Assessing the Ability of Large Language Models to Give Learning Recommendations with Knowledge Space Theory

Peter Steiner*, Jan Hochweber

St.Gallen University of Teacher Education, Notkerstrasse 27, 9000 St.Gallen, Switzerland

Abstract

Generating personalized learning recommendations for next feasible learning steps represents a central challenge in formative assessment. While Large Language Models (LLMs) offer promising opportunities for generating such recommendations, statistically valid and theoretically grounded methods for evaluating their quality remain lacking. Therefore, this contribution proposes Knowledge Space Theory (KST) as a psychometric framework to benchmark LLM-generated next-step recommendations and addresses two questions: (1) Can LLMs reliably identify next feasible learning steps, and (2) do different LLMs systematically differ in this ability?

In a pilot study, two domains of linear functions with well-defined prerequisite structures were modelled as KST knowledge spaces. For 1,000 simulated students per domain, KST-derived outer fringes (sets of next feasible problems) were contrasted with recommendations from six LLMs of three providers.

Results reveal systematic deviations of all models from KST-based recommendations, with relative errors ranging from 18% to 47%. Friedman tests showed statistically significant performance differences between models on different distance metrics (Kendall's W = 0.25–0.37). Pairwise Wilcoxon tests demonstrated that smaller models perform significantly worse than their larger counterparts, with Gemini 2.5 Pro consistently achieving the best performance. Perfect-match rates ranged from 10% to 54% depending on the domain. The study demonstrates that KST provides a suitable framework for theory-driven evaluation of LLM-generated learning recommendations, but reveals that current LLMs remain insufficient for reliable formative feedback.

Keywords

Formative Assessment, AI in Education, Knowledge Space Theory, Educational Data Science

CEUR-WS.org/Vol-4330/paper3.pdf

1. Motivation

Formative feedback plays a critical role in supporting student learning by providing information that guides learners through their educational journey. A fundamental aspect of effective formative feedback is offering actionable recommendations about the next feasible learning steps tailored to each learner's current knowledge state. However, teachers, constrained by time and class size, often cannot consider the individual prerequisite knowledge each student brings, making personalized learning recommendations a challenging task.

Traditional digital learning systems, such as intelligent tutoring systems, have addressed this challenge by relying on rule-based structures and predefined learning pathways to suggest subsequent learning activities. While these systems have demonstrated effectiveness, their rigid frameworks limit their ability to accommodate the diverse and evolving learning needs of individual students. The recent advancements in generative artificial intelligence, particularly Large Language Models (LLMs), herald new opportunities for dynamic and context-aware generation of personalized learning recommendations and formative feedback.

Despite this potential, a significant gap remains in the availability of statistically valid and theoretically grounded methods to benchmark and validate AI-driven learning recommendation systems that focus specifically on the validity of recommended learning steps. In this contribution, knowledge space theory (KST, [1]) is proposed for this purpose. To investigate the effectiveness of this statistical framework, a

LLM Psychometrics: Applying Assessment Theories to Evaluate LLMs and LLM-Supported Learners, LAK 2026 Workshop
*Corresponding author.
peter.steiner@phsg.ch (P. Steiner); jan.hochweber@phsg.ch (J. Hochweber)
0009-0001-5491-9464 (P. Steiner); 0000-0002-7855-2097 (J. Hochweber)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-15

pilot study was conducted focusing on (a) the accuracy with which LLMs can identify next feasible learning steps and (b) performance differences across models.

2. Theory

KST is a mathematical framework designed for efficient knowledge assessment and representation, with a primary focus on generating descriptive and formative assessment rather than calculating normative student scores [1]. It is mathematically closely related to Cognitive Diagnostic Models (CDM, [2, 3]). Furthermore, KST has been identified as one of the most efficient frameworks for adaptive assessment systems [4] and used commercially in the American learning platform ALEKS [5].

Assessment in KST is conducted on a Domain Q, which is a (nonempty) set of problems within a specific field of knowledge (e.g., algebra). The knowledge state K of a student is the subset of the problems in Q that s/he has mastered. The collection of all practically feasible knowledge states $\mathscr{K}$ is called the knowledge structure. It is formally denoted as the pair $(Q, \mathscr{K})$, where $\mathscr{K}$ is a subset of the power set of Q, which minimally contains the empty set (student has not mastered any problems) and Q itself (student has mastered all problems). Whether a knowledge state is feasible or not is determined by a relation defined on the Domain Q. While multiple of these relations and their properties have been discussed [6, 7, 8], the surmise relation ($\preceq$) is the most widely used. For problems $a, b \in Q$, the relation $a \preceq b$ holds, if mastering a is a prerequisite for mastering b. This relation is reflexive and transitive, and therefore implies a quasi-order on Q. A knowledge structure $(Q, \mathscr{K})$ is called a knowledge space when $\mathscr{K}$ is closed under union. Since this assumption reflects the reality (if a student can master two sets of problems independently, s/he can also master all the problems), knowledge spaces are prevalently applied in student assessment.

The most important feature of KST for this contribution is that students' next feasible learning steps can be directly derived from the quasi-order on the domain Q. The set of all feasible learning steps based on a given learning state K is defined as the outer fringe of K:

$$\text{Outer}(K) = \{\, a \in Q \setminus K \mid K \cup \{a\} \in \mathscr{K} \,\}.$$

Therefore, $a \in Q$ is a next feasible learning step for a student with the learning state K, if s/he has not already mastered a ($a \in Q \setminus K$) and K with a is a learning state (i.e., all problems $b \in Q$ for which the relation $b \preceq a$ holds are already in K).

Summarizing, KST provides a well-tested statistical framework that allows to determine next feasible learning steps given a student's learning state within a valid knowledge structure. As a basis to discuss its applicability to evaluate LLMs' ability to propose suitable next learning steps, a pilot study was conducted to answer the following research questions:

- Can LLMs reliably identify next feasible learning steps given a student's learning state?
- Which model performs best at identifying next feasible learning steps given a student's learning state?

3. Study Design

At first, two Domains Q_1 and Q_2 were identified. Both are thematically homogenous subsets from the domain of linear functions, taken from a collection of 124 problems developed in the SaTiM project (Grünig et al., 2025). Domain Q_1 pertains to linear modelling, Q_2 pertains to identifying the slope from a graph of a linear function. Both domains are provided in the supplementary material to this contribution (https://doi.org/10.5281/zenodo.17818594). These two domains have been chosen specifically because there is a large proportion of theoretically obvious surmise relations among the problems (e.g., a student needs to be able to draw a slope triangle to the graph of a linear function before being able to identify the slope from a given graph of a linear function). Hence, the next feasible learning steps derived from the

corresponding knowledge structures may reasonably be used as a basis to evaluate recommendations of next learning steps generated (in the present context) by LLMs.

For this study, the order of the domains was identified by queries of domain experts, using the QUERY Algorithm implemented in the Prerex Tool [9, 10] and a guided consensus procedure. From the identified orders, the knowledge spaces of the domains were derived. Subsequently, student response data was simulated for both spaces with the kmsimulate function from kstMatrix package [11]. To obtain the data used for the reported analyses, the knowledge states and the outer fringes for each simulated student were calculated (see [12], for more detailed information about the assessment procedure). The resulting dataset that contains the students' response patterns and their outer fringes is the reference against which the LLMs will be compared.

For this comparison, six well-established LLMs from three different providers (gpt-4o, gpt-4o-mini, claude-sonnet-4-5, claude-3-haiku, gemini-2.5-pro, gemini-flash) were selected. Structured output was obtained by employing the instructor package [13] in python. The same naive prompt was used for both domains, in order to replicate the typical teacher and student usage of LLMs. The prompt was separated into a) a system prompt: "You are provided with a catalogue of problem types (K1–K7) with descriptions. Your task is to recommend problem types for a student to master next. Rules: Only choose the problem types that can be learned immediately (only based on the problem types that the student has already mastered). Catalogue: [Problemcatalogue]" and b) a user prompt: "The student has already mastered: [list of mastered problem types].Which problem types should the student try to master now?". To evaluate the ability of an LLM to recommend next feasible learning steps, the LLM recommendations were compared with the outer fringes identified through the KST analysis. For naive comparison, the difference of the two sets was identified by calculating the absolute value of the symmetric difference [14]

$$D(A, B) = |(A \setminus B) \cup (B \setminus A)|.$$

It must be noted that because the LLMs are strictly instructed not to output elements that are already in the knowledge state K as well as to only output elements from Q, the choice is limited to $Q \setminus K$. Therefore, the maximum error can be calculated by $|Q \setminus K| = |Q| - |K|$, since K is a subset of Q. Consequently, we can norm the distance of two sets A, B with regard to a knowledge state K and Domain Q by dividing by the maximum error:

$$D_N(A, B) = \frac{D(A, B)}{|Q| - |K|} = \frac{|(A \setminus B) \cup (B \setminus A)|}{|Q| - |K|}.$$

Another index for normed set differences is the Jaccard difference [14], which norms the symmetric set difference by the cardinality of the union of the two sets. Applying it to the data at hand, only the cardinality of the LLM-generated and the KST-generated outer fringe sets are considered. Since this comparison can also be understood as a precision–recall classification problem, we additionally computed the F1-score as a balanced measure of predictive performance [15].

4. Method

To evaluate whether LLMs can reliably identify next feasible learning steps, we first examined the descriptive statistics for the three continuous accuracy metrics evaluated in this study (relative distance, Jaccard distance, and F1-score). In addition, we also considered the perfect-match rate, a binary measure denoting whether a model's predicted outer fringe exactly matched the KST outer fringe.

For each of the $1,000$ simulated students, accuracy values were computed separately for each of the six LLMs across all four metrics. Subsequently, mean values, standard deviations, standard errors, and 95% confidence intervals were derived for every model–metric combination. To evaluate whether LLMs differ in their ability to correctly identify next feasible learning steps, we conducted a set of non-parametric statistical procedures. Prior to the main analyses, we assessed the distributional properties of the evaluation metrics (relative distance, Jaccard distance, F1 score, perfect-match rate) using the Shapiro–Wilk test for each of the models. All tests indicated significant departures from normality (all $p < .001$), which—given the large sample size—suggests systematic non-normality rather than

sample artefacts. Because each student was evaluated by every LLM, the resulting data structure is within-subjects (repeated measures), making non-parametric rank-based tests the appropriate choice.

Since the results are not normally distributed, the overall differences between models were tested by employing the Friedman test [16], the non-parametric equivalent of a repeated-measures ANOVA. For each metric, the Friedman test evaluates whether the median ranks differ across the six LLMs. To quantify the practical relevance of these differences, we calculated Kendall's W as an effect size measure. This coefficient reflects the degree of concordance between model rankings and ranges from 0 (no agreement) to 1 (perfect agreement). In line with common conventions, values around 0.10 can be interpreted as small, 0.30 as moderate, and 0.50 as large effect sizes.

Where the Friedman test was significant, we performed pairwise Wilcoxon signed-rank tests [17] to identify which models differed from one another. To control for the increased risk of a type I error when making multiple statistical comparisons, p-values were adjusted using the Holm correction [18]. For pairwise comparisons, we additionally report effect sizes (r), calculated as $Z/\sqrt{N}$, allowing an interpretation analogous to correlation magnitudes ($r \approx .10$ small, .30 medium, .50 large; Cohen, 1988).

To obtain initial insights into the generalizability of our results, we ran all analyses for both Domains Q_1 and Q_2 and compared the results.

5. Results

For the relative distance and Jaccard distance, which quantify how far an LLM's predicted outer fringe deviates from the KST-derived fringe, all models show non-negligible errors for both domains, with mean values for the relative distance ranging from 18 to 47% (see Figure 1). This suggests that LLM learning recommendations diverged from the next steps suggested by KST for at least 18% of the students. The F1 scores likewise reveal only moderate agreement, indicating that overlap between predicted fringes and the fringes derived from expert judgment is limited. Furthermore, the perfect-match rate, the strictest criterion, shows that exact matches occur in only about 20 to 35% of cases, depending on the model for Q_1 and about 10 to 54% for Q_2.

Taken together, these results suggest that even with a naive prompt, LLMs occasionally approximate the correct next feasible learning steps. Nevertheless, their reliability in this setting can be seen as insufficient to provide teachers and students with actionable formative feedback.

In both domains and across the three difference metrics (relative distance, Jaccard distance, and F1 score), the Friedman tests revealed statistical significant differences—with small to medium effect sizes—between different LLMs' capability to recommend next feasible learning steps in line with results from KST (see Table 1). The perfect-match rate exhibited a substantial difference in effect size between the two domains, indicating domain-specific variation in how strictly models can reproduce the KST-derived outer fringe.

Since the Friedman tests were statistically significant across all metrics and both domains, Wilcoxon signed-rank tests were conducted for all pairwise model comparisons within each domain. While the

Table 1
Friedman Test Results and Effect Sizes Across Models for All Metrics

Metric	Domain	χ^2	df	p	Kendall's W
Relative distance	D1	1432.89	5	$< .0001$	0.287
Relative distance	D2	1837.60	5	$< .0001$	0.368
Jaccard distance	D1	1232.02	5	$< .0001$	0.246
Jaccard distance	D2	1307.71	5	$< .0001$	0.262
F1 score	D1	1306.69	5	$< .0001$	0.261
F1 score	D2	1307.71	5	$< .0001$	0.262
Perfect match	D1	368.42	5	$< .0001$	0.074
Perfect match	D2	1353.05	5	$< .0001$	0.271

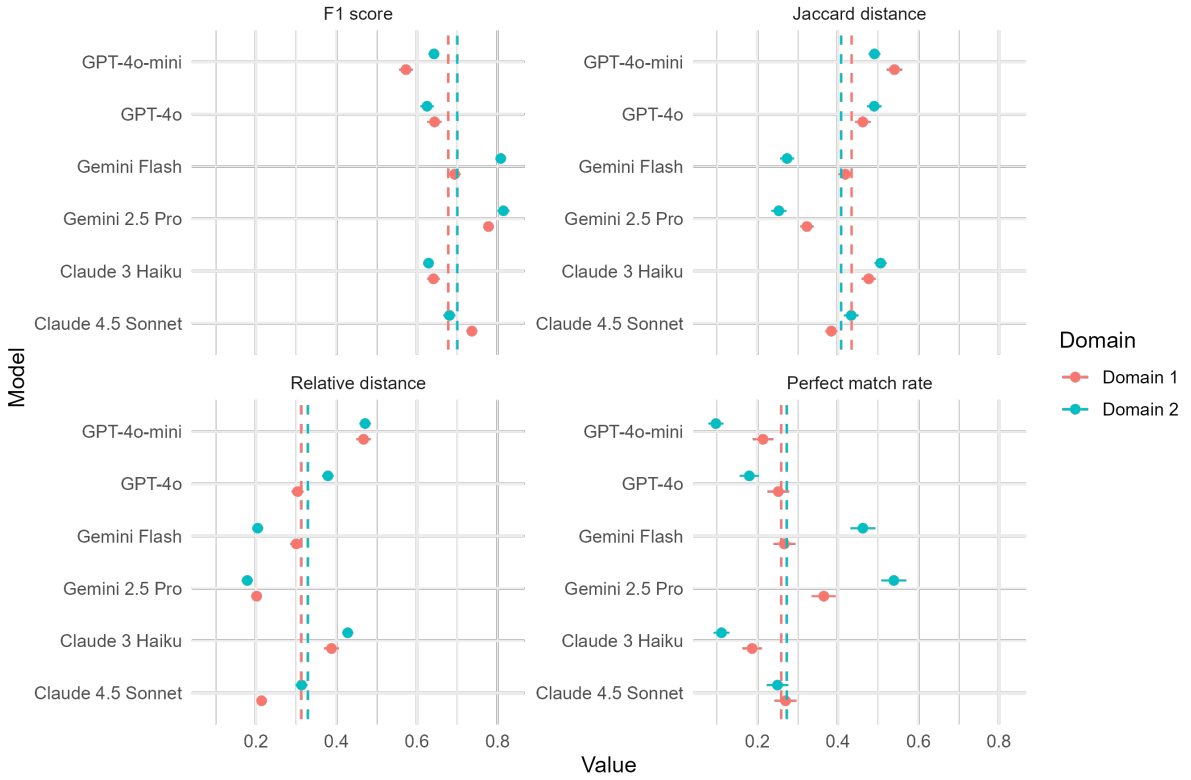


Figure 1: Model comparison across the four different metrics and the two domains. Depicted are the mean values per model with 95% confidence intervals. The dashed lines show the means across all models.

complete dataset is available in the supplementary material (https://doi.org/10.5281/zenodo.17818594), 2 presents an illustrative example for Q_1 based on the relative distance metric. As shown, all smaller models (Gemini Flash, GPT4o-mini, Claude 3 Haiku) perform significantly worse—typically with medium-sized effects—than their larger counterparts within the same provider family. Because this pattern is theoretically expected, it also strengthens the validity of the proposed approach to evaluate LLMs' performance. Interestingly, in this example, no significant difference was observed between Gemini 2.5 Pro and Claude 4.5 Sonnet. The largest effect in the illustrated example ($r = 0.70$) occurs between GPT4o-mini and Gemini 2.5 Pro, and this contrast is even stronger in Q_2 ($r = 0.75$).

The pattern of results—considered together with the descriptive means—suggests an approximate ranking of model performance (see Figure 3). However, this ranking must be interpreted with caution, as not all pairwise differences contributing to this order reached statistical significance (e.g., Gemini 2.5 Pro and Claude 4.5 Sonnet for the relative distance in Q_1). Across both domains, the Gemini 2.5 Pro model consistently outperformed all other models on all metrics. In contrast, GPT-4o, GPT-4o-mini, and Claude 3 Haiku exhibited consistently lower performance.

6. Limitations

Although the present findings offer important insights, they should be interpreted in light of several limitations. First, all data used in this study were obtained in a highly controlled and somewhat artificial setting that relied on a naïve prompt. Consequently, conclusions regarding the first research question, in particular, should be interpreted with caution, as prompt design most likely has an influence on models' performance independently of their actual reasoning capabilities. Since the implementation of the metrics developed for this contribution is novel, the results cannot yet be interpreted in relation to existing findings. At present, no established benchmarks, reference distributions, or external validation studies exist for these metrics, which limits the extent to which the reported statistics can be

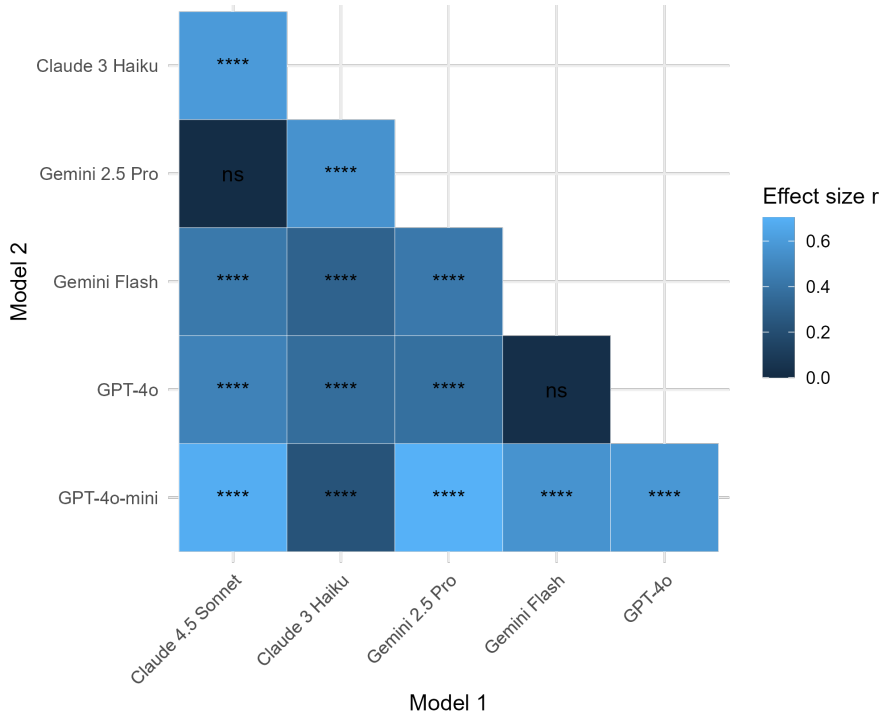


Figure 2: Heatmap of pairwise Wilcoxon signed-rank comparisons between all LLMs. Each cell displays the Holm adjusted significance level of the pairwise test (ns = not significant; * p < .05; ** p < .01; *** p < .001; **** p < .0001). The colour intensity represents the corresponding effect size *r* with darker shades indicating smaller effects.

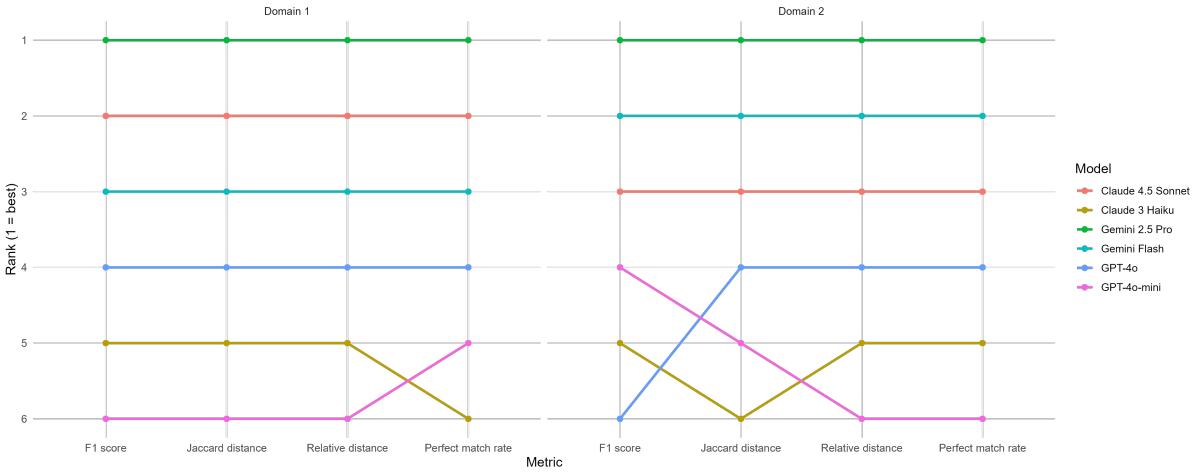


Figure 3: Rankings of all six LLMs across the four accuracy metrics for both domains. Lower values indicate better ranks (1 = best). Each line represents one model's relative position compared to the others. The left panel shows rankings for Domain 1 and the right panel for Domain 2, illustrating how model performance patterns differ across metrics and domains.

contextualized. As a consequence, it remains unclear whether the LLMs' observed performance should be considered strong, weak, or rather typical when compared to other domains, tasks, or diagnostic frameworks.

The present study focused exclusively on three major LLM providers, meaning that other recently emerging high-performing systems—such as Grok or DeepSeek—were not included in the evaluation. Moreover, at the time of data collection, no API integration for GPT-5 was available, which prevented its inclusion in the comparative analyses. As a result, the ranking observed in this study—where Gemini 2.5 Pro achieved the highest overall performance—should not be generalized to the full landscape of

currently available LLMs. Future research should expand the comparison to include these additional models, in order to obtain a more comprehensive and up-to-date assessment of LLM capabilities.

7. Further Research

The present pilot study provides an initial demonstration that KST can serve as a theoretically grounded and statistically rigorous framework for evaluating the quality of LLM-generated learning recommendations. However, several avenues for further research emerge from the findings and limitations of this contribution.

First, the analyses should be extended beyond the two domains investigated here. While Q_1 and Q_2 provide contrasting examples with clear prerequisite structures, it remains an open question whether the proposed evaluation approach generalises to broader or more heterogeneous domains. Consequently, future research should formally analyse cross-domain model behaviour to determine whether observed performance patterns and model rankings persist across different knowledge structures.

Furthermore, as discussed in the limitations, the evaluation pipeline should be expanded to incorporate a broader and more up-to-date set of LLMs. A comprehensive follow-up study should therefore reassess the ranking of models once these systems can be included. Such an extension will be essential for determining whether the top performance of Gemini 2.5 Pro observed in this study persists in the wider ecosystem of state-of-the-art models.

Finally, further research should also investigate alternative prompting conditions systematically. The present work deliberately employed a naïve prompt to mimic realistic teacher-and-student usage scenarios. However, it is crucial to examine the extent to which more sophisticated prompting strategies (e.g., chain-of-thought prompting) influence the alignment between LLM output and the psychometrically grounded KST outer fringes, in order to better understand the true potential of LLMs in this context.

Together, these directions open a pathway toward establishing a general, theory-driven framework for benchmarking AI-generated learning recommendations and integrating knowledge-based psychometric models with modern generative AI systems.

8. Conclusion

This pilot study demonstrates that KST provides a principled and statistically rigorous framework for evaluating the validity of LLM-generated learning recommendations. Across two well-defined domains, all investigated models showed systematic deviations from the psychometrically grounded outer fringes, with only moderate overlap and relatively low perfect-match rates. While the analyses revealed clear performance differences between models—with Gemini 2.5 Pro consistently outperforming the others—the overall reliability of LLM-generated next-step recommendations appears insufficient for formative feedback in their current form. Nonetheless, the results highlight the potential of combining KST with modern generative AI to build transparent, interpretable, and benchmarkable recommendation systems.

9. Acknowledgments

The work on this contribution was funded by the Swiss National Science Foundation (SNSF), grant number 100019_219884.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.1 in order to check grammar and spelling. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J.-P. Doignon, J.-C. Falmagne, Spaces for the assessment of knowledge, International Journal of Man-Machine Studies 23 (1985) 175–196. URL: https://linkinghub.elsevier.com/retrieve/pii/S0020737385800316. doi:10.1016/S0020-7373(85)80031-6.

[2] J. Heller, L. Stefanutti, P. Anselmi, E. Robusto, On the Link between Cognitive Diagnostic Models and Knowledge Space Theory, Psychometrika 80 (2015) 995–1019. URL: https://www.cambridge.org/core/product/identifier/S0033312300002349/type/journal_article. doi:10.1007/s11336-015-9457-x.

[3] S. Noventa, J. Heller, A. Kelava, Toward a unified perspective on assessment models, part I: Foundations of a framework, Journal of Mathematical Psychology 122 (2024) 102872. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022249624000415. doi:10.1016/j.jmp.2024.102872.

[4] A. Ihichr, O. Oustous, Y. E. B. E. Idrissi, A. A. Lahcen, A Systematic Review on Assessment in Adaptive Learning: Theories, Algorithms and Techniques, International Journal of Advanced Computer Science and Applications 15 (2024). URL: http://thesai.org/Publications/ViewPaper?Volume=15&Issue=7&Code=ijacsa&SerialNo=85. doi:10.14569/IJACSA.2024.0150785.

[5] E. Cosyn, H. Uzun, C. Doble, J. Matayoshi, A practical perspective on knowledge space theory: ALEKS and its data, Journal of Mathematical Psychology 101 (2021) 102512. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022249621000134. doi:10.1016/j.jmp.2021.102512.

[6] I. Düntsch, G. Gediga, Skills and knowledge structures†, British Journal of Mathematical and Statistical Psychology 48 (1995) 9–27. URL: https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/j.2044-8317.1995.tb01047.x. doi:10.1111/j.2044-8317.1995.tb01047.x.

[7] J.-C. Falmagne, J.-P. Doignon, Learning Spaces: Interdisciplinary Applied Mathematics, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. URL: https://link.springer.com/10.1007/978-3-642-01039-2. doi:10.1007/978-3-642-01039-2.

[8] M. Koppen, J.-P. Doignon, How to build a knowledge space by querying an expert, Journal of Mathematical Psychology 34 (1990) 311–331. URL: https://linkinghub.elsevier.com/retrieve/pii/0022249690900358. doi:10.1016/0022-2496(90)90035-8.

[9] M. Kambouri, M. Koppen, M. Villano, J.-C. Falmagne, Knowledge assessment: tapping human expertise by the QUERY routine, International Journal of Human-Computer Studies 40 (1994) 119–151. URL: https://linkinghub.elsevier.com/retrieve/pii/S1071581984710068. doi:10.1006/ijhc.1994.1006.

[10] P. Steiner, J. Hochweber, M. Kickmeier-Rust, S. Leininger, Prerex: Conducting expert queries for hierarchical competency structures automatically, in: National Council on Measurement in Education (NCME), Los Angeles, CA, United States, 2026. Conference presentation.

[11] C. Hockemeyer, P. Steiner, W. Wong, kstMatrix: Basic Functions in Knowledge Space Theory Using Matrix Representations, Department of Psychology, University of Graz, 2025. URL: https://CRAN.R-project.org/package=kstMatrix. doi:10.32614/CRAN.package.kstMatrix.

[12] P. Steiner, C. Hockemeyer, J. Hochweber, M. Kickmeier-Rust, Structuring and Assessing Knowledge with Knowledge Space Theory, in: M. L. Wilson, A. Tawfik, A. Ritzhaupt (Eds.), Quantitative Methods in Educational Research: Concepts and Applications, EdTech Books, 2026. Accepted book chapter.

[13] G. Mauro, Instructor: Structured outputs for llms, 2023. URL: https://github.com/instructor-ai/instructor, python package.

[14] J. C. Gower, P. Legendre, Metric and Euclidean properties of dissimilarity coefficients, Journal of Classification 3 (1986) 5–48. URL: http://link.springer.com/10.1007/BF01896809. doi:10.1007/BF01896809.

[15] C. J. v. Rijsbergen, Information retrieval, 2. ed., repr ed., Butterworth, London, 1981.

[16] M. Friedman, The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance, Journal of the American Statistical Association 32 (1937) 675–701. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522. doi:10.1080/01621459.1937.10503522.

[17] F. Wilcoxon, Individual Comparisons by Ranking Methods, Biometrics Bulletin 1 (1945) 80. URL: https://www.jstor.org/stable/10.2307/3001968?origin=crossref. doi:10.2307/3001968.

[18] S. Holm, A Simple Sequentially Rejective Multiple Test Procedure, Scandinavian Journal of Statistics 6 (1979) 65–70. URL: http://www.jstor.org/stable/4615733, publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley].