answer grading, generated feedback, and classroom programming assistants show the potential of model-supported educational workflows, while also demonstrating that usefulness depends on how the model

output enters a human process [27–31]. An analogy score therefore requires validation against the intended screening decision rather than acceptance on the basis of a plausible rubric alone.

2.4 Decision policies, silver labels, and language

Recent judge-evaluation research has moved beyond raw scores toward robustness, uncertainty, and calibration. Prompt-variation studies show that even non-adversarial wording can change decisions [32]. Calibration and small-sample inference work further cautions against treating ordinal judge scores as probabilities without a held-out calibration procedure [33–35]. The three prompts in this study do more than vary wording: they define distinct policies. Their differences are therefore analyzed as policy dependence, not as an estimate of paraphrase variance.

The original-panel scores are used here as silver labels rather than expert truth. Label-efficient methods, including in-context semi-supervised learning, illustrate how indirect supervision can organize a larger sample [36]. The value of such labels nevertheless depends on what they encode. If a length rule or TF–IDF model predicts them well, that result describes the scoring behavior of the panel; it does not validate the predictor as an educational judge.

Language comparisons require similar care. Simulated-student and proxy-learner work argues that model-based educational outcomes should be validated against human or item-level evidence [37, 38]. Studies of Hindi and multilingual evaluation show that language variety, script, and benchmark construction can materially affect results [39, 40]. The Hindi probes used here are translations of the English probes. They test whether the judge follows the reference-grounded task on those translated cases, but they cannot establish broad transfer to Hindi teaching materials.

3 Method

3.1 Scope and evaluation questions

We study reference-grounded screening of AI-generated school-level STEM analogies. Each judged item contains a target concept, a standard explanation, and a candidate analogy. The automatic decision is binary: accept the item under the stated policy, or route it away from automatic acceptance. The latter outcome may motivate revision, rejection, or human review, but the experiment does not distinguish those downstream actions.

The screen is evaluated on planted probe tiers and on ordinary generated analogies. The probes provide known construction labels, while the ordinary items provide a broader sample without teacher or learner ground truth. Original-panel scores are used only for a secondary silver-label comparison. We do not interpret agreement with those scores as instructional validity.

3.2 Data

The evaluated ordinary items are the released qwen2.5:3b generations from the public repository associated with Meduri et al. [1]. The English set contains 256 items: 64

concepts crossed with four source-domain strategies. The Hindi set contains 80 items: 20 concepts crossed with the same four strategies. The strategies are unconstrained, everyday, sports, and cooking.

The planted set contains 20 English and 20 Hindi items. Within each language, five concepts appear at four tiers: good, decent, vague, and wrong. The wrong tier contains an explicit contradiction or incorrect relation relative to the provided explanation. The Hindi cases are direct translations of the English probe structures, so the full set contains five unique planted misconception structures rather than ten independent language-specific structures.

Data extraction used public repository commit b884f038a75a57b46ec9a0d0acb31f d827bee097. The repository snapshot did not contain a separate code or data license file. The associated preprint is distributed under CC BY 4.0, but the reuse terms for repository files were not stated separately.

3.3 Open judge and decision policies

All new judgments used one locally served Qwen3-14B model through an OpenAI-compatible endpoint. Endpoint metadata identified the served snapshot root as 40c069 824f4251a91eefaf281ebe4c544efd3e18. Sampling used temperature 0 and a maximum of 220 generated tokens. Each call used a deterministic seed derived from the item identifier and policy name. The model received the concept, standard explanation, analogy, and the selected decision policy.

The **binary-strict policy** asked whether the analogy was safe for a student and directed the model to reject conceptual errors, over-equivalence, or hidden misleading mappings. The **rubric policy** requested five teaching-analogy dimensions and accepted an item only when conceptual accuracy and non-misleadingness were both at least 4. The **minimal policy** asked whether the analogy was acceptable for teaching without creating a misconception. All policies requested structured JSON containing an acceptance decision and score.

Because the policies encode different standards, agreement among them measures decision stability under policy choice. It does not isolate sensitivity to semantically equivalent prompt paraphrases.

3.4 Non-LLM baselines

The non-LLM baselines predict a binary silver label derived from the original automatic panel. An item is labeled silver-accept when the released mean dimensional score is at least 4.0. This cutoff is a diagnostic convention and was not fitted to teacher judgments.

The length-only baseline rescales word count and thresholds at the median. The feature model is logistic regression over analogy-text term-frequency–inverse-document-frequency (TF–IDF) features, word count, character count, sentence count, mean word length, an analogy-marker cue, subject, and source-domain strategy. The main reported result uses stratified five-fold item-level cross-validation with a fixed seed. A secondary analysis uses stratified grouped cross-validation by concept so that different strategy versions of the same concept do not appear on both sides of a fold.

The non-LLM models are evaluated on the 256 English ordinary items. Qwen results are reported both for the pooled set of 336 ordinary English and Hindi items and for the matched 256-item English subset. This separation prevents the near-uniform Hindi rejections from being confused with a direct comparison against the English-only feature models.

3.5 Metrics and statistical analysis

For planted probes, the primary summaries are tier-wise acceptance rate and mean score. Wrong-probe rejection is also reported by language-policy cell with Wilson confidence intervals because each cell contains only five wrong items. Repeated judgments of the same five structures under multiple policies are not counted as new misconception cases.

For ordinary items, we report parse rate, acceptance rate, and the proportion of items whose binary decision differs across policies. Under the rubric policy, source-domain scores are compared within the 64 English concepts using a Friedman test. Because that test only establishes that at least one condition differs, paired Wilcoxon signed-rank tests with Holm correction are used for the six pairwise contrasts.

Silver-label comparisons report accuracy, macro-F1, area under the receiver-operating-characteristic curve (AUC), and Brier score where available. The Brier values for Qwen use a linear transformation of the ordinal score and are descriptive; they are not evidence that the scores are validated probabilities.

4 Results

4.1 Execution and parseability

The experiment evaluated 376 items under three policies, producing 1,128 Qwen3-14B calls. All 1,128 responses parsed through the standardized JSON parser, and the parsed records contain no API errors. The sum of recorded model-call wall times was 756.2 seconds.

This 100.0% parse rate establishes output-format compliance for the served model and prompts used here. It does not establish conceptual accuracy. The remaining analyses therefore examine the content and stability of the resulting decisions.

4.2 Planted-probe behavior

Qwen3-14B rejected all five wrong probes in every language-policy cell. The observed rejection rate is 1.000 in each cell, but the Wilson lower bound is 0.566 because each cell contains only five wrong cases. The repeated policy and language observations should not obscure that small denominator.

Table 1 reports the tier-wise result pooled over languages and policies. Good and decent probes were accepted in all 30 observations for each tier. Wrong probes were rejected in all 30 observations. Vague probes were accepted once in 30 observations. The screen therefore separated the explicit contradictions from the good and decent items, while also routing almost all deliberately vague items away from automatic acceptance.

Table 1 Tier-wise planted-probe results pooled over five concepts, two languages, and three decision policies. The Hindi cases translate the English structures and are not independent misconception types.

Probe tier	Observations	Accept rate	Mean score
Good	30	1.000	4.733
Decent	30	1.000	4.433
Vague	30	0.033	2.167
Wrong	30	0.000	1.433

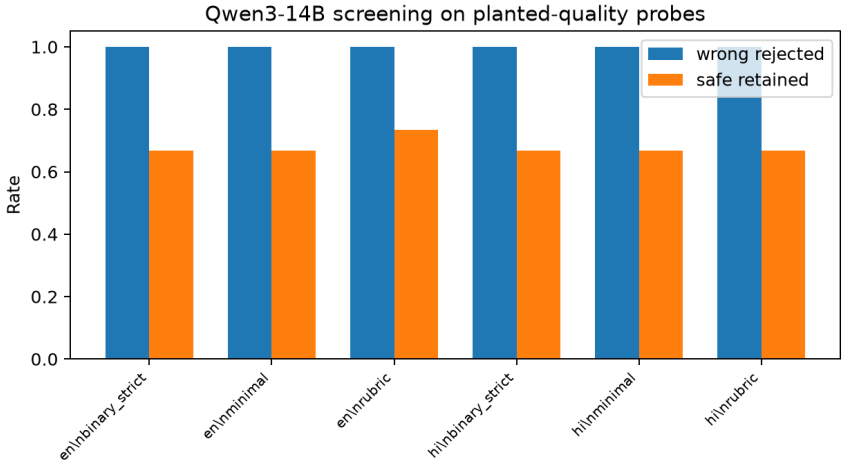


Fig. 1 Acceptance rates for the four planted-probe tiers. Values pool five concepts, two languages, and three decision policies. The Hindi cases translate the English probe structures and are not independent misconception types.

Figure 1 visualizes the same tier separation. The result supports detection of these explicit reference contradictions. It does not measure performance on subtle mapping errors, omitted boundary conditions, plausible distractors, culturally unsuitable source domains, or naturally occurring teacher-identified misconceptions.

4.3 Behavior on ordinary Hindi items

The translated Hindi probes parsed and the planted wrong cases were rejected, but the 80 ordinary Hindi generations were almost uniformly routed away from acceptance. The binary-strict policy accepted 1.25% of these items; the minimal and rubric policies accepted 0.0%. The policy-switch rate was 1.25% for Hindi, compared with 18.4% for English, because nearly all ordinary Hindi items received the same negative outcome.

The data do not identify the cause. Possible contributors include the quality of the released Hindi generations, the judge's language behavior, an English-centered rubric, register or script effects, and interactions among these factors. The result is therefore

a finding about this fixed item set and judge, not evidence of general Hindi screening performance.

4.4 Source-domain scores

For the 256 English ordinary items under the rubric policy, mean scores were 3.906 for unconstrained generation, 3.828 for everyday analogies, 3.688 for sports analogies, and 3.500 for cooking analogies. The Friedman test over 64 concepts gave $\chi^2 = 11.83$ and $p = 0.0080$, indicating an overall difference among the four conditions.

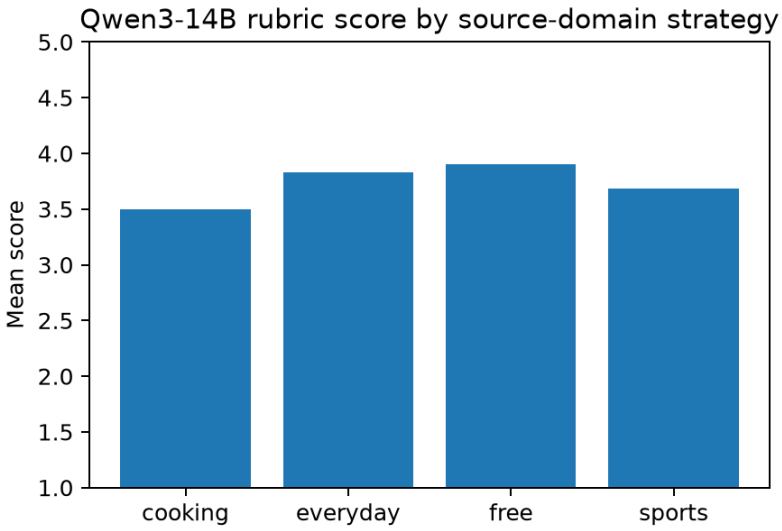


Fig. 2 Mean Qwen3-14B rubric scores for the four source-domain strategies on 256 English analogies. Error bars are 95% confidence intervals over the 64 concept-level observations.

The corrected pairwise tests narrow the interpretation of Figure 2. Unconstrained analogies scored above cooking analogies by 0.406 points (Holm $p = 0.0128$), but the unconstrained–everyday difference of 0.078 was not supported (Holm $p = 0.531$), nor was the unconstrained–sports difference of 0.219 (Holm $p = 0.142$). Everyday analogies also scored above cooking analogies by 0.328 points (Holm $p = 0.0415$). Thus, the evidence primarily identifies lower scores for the cooking condition under this judge. It does not establish a general ordering in which unconstrained generation is superior to every constrained strategy.

4.5 Policy-dependent acceptance

Across the 336 ordinary items judged under all three policies, the binary decision was unanimous for 288 items and differed for 48, a pooled switch rate of 14.3%. The language-specific rates were 18.4% for English and 1.25% for Hindi.

Acceptance levels also varied by policy. For English items, binary-strict acceptance was 53.1% for unconstrained, 57.8% for everyday, 45.3% for sports, and 32.8% for cooking analogies. Under the rubric policy, the corresponding rates were 73.4%, 68.8%, 57.8%, and 50.0%. The minimal policy produced another stated screening rule.

These differences should not be described as instability under harmless rewording. The policies ask different questions and implement different thresholds. Their disagreement instead demonstrates that the operating definition of "acceptable" is a substantive part of an automatic screen and must be specified with the reported result.

4.6 Silver-label comparisons

Table 2 compares Qwen decisions and non-LLM models against the original automatic panel's silver labels. The pooled Qwen rows include 336 English and Hindi ordinary items. The matched English Qwen row and both feature-model rows use the same 256 English items.

Table 2 Performance against original-panel silver labels. Pooled Qwen rows and English-only rows are shown separately because the feature models were evaluated only on English items. CV denotes cross-validation.

Screen	Evaluation pool	n	Accuracy	Macro-F1	AUC	Brier
Qwen3 binary strict	Pooled ordinary items	336	0.542	0.541	0.594	0.293
Qwen3 minimal	Pooled ordinary items	336	0.580	0.580	0.608	0.273
Qwen3 rubric	Pooled ordinary items	336	0.565	0.564	0.576	0.287
Qwen3 minimal	English ordinary items	256	0.641	0.632	0.648	0.243
TF–IDF/metadata logistic	English, item-level CV	256	0.711	0.706	0.754	0.200
TF–IDF/metadata logistic	English, grouped by concept	256	0.668	0.661	0.741	0.206
Length-only	English	256	0.656	0.654	0.700	—

The English TF–IDF/metadata model reached macro-F1 0.706 under item-level cross-validation, compared with 0.632 for the best English Qwen policy. Grouping folds by concept reduced the feature model to 0.661, showing that the item-level estimate benefited from allowing other source-domain versions of a concept to appear in training. Length alone reached macro-F1 0.654. Figure 3 summarizes these English-only macro-F1 values.

These results do not show that a lexical model assesses educational quality better than Qwen3-14B. The target comes from an automatic panel rather than teachers or learners. The supported conclusion is that the silver labels contain substantial signal associated with lexical, length, subject, and strategy features. A proposed educational judge should therefore be compared with simple surface baselines before its score is interpreted as a measure of conceptual or pedagogical quality.

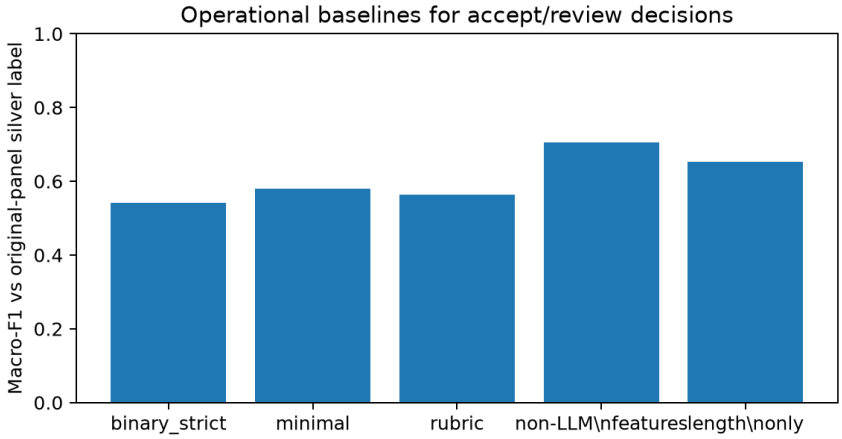


Fig. 3 Macro-F1 against the original automatic panel's silver labels. The English-only comparison includes Qwen3 minimal, a length-only rule, and TF–IDF/metadata logistic regression under item-level and concept-grouped cross-validation.

5 Discussion

5.1 A narrow role for reference-grounded screening

The clearest result is the separation of the planted tiers. On the five available misconception structures, the judge accepted every good and decent observation, rejected every wrong observation, and routed almost every vague observation away from acceptance. This pattern is consistent with a conservative first-pass screen when a standard explanation is available and the error is an explicit contradiction.

The role is narrower than general analogy evaluation. Reference-grounded screening asks whether a candidate conflicts with a supplied explanation. Pedagogical evaluation asks whether the comparison helps a learner map the intended relations, recognize the limits of the comparison, and avoid misconceptions in context. The first task can support the second, but it cannot replace teacher judgment or learner evidence. The present experiment evaluates the first task.

The treatment of vague probes is also instructive. A pooled binary division between "wrong" and "safe" would count vague items as safe and make the screen appear to reject desirable material. A classroom workflow may instead prefer to send a vague analogy for revision because it provides too little relational guidance. Tier-level reporting therefore gives a more useful description than a single safe-retention rate.

5.2 Policy is part of the measurement

The 14.3% pooled switch rate shows that the chosen policy affects the decision, especially on English items where acceptance was not already near zero. This is expected: binary-strict, rubric, and minimal prompts define different standards. The relevant lesson is not that one prompt is unreliable because it disagrees with another. It is that

an evaluation must publish and justify the policy that turns a model response into an action.

A stronger operational study would define three actions—accept, review, and reject—fit any thresholds on a calibration set, and evaluate false-accept risk and review load on a locked test set. The current experiment does not perform that procedure. Its binary non-accept outcome combines cases that may deserve revision, rejection, or human review.

5.3 Source domains and silver-label signal

The source-domain analysis illustrates why omnibus significance should be followed by pairwise interpretation. Although the Friedman test detected an overall difference, only contrasts involving cooking survived Holm correction. The result is compatible with cooking being a less suitable source domain for some concepts in this sample, but it does not establish an across-the-board benefit for unconstrained generation.

The silver-label analysis raises a separate measurement issue. Item-level TF–IDF/metadata prediction exceeded the matched English Qwen result, and concept grouping reduced but did not remove that advantage. This means that the automatic panel's labels are partly predictable from surface and metadata features. It does not identify which features are causal, nor does it show that those labels are wrong. It does show that agreement with the panel cannot, by itself, establish conceptual validity.

Together, the findings suggest a practical evaluation sequence for future work: begin with explicit misconception probes; report each probe tier; define the decision policy before testing; compare with surface baselines; separate languages and evaluation pools; group related concepts across splits; and obtain teacher or learner labels before making claims about instructional quality.

6 Limitations

The study has no new human ground truth. The five unique wrong probes provide known explicit contradictions, but the ordinary analogies have no blinded STEM-teacher labels, adjudicated error types, learner ratings, or measured learning outcomes. The original-panel threshold of 4.0 is an automatic silver label, not an expert standard.

The probe set is small and structurally simple. Hindi probes translate the same five English misconception structures and do not constitute independent multilingual evidence. The judge receives the standard explanation, and there is no reference-free ablation. The experiment therefore does not test subtle missing relations, plausible but misleading mappings, omitted boundary conditions, culturally unsuitable analogies, or naturally authored Hindi misconceptions.

Model and prompt coverage are limited to one Qwen3-14B snapshot, one deterministic response for each item-policy pair, one older qwen2.5:3b generator family for ordinary items, and three policies that are not controlled paraphrases. There are no judge-family, model-size, quantization, decoding-repetition, or controlled prompt-paraphrase comparisons. The near-total rejection of ordinary Hindi items is not diagnostically separable into generator, judge, rubric, or language effects.

The statistical analyses are exploratory. Qwen and feature baselines are comparable only on the explicitly matched English subset. Item-level cross-validation allows related versions of a concept across folds; the grouped result reduces that leakage but still predicts silver labels. No held-out calibration split, reliability analysis, risk–coverage curve, or pre-specified three-action policy is available. Finally, raw responses, parsed tables, seeds, dataset commit hashes, and endpoint snapshot metadata are recorded, but dependencies, serving arguments, hardware, and quantization are not fully pinned, and the repository files lack a separately stated reuse license.

7 Conclusion

A single open LLM can provide a useful but narrow reference-grounded screen for AI-generated STEM teaching analogies. In this case study, Qwen3-14B produced parseable output for all 1,128 calls and separated five explicit planted misconception structures from good and decent probes under three decision policies. It also rejected nearly all vague probes, almost uniformly rejected the ordinary Hindi set, produced policy-dependent decisions on English items, and assigned lower scores chiefly to cooking analogies. Simple lexical and metadata models predicted the automatic-panel silver labels strongly, including under concept-grouped cross-validation.

The evidence supports conservative screening for explicit contradictions when a standard explanation is available. It does not validate automatic scores as measures of instructional quality, broad language transfer, or estimated risk. Those claims require harder misconception cases, native-language evaluation, multiple judges and decoding runs, leakage-resistant splits, and teacher or learner ground truth.

Data availability statement

The public source data are available from https://github.com/Sriharsha-Meduri/Automatic-Analogy-Generation-STEM at commit b884f038a75a57b46ec9a0d0acb31fd827bee097. The associated preprint is available at doi:10.20944/preprints202610.0397.v1. Derived tables and figures are included with this manuscript package.

Code availability

Code is available from the corresponding author and will be released at publication.

CRediT author contribution statement

Lauren Mercer, Maya Kalyani Reed, Hoang Justin Tran, Alexander J. Garnier: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—original draft, Writing—review and editing, Visualization, and Project administration.

Funding

This research received no external funding.

Declaration of competing interest

The author declares no competing interests.

Artificial Intelligence Use Statement (Declaração de uso de Inteligência Artificial)

During the preparation of this work the authors used artificial intelligence (AI)-assisted tools for literature search, code development and drafting. The authors reviewed and edited the content and take full responsibility for it.

References

[1] Meduri, S., Rathore, M.P.S., Kalsi, G.: Evaluating and Assuring AI-Generated STEM Teaching Analogies with Small Models. MDPI AG (2026). https://doi.org/10.20944/preprints202610.0397.v1 . http://dx.doi.org/10.20944/preprints202610.0397.v1

[2] Fan, J., Gao, B., Jin, H., Jiang, L.: Ucc: Uncertainty guided cross-head cotraining for semi-supervised semantic segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9937–9946. IEEE, ??? (2022). https://doi.org/10.1109/cvpr52688.2022.00971 . http://dx.doi.org/10.1109/cvpr52688.2022.00971

[3] Fan, J., Liang, Y., Liu, L., Huang, S., Zhang, L.: Rca-noc: Relative contrastive alignment for novel object captioning. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15464–15474. IEEE, ??? (2023). https://doi.org/10.1109/iccv51070.2023.01422 . http://dx.doi.org/10.1109/iccv51070.2023.01422

[4] Liang, Y., Yang, Z., Tang, Y., Fan, J., Li, Z., Wang, J., Torr, P.H.S., Huang, S.-L.: Luna: Language as continuing anchors for referring expression comprehension. In: Proceedings of the 31st ACM International Conference on Multimedia. MM '23, pp. 5174–5184. ACM, ??? (2023). https://doi.org/10.1145/3581783.3612584 . http://dx.doi.org/10.1145/3581783.3612584

[5] Liu, L., Kong, T., Zhu, M., Fan, J., Fang, L.: Clickseg: 3d instance segmentation with click-level weak annotations. arXiv preprint arXiv:2307.09732 (2023)

[6] Chung, M.-Y., Fan, J., Ye, H., Wang, Q., Shen, W.-C., Yu, C.-M., Chen, P.-Y., Kuo, S.-Y.: Model reprogramming demystified: A neural tangent kernel perspective. arXiv preprint arXiv:2506.00620 (2025)

[7] Yu, T., Ekbote, C., Morozov, N., Herrera, A., Fan, J., Dominguez Mantes, A., Novello, S., Lashuel, H., Frossard, P., d'Ascoli, S., La Manno, G., Brbic, M.:

Tissue Reconstruction with Generative AI. openRxiv (2025). https://doi.org/10.1101/2025.02.13.638045 . http://dx.doi.org/10.1101/2025.02.13.638045

[8] Gentner, D.: Structure-mapping: A theoretical framework for analogy*. Cognitive Science **7**(2), 155–170 (1983) https://doi.org/10.1207/s15516709cog0702_3

[9] Duit, R.: On the role of analogies and metaphors in learning science. Science Education **75**(6), 649–672 (1991) https://doi.org/10.1002/sce.3730750606

[10] Dagher, Z.R.: Review of studies on the effectiveness of instructional analogies in science education. Science Education **79**(3), 295–312 (1995) https://doi.org/10.1002/sce.3730790305

[11] Gray, M.E., Holyoak, K.J.: Teaching by analogy: From theory to practice. Mind, Brain, and Education **15**(3), 250–263 (2021) https://doi.org/10.1111/mbe.12288

[12] Clement, J.: Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics. Journal of Research in Science Teaching **30**(10), 1241–1257 (1993) https://doi.org/10.1002/tea.3660301007

[13] Pacaci, C., Ustun, U., Ozdemir, O.F.: Effectiveness of conceptual change strategies in science education: A meta-analysis. Journal of Research in Science Teaching **61**(6), 1263–1325 (2023) https://doi.org/10.1002/tea.21887

[14] Boteanu, A., Chernova, S.: Solving and explaining analogy questions using semantic networks. Proceedings of the AAAI Conference on Artificial Intelligence **29**(1) (2015) https://doi.org/10.1609/aaai.v29i1.9400

[15] Jiayang, C., Qiu, L., Chan, T.H., Fang, T., Wang, W., Chan, C., Ru, D., Guo, Q., Zhang, H., Song, Y., Zhang, Y., Zhang, Z.: StoryAnalogy: Deriving Story-level Analogies from Large Language Models to Unlock Analogical Understanding. arXiv (2023). https://doi.org/10.48550/ARXIV.2310.12874 . https://arxiv.org/abs/2310.12874

[16] Wijesiriwardene, T., Wickramarachchi, R., Gajera, B.G., Gowaikar, S.M., Gupta, C., Chadha, A., Reganti, A.N., Sheth, A., Das, A.: ANALOGICAL – A Novel Benchmark for Long Text Analogy Evaluation in Large Language Models. arXiv (2023). https://doi.org/10.48550/ARXIV.2305.05050 . https://arxiv.org/abs/2305.05050

[17] Bhavya, B., Xiong, J., Zhai, C.: Analogy generation by prompting large language models: A case study of instructgpt. In: Proceedings of the 15th International Conference on Natural Language Generation, pp. 298–312. Association for Computational Linguistics, ??? (2022). https://doi.org/10.18653/v1/2022.inlg-main.25 . http://dx.doi.org/10.18653/v1/2022.inlg-main.25

[18] Ding, Z., Srinivasan, A., Macneil, S., Chan, J.: Fluid transformers and creative analogies: Exploring large language models' capacity for augmenting cross-domain analogical creativity. In: Creativity and Cognition. CandC '23, pp. 489–505. ACM, ??? (2023). https://doi.org/10.1145/3591196.3593516 . http://dx.doi.org/10.1145/3591196.3593516

[19] Shao, Z., Yuan, S., Gao, L., He, Y., Yang, D., Chen, S.: Unlocking scientific concepts: How effective are llm-generated analogies for student understanding and classroom practice? In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. CHI '25, pp. 1–19. ACM, ??? (2025). https://doi.org/10.1145/3706598.3714313 . http://dx.doi.org/10.1145/3706598.3714313

[20] Barakat, M., Kochmar, E.: Teaching through analogies: A modular pipeline for educational analogy generation. In: Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026), pp. 863–898. Association for Computational Linguistics, ??? (2026). https://doi.org/10.18653/v1/2026.bea-1.59 . http://dx.doi.org/10.18653/v1/2026.bea-1.59

[21] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena. arXiv preprint arXiv:2306.05685 (2023)

[22] Chiang, C.-H., Lee, H.-y.: Can Large Language Models Be an Alternative to Human Evaluations? arXiv (2023). https://doi.org/10.48550/ARXIV.2305.01937 . https://arxiv.org/abs/2305.01937

[23] Huidrom, R., Belz, A.: Ask me like i'm human: Llm-based evaluation with for-human instructions correlates better with human evaluations than human judges. In: Proceedings of the 4th Table Representation Learning Workshop, pp. 98–108. Association for Computational Linguistics, ??? (2025). https://doi.org/10.18653/v1/2025.trl-1.9 . http://dx.doi.org/10.18653/v1/2025.trl-1.9

[24] Wang, Y., Yu, Z., Zeng, Z., Yang, L., Wang, C., Chen, H., Jiang, C., Xie, R., Wang, J., Xie, X., Ye, W., Zhang, S., Zhang, Y.: PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. arXiv (2023). https://doi.org/10.48550/ARXIV.2306.05087 . https://arxiv.org/abs/2306.05087

[25] Kasneci, E., Sessler, K., Kuchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Gunnemann, S., Hullermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., Kasneci, G.: Chatgpt for good? on opportunities and challenges of large language models for education. Learning and Individual Differences **103**, 102274 (2023) https://doi.org/10.1016/j.lindif.2023.102274

[26] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., Gasevic, D.: Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology

55(1), 90–112 (2023) https://doi.org/10.1111/bjet.13370

[27] Qu, F., Jia, X., Wu, Y.: Asking questions like educational experts: Automatically generating question-answer pairs on real-world examination data. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 2583–2593. Association for Computational Linguistics, ??? (2021). https://doi.org/10.18653/v1/2021.emnlp-main.202 . http://dx.doi.org/10.18653/v1/2021.emnlp-main.202

[28] Kortemeyer, G.: Performance of the pre-trained large language model gpt-4 on automated short answer grading. Discover Artificial Intelligence **4**(1) (2024) https://doi.org/10.1007/s44163-024-00147-y

[29] Bernius, J.P., Krusche, S., Bruegge, B.: Machine learning based feedback on textual student answers in large courses. Computers and Education: Artificial Intelligence **3**, 100081 (2022) https://doi.org/10.1016/j.caeai.2022.100081

[30] Lyu, W., Wang, Y., Chung, T.R., Sun, Y., Zhang, Y.: Evaluating the effectiveness of llms in introductory computer science education: A semester-long field study. In: Proceedings of the Eleventh ACM Conference on Learning @ Scale. L@S '24, pp. 63–74. ACM, ??? (2024). https://doi.org/10.1145/3657604.3662036 . http://dx.doi.org/10.1145/3657604.3662036

[31] Kazemitabaar, M., Ye, R., Wang, X., Henley, A.Z., Denny, P., Craig, M., Grossman, T.: Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. CHI '24, pp. 1–20. ACM, ??? (2024). https://doi.org/10.1145/3613904.3642773 . http://dx.doi.org/10.1145/3613904.3642773

[32] Bhat, S., Varma, V.: All prompts are created equal? evaluating robustness of llm judges against non-adversarial prompt variations. In: Findings of the Association for Computational Linguistics: ACL 2026, pp. 38730–38745. Association for Computational Linguistics, ??? (2026). https://doi.org/10.18653/v1/2026.findings-acl.1929 . http://dx.doi.org/10.18653/v1/2026.findings-acl.1929

[33] Radharapu, B., Saxena, E., Li, K., Whitehouse, C., Williams, A., Cancedda, N.: Calibrating llm judges: Linear probes for fast and reliable uncertainty estimation. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pp. 193–216. Association for Computational Linguistics, ??? (2026). https://doi.org/10.18653/v1/2026.acl-industry.14 . http://dx.doi.org/10.18653/v1/2026.acl-industry.14

[34] Li, H., Chen, J., Ai, Q., Chu, Z., Zhou, Y., Dong, Q., Liu, Y.: Calibraeval: Calibrating prediction distribution to mitigate selection bias in llms-as-judges. arXiv preprint arXiv:2410.15393 (2024)

[35] Arawjo, I.: How to run statistics over llm judges and trust the results: Calibrated inference for small-sample ai evaluation with evalstats. arXiv preprint arXiv:2609.35815 (2026)

[36] Fan, J., Rosu, P., Wang, A.T., Li, Z.M., Carin, L., Cheng, X.: In-context semi-supervised learning. arXiv preprint arXiv:2512.15934 (2025)

[37] Scarlatos, A., Fernandez, N., Ormerod, C., Lottridge, S., Lan, A.: Smart: Simulated students aligned with item response theory for question difficulty prediction. arXiv preprint arXiv:2507.05129 (2025)

[38] Yuan, Z., Xiao, Y., Li, M., Xuan, W., Tong, R., Diab, M., Mitchell, T.: Towards valid student simulation with large language models. arXiv preprint arXiv:2601.05473 (2026)

[39] Gharami, K., Muhtaseem, Q.S., Gupta, D., Elluri, L., Moni, S.S.: Modeling romanized hindi and bengali: Dataset creation and multilingual llm integration. In: 2025 IEEE International Conference on Big Data (BigData), pp. 6569–6576. IEEE, ??? (2025). https://doi.org/10.1109/bigdata66926.2025.11401059 . http://dx.doi.org/10.1109/bigdata66926.2025.11401059

[40] Kumar, P.: Do Hindi LLM Benchmarks Agree? A Cross-benchmark Consistency Study and the Case for Harmonized Evaluation. Elsevier BV (2026). https://doi.org/10.2139/ssrn.7620581 . http://dx.doi.org/10.2139/ssrn.7620581