

Comparing the Ability of LLMs to Identify Surmise Relations Between Competencies

Peter Steiner^{1*}, Jamilla Soares Lobo^{2*}, Rafael Mello^{2,3}

¹St. Gallen University of Teacher Education

²Centro de Estudos e Sistemas Avançados do Recife (CESAR School)

³Universidade Federal Rural de Pernambuco

peter.steiner@phsg.ch, jsl@cesar.school, rafael.mello@ufrpe.br

Abstract. Adaptive instruction tailored to learners' knowledge states requires fine-grained assessment that can be interpreted regarding the next feasible learning steps. Modern psychometric frameworks leverage graphs of competencies connected by surmise relations. However, both data- and expert-based elicitation of these surmise relations scale only to a limited extent. This contribution investigates whether Large Language Models (LLMs) can serve as a scalable alternative by evaluating their performance in a zero-shot setting. Therefore, the results reveal a clear performance gap between commercial and open-source models, with Claude Haiku 4.5 and Gemini 2.5 flash achieving the best results. The findings suggest that top-tier LLMs can partially automate the elaboration of surmise relations, though human-in-the-loop validation remains necessary, particularly for semantically ambiguous relation types.

Keywords: Automated curriculum analysis, Learning progressions, Prerequisite relations, Competence Structures

1. Introduction

Adapting instruction to individual students' needs requires reliable information about their current learning state and their next feasible learning steps [Anselmi et al. 2024, Wu et al. 2024]. However, many practitioners find that conventional scale-based assessment methods, such as Item Response Theory (IRT) [de Ayala 2022], fall short of informing teaching meaningfully [Bez et al. 2025, Huff and Goodman 2007]. This limitation is further emphasized when such methods serve as the psychometric foundation for digital technologies. For example, intelligent tutoring systems that aim to recommend next learning steps automatically [Thompson and Clark 2024, Käser et al. 2017].

Consequently, several recent psychometric frameworks, including Knowledge Space Theory (KST) [Doignon and Falmagne 1985], Hierarchical Diagnostic Classification Models [Templin and Bradshaw 2014], and Dynamic Bayesian Networks [Käser et al. 2017], conduct student assessment within hierarchical competency structures. A competence structure is a set of competencies that is structured by a surmise relation that specifies prerequisite dependencies among competencies: if competency *A* is a prerequisite for *B*, proficiency in *A* supports the likelihood of success in *B*.

*These authors contributed equally to this work and share first authorship.

Through this relation, a quasi-order on the competence set can be established. Therefore, a competence structure can be visualized as a directed acyclic graph, with nodes representing the theoretically possible learning states and edges representing the learning steps. Assessments based on these structures have the potential to be more efficient [Ihichr et al. 2024, Anselmi et al. 2024] and are inherently more interpretable: their results directly encode which competencies a student has mastered and which learning steps are immediately feasible. This makes the assessment outcomes actionable for teachers and digital learning technologies without further inference.

Despite the advantages, competence structures have not found widespread application [Ihichr et al. 2024]. Their primary limitation is the cost of elaboration [Segedinac et al. 2018, de Chiusole et al. 2024]. In some contexts, competence structures have been established by expert teams. However, such efforts are resource-intensive and difficult to scale. Curricula differ substantially across countries in scope, sequence, and standards, making transfer of existing structures to other contexts generally infeasible [Cosyn and Thiéry 2000].

In this paper, we investigate whether recent advances in large language models (LLMs) can address this gap. Specifically, we examine whether LLMs can reliably identify pairwise surmise relations among mathematical competencies in a zero-shot setting, when evaluated against expert-established ground truth from the Learning Commons Knowledge Graph. We further analyze how model size and model type (commercial vs. open source) affect performance. This provides the first systematic evidence on the conditions under which LLM-based methods can serve as a scalable alternative to existing approaches.

2. Related Work

Although using competence structures is resource-intensive, they are already operationalized in large-scale adaptive assessment systems. The learning platform ALEKS, owned by McGraw-Hill, applies KST to drive both item selection and the recommendation of next learning steps: at any point during assessment, the set of immediately learnable topics can be read out directly from the competence structure given the student's current learning state [Cosyn et al. 2021].

Dynamic Learning Maps (DLM) takes a similar approach for students with significant cognitive disabilities, representing competencies and their dependencies as a fine-grained competence structure. The map underpins both summative and formative assessment across multiple US states, with the learning progression guiding inferences about student mastery and informing subsequent instructional decisions [Thompson and Nash 2022].

Both systems demonstrate that competence structures can be reliably deployed at scale. Critically, in both cases the validity of the assessment is contingent on the quality of the underlying structure—making the reliable elaboration of surmise relations a foundational methodological problem.

Existing methods can be broadly categorized into data-driven and expert-based approaches [Segedinac et al. 2018], each with distinct strengths and limitations. Inductive Item Tree Analysis (IITA) [Sargin and Ünlü 2009, Ünlü and Schrepp 2021] is a well-

validated method for deriving surmise relations directly from binary response data. Several variants have been proposed, each with well-characterized operating properties, and the method has demonstrated reliable performance across a range of empirical applications. Despite this, IITA is subject to three limitations that constrain its applicability in practice. First, stable estimates require large samples, which are often unavailable in realistic school assessment contexts. Second, all currently published variants lack support for missing data, rendering them inapplicable in multi-booklet designs or other settings with incomplete response matrices. Third, IITA operates exclusively on observable binary item responses and cannot be applied to latent competencies that are not directly measurable through individual items. Together, these constraints mean that data-driven approaches are frequently inapplicable precisely in the contexts where competence structures are most needed.

Most existing competence structures, including the Learning Commons Knowledge Graph, have been established by content expert teams, a process that is both expensive and time-intensive. The QUERY algorithm [Koppen and Doignon 1990, Kambouri et al. 1994] provides a structured approach to this task by eliciting pairwise relations through systematic expert queries. However, the number of required queries scales quadratically with the number of competencies, such that expert fatigue becomes a practical constraint as granularity increases [Cosyn and Thiéry 2000].

Neither data-driven nor expert-based methods therefore provide a scalable and reliable solution for establishing competence structures. Recent advances in machine learning, and particularly in large language models, offer one potential alternative. Several studies have addressed related problems using natural language processing, representation learning, and graph-based machine learning techniques. [Roy et al. 2019] proposed PREREQ, a supervised learning approach that combines latent concept representations learned from educational resources with a Siamese neural network to infer concept surmise relations. [Pan et al. 2017] framed surmise relation learning as a supervised classification problem based on semantic, contextual, and structural features. [Zhang et al. 2025] proposed a graph-based global knowledge relation optimization model for learning concept surmise relations jointly with other document- and concept-level relations. While these contributions demonstrate that surmise relations can be approximated from educational texts and structured learning resources, they rely on labeled concept relations, engineered features or graph constructions, and non-trivial model training procedures.

The present paper takes a different approach by investigating whether competence structures can be established through direct zero-shot LLM prompting, without training data, feature engineering, or task-specific pipelines. Closest to this aim, [Le and Abel 2025] recently examined whether LLMs can predict prerequisite skills in a zero-shot setting. However, their task is formulated as an open-ended generation problem: given a target skill, the model is asked to produce a list of possible prerequisite skills. In contrast, the present study treats competence-structure construction as a pairwise relation-classification problem, which corresponds more closely to established qualitative elicitation procedures such as QUERY. To our knowledge, no existing work has systematically evaluated under which conditions LLMs can perform this pairwise classification task against expert-established ground truth at scale.

Several factors are likely to moderate performance, including model size, model

type, and prompting strategy. As a first and most practically relevant baseline, this study focuses on the zero-shot setting, which requires no labeled examples and therefore preserves the scalability and cost-efficiency that motivate the approach. Within this setting, we further contrast commercial models with open-source alternatives: while commercial models generally offer superior performance, their use in educational contexts raises concerns around data privacy, cost, and long-term availability—considerations that may favor open-source alternatives if their performance proves sufficient.

2.1. Research Question

This study aims to investigate whether LLMs can serve as a scalable alternative to expert-based and data-driven methods for establishing competence structures. Specifically, we examine the extent to which LLMs can reliably identify pairwise surmise relations among mathematical competencies in a zero-shot setting, and whether performance varies systematically across model sizes and families. To this end, the following research questions guide the study:

RESEARCH QUESTION 1 (RQ1):

Can LLMs reliably identify pairwise surmise relations among mathematical competencies in a zero-shot setting?

RESEARCH QUESTION 2 (RQ2):

Which relation category is the most challenging for LLMs to classify correctly, and how do structural factors in the model (size, type) affect this?

3. Method

3.1. Data Understanding

This study uses data from the Learning Commons Knowledge Graph API¹, a structured repository of academic standards and their interrelationships maintained through expert curation and cross-state alignment efforts. Since standards such as the CCSSM specify what students are expected to understand and be able to do, we operationalize individual standards as competency-like units in this study. We use the term competencies to emphasize their role as assessable units of knowledge or skill that can be related through a surmise relation. The Learning Commons API exposes two primary relationship types between competencies: `buildsTowards`, a directed relation indicates that mastery of one competency may support success on another; and `relatesTo`, a symmetric relation indicating meaningful conceptual overlap without implying a specific sequence or clear directed surmise relation.

Data collection focused on mathematics competencies under the multi-state jurisdiction, restricted to items of normalized statement type competencies (excluding grouping headers). They were retrieved through nine thematic search queries corresponding to the major content domains established by the NCTM principles and standards [NCTM 2000] and the CCSS-M domain structure [of Chief State School Officers 2010]: number and operations, fractions, ratios and proportional reasoning, expressions and

¹<https://docs.learningcommons.org/knowledge-graph/entity-and-relationship-reference/academic-standards>

equations, functions, geometry, statistics and probability, measurement, and base-ten operations. This domain-aligned retrieval strategy reflects a core premise of competence structures research—that meaningful surmise relations are concentrated within coherent mathematical domains rather than across them [Daro et al. 2011], making thematic coverage a theoretically motivated choice rather than an arbitrary one. Queries were executed across four grade-level groups (grades 3–5, 6–8, 9–12, and unfiltered) to ensure broad grade coverage.

The initial collection yielded 479 seed competencies, which were subsequently expanded to 509 nodes after relation traversal added competencies referenced as targets of buildsTowards or relatesTo edges. The Knowledge Graph contains 945 direct edges in total, of which 694 are directed buildsTowards edges and 132 are undirected relatesTo edges. Each standard is described by a statement code following the Common Core naming convention (e.g., 4.NF.A.2), a natural-language description, and metadata fields including grade level, mathematical domain, cluster, and jurisdiction.

3.2. Data Preparation

From the retrieved competencies and their relationships, a graph structure was constructed using NetworkX library, where the relations were direct, as for this study, we do not consider the transitivity that might appear in distant relationships.

The benchmark dataset was then constructed by enumerating unordered pairs from the 479 seed competencies. Given the combinatorial scale of all possible pairs ($\binom{479}{2} = 114,481$), a balanced sampling strategy was adopted: all positive pairs (those with a known relation) were retained, and an equal number of negative pairs (noRelation) were randomly sampled. This resulted in a dataset of approximately 1,535 pairs. Each pair (A, B) was assigned one of four ground truth labels:

- **buildsTowards_A_to_B:** A direct buildsTowards edge exists from A to B .
- **buildsTowards_B_to_A:** A direct buildsTowards edge exists from B to A .
- **relatesTo:** An undirected relatesTo edge exists between A and B .
- **noRelation:** No direct relation exists between A and B in either graph.

We selected a subset of 20% from the labeled dataset to run the experiments, with a fixed random seed of 42 to ensure reproducibility. The test set comprises 307 pairs with the following class distribution:

Table 1. Relation Distribution

Label	Count	Proportion
noRelation	153	49.8%
buildsTowards_A_to_B	73	23.8%
buildsTowards_B_to_A	57	18.6%
relatesTo	24	7.8%

All LLM evaluations were conducted exclusively on this test set. The test pairs were serialized as hydrated JSON objects containing the full standard metadata for both competencies, the computed structural features, and the ground truth label.

3.3. LLM Models and Prompt

This study employed five LLMs to evaluate the relationships among the academic competencies in the benchmark dataset. To ensure a representative sample of the current LLM landscape, our selection features proprietary models from prominent AI companies and resource-friendly open-source LLMs. This approach aligns with literature recommendations that stress the importance of comparing different model families in evaluating educational solutions [Singh et al. 2024]. Our evaluation framework specifically incorporates five distinct LLM implementations:

- **OpenAI GPT-5-mini²**: chosen for its speed and cost-efficiency compared to GPT-5. Ideal for well-defined, lightweight tasks and structured prompts [OpenAI 2026].
- **Google Gemini 2.5 Flash³**: included as a light counterpoint (flash), the model was built for large-scale testing and thinking capabilities. Offers next-gen multimodal capabilities, fast inference, and a massive 1M-token context window [Google Cloud 2026].
- **Anthropic Claude 4.5 Haiku⁴**: an advanced and the fastest model featuring major upgrades in instruction-following, used for real-time, low-latency tasks [Anthropic 2025].
- **Gemma 3 4B⁵**: a multimodal model highly customizable via open weights, and ideal for local deployment in reasoning, considered the state-of-the-art open model from [Google 2026].
- **Qwen 2.5 1.5B⁶**: known to be a versatile and multilingual open-source model with exceptional performance in structured data generation and long-context support up to 128K tokens [Qwen 2024].

They have facilitated support for the LangChain framework [Chase 2023], which provides consistent interfaces and abstractions across various model providers. For competence relation classification, the best practice is to configure all models with temperature settings equals to zero. This configuration prioritizes consistency and reliability in the evaluation process over the generation of creative or varied responses.

Regarding prompt design, the system message was structured by assuming a role, understanding the labels and the expected output format guided by [Li et al. 2023], and it appears below:

You are a very strict AI evaluator specializing in classifying the type of the relationship between two academic competencies. Classify the relationship using exactly one of these labels:

buildsTowards_A_to_B: Competence A provides prerequisite knowledge or conceptual foundation for competence B.

buildsTowards_B_to_A: Competence B provides prerequisite knowledge or conceptual foundation for competence A.

²<https://platform.openai.com/docs/models>

³<https://cloud.google.com/vertex-ai/generative-ai/docs/models>

⁴<https://platform.claude.com/docs/en/about-claude/models/overview>

⁵<https://ollama.com/library/gemma3:4b>

⁶<https://ollama.com/library/qwen2.5>

```

relatesTo: The competencies are meaningfully conceptually related, but there is no
           clear prerequisite direction.
noRelation: There is no direct learning progression or meaningful conceptual
            relationship.

Return only valid JSON with this schema: {"label": "one of the allowed labels"}

```

The label descriptions are taken directly from the Learning Commons website. The user message presents each competence pair with a structured prompt that contains two categories of information. For each competence, the model receives its identifier code, subject area, grade level, mathematical domain, cluster, and full natural-language description. These fields are complemented by four precomputed structural comparison features derived from the pair:

- Same subject: whether both competencies belong to the same subject area
- Same domain: whether both competencies share the same mathematical domain
- Same cluster: whether both competencies fall within the same curriculum cluster
- Grade distance (B minus A): the signed difference in grade level between competence B and competence A, capturing the directional progression signal encoded in grade sequencing

3.4. Evaluation Metrics

Model performance was assessed using three complementary measures [Opitz 2024, Ocheja et al. 2024]. Accuracy measures the proportion of correctly classified instances, but can be misleading under class imbalance, a known limitation given that noRelation accounts for 49.8% of the dataset. Macro F1 computes the unweighted mean of per-class F1 scores, giving equal weight to all classes regardless of frequency; it serves as the primary indicator of performance on minority classes, particularly relatesTo. Cohen's Kappa (κ) measures agreement with ground truth corrected for chance, with values above 0.60 conventionally indicating substantial agreement. To address RQ2, per-class scores were calculated separately for each relation category, allowing direct comparison of model behaviour across different families. family models. Also, Confusion matrices were additionally reported for each model to make the distribution of misclassification patterns more observable.

4. Results

4.1. RQ1: Identification of surmise relations

Table 2 reports the classification performance of the five evaluated models on the learning progression relationship detection task. Claude Haiku 4.5 achieved the highest Macro F1 of 0.603, indicating the most balanced performance across all relation categories. Gemini 2.5 Flash achieved the highest accuracy (0.697) and Cohen's κ (0.572), reflecting strong overall agreement with ground truth. The two smaller, locally-run models, Gemma3 4b and Qwen2.5 1.5b, performed substantially below the commercial LLMs, with Macro F1 scores of 0.395 and 0.234 and κ values of 0.320 and 0.122, respectively. For Qwen2.5 1.5b in particular, these figures suggest that its accuracy of 0.521 is largely an artifact of class imbalance rather than genuine classification ability.

Table 2. Classification performance of LLMs on the learning progression relationship detection task ($n = 307$ pairs)

Model	Accuracy	Macro F1	Cohen's Kappa
Claude Haiku 4.5	0.691	0.603	0.558
Gemini 2.5 Flash	0.697	0.553	0.572
GPT-5 Mini	0.625	0.529	0.490
Qwen2.5 1.5b	0.521	0.234	0.122
Gemma3 4b	0.511	0.395	0.320

4.2. RQ2: Extraction of different relationship categories

Table 3 reports the per-class F1 scores for all five models. Three consistent patterns emerge across model families. Even Claude Haiku 4.5, the best-performing model on Macro F1, achieved an F1 of only 0.246 for *relatesTo*, compared to 0.802 for *noRelation*. Gemini 2.5 Flash and Qwen2.5 1.5b failed entirely on this class ($F1 = 0.000$), while GPT-5 Mini produced only marginal predictions ($F1 = 0.035$). This pattern is directly observable in the confusion matrices: for Gemini (Figure 1b) and GPT-5 (Figure 1c), the *relatesTo* column is almost entirely empty.

For models that do predict *buildsTowards* relationships, confusing the direction, predicting $BT(A \rightarrow B)$ when the ground truth is $BT(B \rightarrow A)$, or vice versa, constitutes the next most frequent error type. This is visible in Claude's confusion matrix, where off-diagonal cells within the *buildsTowards* block are non-zero. Gemini 2.5 Flash shows a similar pattern, while GPT-5 Mini distributes errors more broadly across both directional categories and *noRelation*.

Table 3. F1-score per class for each evaluated model ($n = 307$ pairs)

Class	Gemini 2.5 Flash	Claude Haiku 4.5	GPT-5 mini	Qwen2.5 1.5b	Gemma3 4b
$BT(A \rightarrow B)$	.751	.677	.718	.000	.180
$BT(B \rightarrow A)$	.729	.686	.743	.220	.355
<i>relatesTo</i>	.000	.246	.035	.000	.236
<i>noRelation</i>	.730	.802	.622	.715	.810
Macro F1	.553	.603	.529	.234	.395
Accuracy	.697	.691	.625	.521	.511

On *relatesTo*, only Claude Haiku 4.5 and Gemma3 4b produce non-zero F1 scores (0.246 and 0.236, respectively), while Gemini 2.5 Flash and GPT-5 Mini fail almost entirely on this class ($F1 = 0.000$ and 0.035). Qwen2.5 1.5b produces zero predictions for $BT(A \rightarrow B)$ and *relatesTo* altogether, with 97.4% of its *buildsTowards* instances classified as *noRelation*.

5. Discussion

The findings of this study indicate that large language models can partially approximate the expert reasoning required to identify pairwise relations between mathematical com-

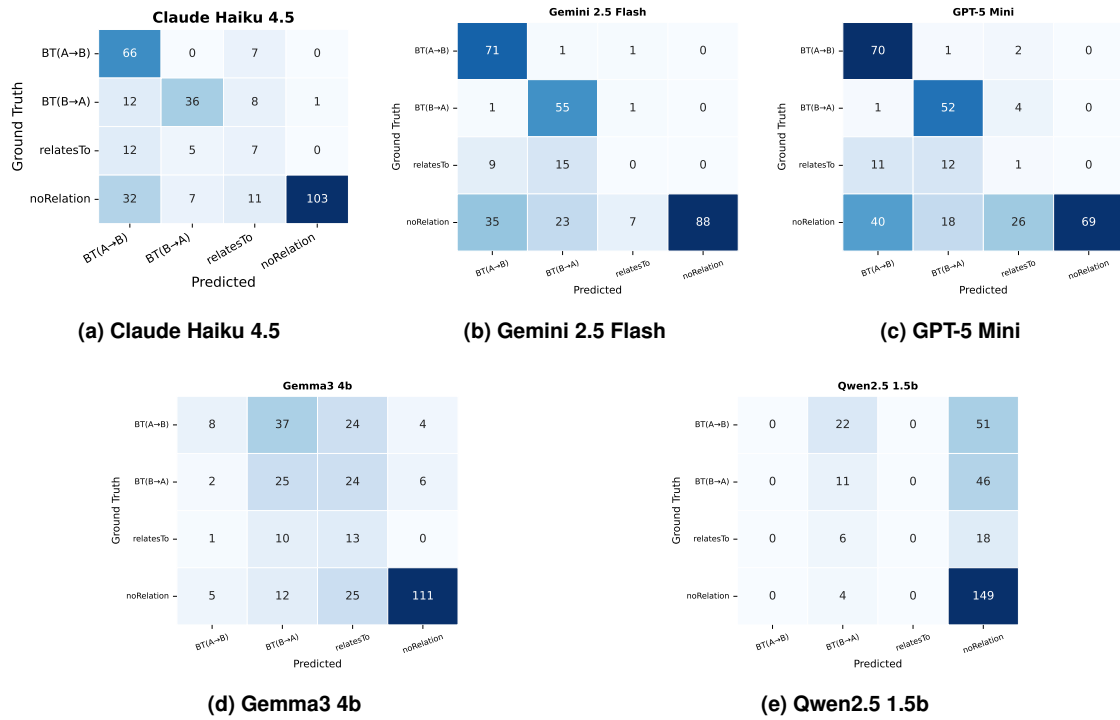


Figure 1. Confusion matrices for all five models

petencies, but that this capability is unevenly distributed and subject to clear structural limits. The performance of Claude Haiku 4.5 and Gemini 2.5 Flash, both achieving Cohen's κ above 0.55, indicates moderate to substantial agreement with expert-established ground truth, suggesting that zero-shot LLM prompting is a viable starting point for automated learning progression construction. This is a meaningful result given the task: identifying surmise relations among mathematical competencies has traditionally required teams of curriculum specialists working through structured elicitation procedures such as the QUERY algorithm [Koppen and Doignon 1990], whose quadratic scaling with the number of competencies makes exhaustive expert review logistically infeasible at scale [Cosyn et al. 2021]. That a general-purpose language model can approximate this reasoning without fine-tuning, domain-specific training data, or engineered features represents a qualitative shift in what automated tools can contribute to curriculum analysis.

At the same time, the results make clear that this capability is unevenly distributed across model families, reflecting a difference in the kind of reasoning each model brings to the task. GPT-5 Mini, Gemma3 4b, and Qwen2.5 1.5b are statistically indistinguishable from one another despite differences in scale and architecture, and their performance profiles suggest majority-class exploitation rather than genuine semantic discrimination, a conclusion supported by near-zero κ values for the two smaller open-source models. This pattern is consistent with findings from [Le and Abel 2025], who similarly observed substantial variation across model families in zero-shot prerequisite skills prediction, and with the broader observation by [Singh et al. 2024] that model family moderates LLM performance on educational tasks more than parameter count alone.

The consistent failure across all models on the *relatesTo* category reflects a deeper challenge that performance metrics alone do not fully capture. This class occupies a

contested space in curriculum design: It asserts conceptual proximity without hierarchical dependency, a distinction that is partially underdetermined by the structural features available in the prompt and that may carry genuine ambiguity even for human annotators [Son and Park 2025]. The structural features provided encode progression signals well for *buildsTowards* relations, where grade distance and cluster membership are informative, but they offer limited purchase on bidirectional conceptual overlap. This aligns with the observation by [Zhang et al. 2025] that global knowledge relation optimization benefits from document- and concept-level signals beyond structural metadata alone.

The finding that Gemma3 4b matches Claude Haiku 4.5 on *relatesTo* (F1 = 0.236 vs. 0.246) while performing substantially worse on *noRelation* points to a dissociation between the ability to detect ambiguous conceptual overlap and the ability to discriminate against unrelated competencies. This dissociation suggests that performance on minority classes is not straightforwardly explained by overall model capability, and that model type moderates class-specific behavior in ways that aggregate metrics do not capture.

The practical takeaway is that LLMs are best positioned not as replacements for curriculum experts, but as a first-pass filter within a human-in-the-loop architecture. Both ALEKS [Cosyn et al. 2021] and Dynamic Learning Maps [Thompson and Nash 2022] demonstrate that the validity of adaptive assessment depends directly on the quality of the underlying competence structure. A model that handles unambiguous cases automatically while flagging borderline decisions for human review could make large-scale curriculum graph construction feasible where exhaustive manual review is not — provided that model confidence is well-calibrated enough to support reliable triage, a question that future work should address directly.

6. Conclusion

The findings of this study demonstrate that large language models can partially automate a task that has traditionally required substantial expert effort, yet this capability is unevenly distributed across model families and subject to clear structural limits that overall performance metrics alone do not capture.

With respect to RQ1, findings demonstrate that top-tier commercial LLMs can partially approximate expert reasoning in this task. Claude Haiku 4.5 reached agreement approaching the substantial threshold with ground truth labels, suggesting that zero-shot LLM prompting is a viable starting point for automated learning progression construction.

Regarding RQ2, the *relatesTo* category proved systematically difficult across all models, including the best-performing one. This failure is not incidental: it reflects a deeper semantic challenge posed by bidirectional, non-hierarchical relations that structural features alone cannot resolve. Addressing this limitation likely requires richer input representations or a reconceptualization of the classification schema itself.

6.1. Limitations and future work

Several limitations should be acknowledged when interpreting these findings. First, the evaluation was conducted exclusively on mathematics competencies from the Common Core State Standards framework, which limits the generalizability of results to other subject domains, curricula, or educational contexts. Although individual standards can reasonably be operationalized as assessable competencies, they are not necessarily equiva-

lent to fine-grained competency statements, and differences in granularity and explicitness may therefore have influenced the identification of prerequisite relations. Second, all models were evaluated in a zero-shot setting with a fixed prompt structure, so performance may differ substantially under other prompt approaches or fine-tuned conditions; the results reported here should therefore be interpreted as a baseline rather than an upper bound on LLM capability for this task.

A further limitation concerns sample size and class imbalance: the test set comprises only 307 pairs, of which just 24 (7.8%) belong to the `relatesTo` class. This severe imbalance limits the robustness of per-class conclusions, particularly for `relatesTo`, where small sample sizes inflate the variance of F1 and κ estimates, and results for this class should therefore be regarded as exploratory rather than conclusive. Negative (`noRelation`) pairs were also sampled at random from the full set of non-related competency pairs, a procedure that may have favored comparisons between competencies from clearly distinct domains and potentially made the `noRelation` class easier to classify than it would be under a more adversarial sampling scheme restricted to same-domain or same-cluster pairs. Finally, commercial models were accessed via API at a specific point in time; as these models are updated continuously by their providers, results may not replicate exactly under future model versions, and the relative ranking of models observed here may shift over time.

Future research should explore prompting strategies, such as few-shot and chain-of-thought, to evaluate whether explicit reasoning guidance improves performance on ambiguous relation types. Second, richer input representations, such as semantic embeddings of competency descriptions or co-occurrence patterns in curriculum documents, may provide the signal needed to resolve `relatesTo` classification. Finally, calibration analysis should be conducted to assess whether model confidence scores are reliable enough to support automated triage between high- and low-certainty predictions in a practical pipeline.

Statement on the Use of Artificial Intelligence

We acknowledge the use of Generative AI to help in drafting and revising this paper, which was revised accordingly, and take full responsibility for its content.

Ethical Concerns

This study relies on a previous collected / published dataset. Accordingly, it is not eligible for assessment by an Institutional Review Board.

Artifacts Availability

The study's artifacts is available in this link: <https://github.com/jamillalobo/learning-progressions>.

References

- Anselmi, P., de Chiusole, D., and Heller, J. (2024). Adaptive assessment. In Heller, J. and Stefanutti, L., editors, *Knowledge Structures: Recent Developments in Theory and Application*, volume 7 of *Advanced Series on Mathematical Psychology*, pages 159–180. World Scientific.

- Anthropic (2025). Introducing Claude Haiku 4.5. <https://www.anthropic.com/news/claude-haiku-4-5>. Accessed: June 1, 2026.
- Bez, S., Burkart, F., Tomasik, M. J., and Merk, S. (2025). How do teachers process technology-based formative assessment results in their daily practice? results from process mining of think-aloud data. *Learning and Instruction*, 97:102100.
- Chase, H. (2023). Langchain. <https://github.com/langchain-ai/langchain>.
- Cosyn, E. and Thiéry, N. (2000). A practical procedure to build a knowledge structure. *Journal of Mathematical Psychology*, 44(3):383–407.
- Cosyn, E., Uzun, H., Doble, C., and Matayoshi, J. (2021). A practical perspective on knowledge space theory: ALEKS and its data. *Journal of Mathematical Psychology*, 101:102512.
- Daro, P., Mosher, F. A., and Corcoran, T. B. (2011). Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction.
- de Ayala, R. J. (2022). *The Theory and Practice of Item Response Theory*. Guilford Press, New York, 2nd edition.
- de Chiusole, D., Spoto, A., and Stefanutti, L. (2024). Innovative methods for building knowledge structures. In Heller, J. and Stefanutti, L., editors, *Knowledge Structures: Recent Developments in Theory and Application*, volume 7 of *Advanced Series on Mathematical Psychology*, pages 87–104. World Scientific.
- Doignon, J.-P. and Falmagne, J.-C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, 23(2):175–196.
- Google (2026). Gemma 3 4b instruction-tuned (Gemma-3-4b-it) model repository. <https://huggingface.co/google/gemma-3-4b-it>. Accessed: June 1, 2026.
- Google Cloud (2026). Gemini 2.5 flash model documentation. <https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/2-5-flash>. Accessed: June 1, 2026.
- Huff, K. and Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In Leighton, J. P. and Gierl, M. J., editors, *Cognitive Diagnostic Assessment for Education: Theory and Applications*, pages 19–60. Cambridge University Press, Cambridge.
- Ihichr, A., Oustous, O., El Idrissi, Y. E. B., and Ait Lahcen, A. (2024). A systematic review on assessment in adaptive learning: Theories, algorithms and techniques. *International Journal of Advanced Computer Science and Applications*, 15(7).
- Kambouri, M., Koppen, M., Villano, M., and Falmagne, J.-C. (1994). Knowledge assessment: tapping human expertise by the QUERY routine. *International Journal of Human-Computer Studies*, 40(1):119–151.
- Käser, T., Klingler, S., Schwing, A. G., and Gross, M. (2017). Dynamic Bayesian networks for student modeling. *IEEE Transactions on Learning Technologies*, 10(4):450–462.

- Koppen, M. and Doignon, J.-P. (1990). How to build a knowledge space by querying an expert. *Journal of Mathematical Psychology*, 34(3):311–331.
- Le, N. L. and Abel, M.-H. (2025). How well do LLMs predict prerequisite skills? Zero-Shot comparison to expert-defined concepts. In *Proceedings of the 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 7245–7250.
- Li, G., Wang, P., and Ke, W. (2023). Revisiting large language models as zero-shot relation extractors. pages 6877–6892.
- NCTM, À. (2000). National council of teachers of mathematics (2000) principles and standards for school mathematics. *Council of Teachers of Mathematics*.
- Ocheja, P., Flanagan, B., Dai, Y., and Ogata, H. (2024). How good is chatgpt in giving adaptive guidance using knowledge graphs in e-learning environments? *arXiv preprint arXiv:2412.03856*.
- of Chief State School Officers, C. (2010). Common core state standards for mathematics. Accessed: 2026-06-01.
- OpenAI (2026). Gpt-5-mini model documentation. <https://developers.openai.com/api/docs/models/gpt-5-mini>. Accessed: June 1, 2026.
- Opitz, J. (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *transactions of the association for computational linguistics*, 12:820–836.
- Pan, L., Li, C., Li, J., and Tang, J. (2017). Prerequisite relation learning for concepts in MOOCs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1447–1456, Vancouver, Canada. Association for Computational Linguistics.
- Qwen (2024). Qwen2.5-1.5B model repository. <https://huggingface.co/Qwen/Qwen2.5-1.5B>. Accessed: June 1, 2026.
- Roy, S., Madhyastha, M., Lawrence, S., and Rajan, V. (2019). Inferring concept prerequisite relations from online educational resources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9589–9594.
- Sargin, A. and Ünlü, A. (2009). Inductive item tree analysis: Corrections, improvements, and comparisons. *Mathematical Social Sciences*, 58(3):376–392.
- Segedinac, M. T., Horvat, S., Rodić, D. D., Rončević, T. N., and Savić, G. (2018). Using knowledge space theory to compare expected and real knowledge spaces in learning stoichiometry. *Chemistry Education Research and Practice*, 19(3):670–680.
- Singh, A., Tan, D., Hepworth, C., and Seeland, M. (2024). Comparing LLM responses for education across model families. *arXiv preprint arXiv:2502.08450*.
- Son, M. and Park, Y. (2025). Difficulties and solutions experienced by content experts in the process of developing a science curriculum ontology. *Brain, Digital, Learning*.
- Templin, J. and Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, 79(2):317–339.

- Thompson, W. J. and Clark, A. K. (2024). Improving instructional decision-making using diagnostic classification models. *Educational Measurement: Issues and Practice*.
- Thompson, W. J. and Nash, B. (2022). A diagnostic framework for the empirical evaluation of learning maps. *Frontiers in Education*, 6:714736.
- Ünlü, A. and Schrepp, M. (2021). Generalized inductive item tree analysis. *Journal of Mathematical Psychology*, 103:102547.
- Wu, X., Sun, S., Xu, T., and Wang, A. (2024). Research on the selection of cognitive diagnosis model from the perspective of experts. *Current Psychology*, 43:13802–13810.
- Zhang, M., Wang, J., Xiao, K., Wang, S., Zhang, Y., Chen, H., and Li, Z. (2025). Learning concept prerequisite relation via global knowledge relation optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1638–1646.