

Large Language Model as teachers' assistants in inquiry-based learning: analyzing student answers to rich mathematical problems

Shai Olsher, Michal Yerushalmy, & Gil Sher

University of Haifa

Abstract: Inquiry-Based Learning (IBL) in mathematics promotes active student engagement in solving complex problems and reflective thinking. Rich problems support multiple solution paths and require higher-order thinking. While digital tools such as models and simulations support IBL, the potential of large language models (LLMs) such as ChatGPT to analyze student responses in this context is still underexplored. This qualitative study examines how ChatGPT analyzes responses to a rich mathematical task solved by high-school students and compares them to expert educators' analyses, assessing alignment with IBL principles. The study aims to deepen understanding of how AI can support IBL by efficiently processing large volumes of student-generated data - especially written responses to rich problems - and reflecting key insights to teachers. Moreover, it may provide a foundation for using GenAI as part of innovative tools for mathematics education.

Keywords GenAI for teaching productivity, PBL Problem Based Learning, LLM assisted assessment, problem solving assessment rubrics, modeling task.

1.1 Introduction

Inquiry-based mathematics classrooms encourage students to explore, model, and reason through complex problems. However, assessing such open-ended work poses a significant challenge for teachers. The diversity of student approaches and representations makes it difficult to capture both the quality of reasoning and the depth of conceptual understanding. At the same time, generative AI (GenAI) tools such as large language models (LLMs) are becoming increasingly capable of solving complex mathematical problems, yet assessing inquiry involves qualities that go far beyond the correctness or efficiency of a solution. The focus of this study is therefore not on what an LLM can solve, but on what it can reveal about students' mathematical thinking.

To examine whether GenAI tools can assist in analyzing students' products in problem-based learning (PBL), it is first necessary to establish a shared standard for what constitutes a rich mathematical task and to describe how experts evaluate such work. Effective assessment of inquiry requires attention to multiple facets of learning, including reasoning, modeling, representation, and reflection. These must be made

explicit if an AI tool is to approximate expert judgment. In this study, we rely on an example of a rich mathematical modeling problem and apply the assessment model and rubrics developed by the Balanced Assessment Project (Schwartz et al., 1995) as a framework for expert evaluation.

The study focuses on the written work of twelve Israeli high-school students who solved a modeling task involving geometric reasoning, algebraic representation, and contextual interpretation. Their responses were analyzed by ChatGPT (version 5.0), and the AI-generated assessments were examined to address the following research questions:

1. How does a LLM analyze students' written responses to a rich mathematical problem according to IBL principles/characteristics?
2. What information given by the LLM tool could inform teachers' formative assessment practices, mainly identifying strengths, and elements that require improvement in students' work?

A central aim of this exploration is to identify whether, and in what ways, information produced by the LLM can support teachers' formative assessment in PBL contexts. In particular, we seek to understand how AI-based analysis might help teachers recognize patterns of reasoning, modeling, and inquiry across multiple student solutions, thereby offering efficient yet pedagogically meaningful insights into students' mathematical understanding.

1.2 Background

To guide large language models (LLMs) toward meaningful assessment of students' written work on rich mathematical problems, it is necessary to understand what expert teachers do when they interpret such work. Teachers do not simply evaluate correctness; they read student solutions as expressions of mathematical thinking and make sense of them through interpretive, dialogic, and reflective processes. While many reform programs and teaching experiments in mathematics education have addressed inquiry-

oriented assessment (e.g., Van den Heuvel-Panhuizen et al., 2020), few have articulated these interpretive practices in a manner that generalizes beyond specific tasks or curricula.

Lampert's *Teaching Problems and the Problems of Teaching* (Lampert 2001) offers one of the most detailed accounts of such practice. She documents how, when examining students' written and verbal activity, she continuously attends to what students do and say, revises her own knowledge of the domain, probes and revoices student ideas, analyzes whether mathematical focus is maintained, and locates individual performances within the broader landscape of the mathematical content she intends to teach. These actions exemplify how teachers engage in formative interpretation: a process of making sense of students' reasoning while preserving mathematical integrity.

Chazan and Ball (see Chazan, 2000) similarly emphasize that teaching through inquiry requires teachers to discern what information about student thinking is essential for providing meaningful scaffolding. Schnepf's reflective accounts in Schnepf and Chazan (Chazan & Schnepf, 2002) further refine this view through three interrelated teaching modes. In the exploratory mode, students investigate rich problems without prior algorithmic instruction while the teacher listens, encourages conjecture, and refrains from evaluating correctness. The hybrid mode involves guiding and structuring students' emerging ideas while maintaining their ownership of reasoning. Finally, in the formal mode, the teacher supports students in consolidating methods, connecting representations, and refining precision and efficiency. These modes outline how formative assessment evolves dynamically as inquiry progresses—a structure that can inform the design of LLM-based analysis.

The Balanced Assessment Project (Schwartz et al., 1995) developed rubrics and frameworks that similarly highlight process-oriented assessment. The project's Mathematical Concepts and Problem Solving (MCAPS) scoring system emphasizes reasoning and representation over correctness. Tasks are designed to have multiple entry points, inviting students to engage with mathematical objects (representations, diagrams, quantities) and actions (transforming, generalizing, justifying). Teachers are supported in using these rubrics to gather evidence of student reasoning and to make instructional decisions based on observed thinking rather than outcomes. This approach underscores the dual purpose of assessment as both a measure

and a formative learning opportunity, and provides a concrete model for how AI might analyze student work in ways that reflect inquiry principles.

Quality in mathematical writing is also central to LLM analysis and interpretation of students' written work. Teledahl et al. (2025) identify four key aspects of strong mathematical writing: defining assumptions and variables, constructing coherent narratives that include relevant calculations, using correct language and representations, and attending to contextual appropriateness. These criteria highlight the complex interplay of semantic and syntactic dimensions that expert teachers implicitly consider when interpreting students' written responses.

Taken together, these perspectives suggest that assessing inquiry requires attention to the processes students engage in, not just the products they produce. This perspective is aligned with the call to digitally assess authentic mathematical work of the students, not only via proxies that are easily assessed algorithmically such as multiple choice or closed answer questions (Olsher et al, 2024; Stacey and Wiliam, 2013). Translating such interpretive sophistication into LLM-based assessment therefore requires models trained to detect and describe reasoning, representation, and reflection in ways consistent with how expert teachers read student work.

Large language models (LLMs) such as ChatGPT inherently operate in tension with the pedagogical aims of inquiry- and problem-based learning (PBL). Whereas PBL emphasizes exploration, reasoning diversity, and the productive struggle involved in solving rich problems, LLMs are designed to provide efficient, correct, and convergent answers. From an educational standpoint, the goal is not to have an AI that tells students what the correct solution is or that evaluates them solely against a predefined procedure. Instead, the value of such models lies in their capacity to recognize and interpret how students reason, even when their processes diverge from canonical solution paths. In this regard, Applebaum (2025) demonstrated that LLMs can approach mathematical puzzles through multiple pathways and even generate alternative problem framings, illustrating a capacity for flexible reasoning rather than mere computation. This finding suggests that the same models capable of exploring diverse strategies might also be used to assess and illuminate students' exploratory thinking—an essential goal of inquiry-based mathematics education.

Recent studies highlight the growing potential of LLMs such as ChatGPT to assist teachers in interpreting and assessing students' mathematical thinking. Traditional assessment systems have focused primarily on correctness and procedural accuracy, whereas LLMs can process written, open-ended explanations that capture aspects of reasoning and reflection central to inquiry-based learning (IBL). The U.S. Department of Education (2023) positions artificial intelligence as a means to augment rather than replace teacher judgment, emphasizing its role in synthesizing large volumes of learner-generated data into actionable insights for formative assessment. Within mathematics education, this potential has been explored through studies that examine how ChatGPT can evaluate reasoning, identify misconceptions, and generate explanatory feedback to support both students and teachers.

Empirical work in this area demonstrates both promise and limitation. Jin and Suh (2024) investigated the use of ChatGPT 4.0 for automated scoring of descriptive mathematical responses in combinatorics. Using well-defined rubrics and prompt engineering aligned with specific assessment criteria, they found a high level of consistency between AI-generated scores and human ratings. This structured approach underscored the importance of rubric-based design for reliability in automated analysis. Similarly, Rigaud Téllez, Rayón Villela, and Blanco Bautista (2024) explored ChatGPT's capacity to generate and evaluate formative assessment tasks in linear algebra. Their findings indicated that while the model could produce and assess valid conceptual items, it sometimes simplified complex reasoning, pointing to the need for teacher moderation in interpretive tasks. Extending beyond automated scoring, Huda, Anwar, Nofverma, and Kurniawan (2025) demonstrated that ChatGPT-mediated dialogue in guided-inquiry calculus lessons enhanced students' metacognitive regulation. Through structured inquiry worksheets, they collected analyzable traces of reasoning that could inform teachers' formative assessment practices. Complementing these approaches, Moed Abu Raya and Olsher (2025) combined LLM-based text analysis with learning analytics to connect students' GeoGebra constructions and written justifications. Their design-based research highlighted how augmenting automated analysis with contextual text interpretation can reveal class-level reasoning patterns and support teachers in providing targeted, formative feedback.

Synthesizing these developments, Pepin, Buchholtz, and Salinas-Hernández (2025) identified assessment and feedback as a key theme in their scoping survey of ChatGPT research in mathematics education. They argue that LLMs' ability to generate explanatory discourse positions them as potential co-analysts of mathematical reasoning, provided that human interpretation and pedagogical framing remain central. Collectively, these studies signal a shift from using AI for automated grading to employing it as an analytic partner capable of surfacing patterns in students' reasoning and reflection. For IBL contexts, where assessment aims to illuminate diverse solution paths and depth of thought, ChatGPT offers a scalable yet teacher-anchored means of transforming student-generated data into formative insights while maintaining the interpretive essence of mathematics assessment.

Building on this line of research, the present study examines how ChatGPT analyzes high-school students' written responses to a rich mathematical task and how its interpretations align with those of expert educators trained in IBL. By comparing these analyses, the study explores how generative AI can support teachers in identifying patterns of reasoning and reflection across large sets of student work, thereby contributing to more responsive and insight-driven assessment practices.

1.3 Research questions

How does a LLM analyze students' written responses to a rich mathematical problem according to IBL principles/characteristics? And what information given by the LLM tool could inform teachers' formative assessment practices, mainly identifying strengths, and elements that require improvement in students' work?

1.4 Methods

This study employed a qualitative research design, aiming to explore how a LLM, in this case, ChatGPT 5.0, analyzes answers of high-school students in Israel to an open-ended mathematical inquiry task. The focus was on analyzing students' written solutions to an optimization problem ("Best bus route", see elaboration below) using problem-solving principles. Rather than measuring correctness alone, the analysis

emphasized the mathematical reasoning, representational choices, and inquiry-oriented strategies students employed.

1.4.1 The task

The inquiry task used in this study was chosen to reflect the type of problem-solving tasks we expect students to encounter in a dynamic educational environment where performing calculations and solving formulas are less the focus, but rather employing mathematical principles and ideas in different scenarios to solve problems. The problem was taken from a set of mathematical assessment problems developed by the Balanced Assessment team (Schwartz et al., 1995).

1.4.2 Matching model

1. You are asked to add a bus line that will serve four cities of similar size and population type. For each of the following requirements: A, B, and C (handle each requirement separately) plan the bus route. The cities are represented by points on the map below.

Decide which of the three routes is the best route.

Describe the criterion that makes this route the best.

Your criterion should be based on a quantitative measure that you devise, and that can be used to calculate how well a route fits for any given set of cities.

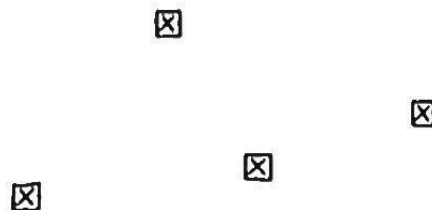


Figure 1. The cities in the “best bus route” problem, reproduced from Schwartz et al., 1995

A. The route must be straight and must pass through at least one of the cities.

B. The route must be straight.

C. The route must be circular.

2. Where could you place an additional city so that the chosen route would serve it in the best way?

Explain your answer.

1.4.3 Population

The participants in this study were 12 Israeli high school students from central Israel. 6 of the participants were 11th-grade students, and 6 were 12th-grade students. All participants learned an advanced-level mathematics course, towards either 4 or 5 point matriculation exam, at the end of their 12th grade. All students learned Algebra and functions, Euclidean geometry and analytic geometry. On the other hand, the Israeli high school curriculum does not include a specific emphasis on problem-solving or mathematical literacy.

1.4.4 Generating LLM assessment

The development and training of the prompt followed an iterative, design-based process that ended up consisting of seven stages. Each stage was guided by pedagogical theory, empirical testing on authentic student work, and progressive refinement based on observed outcomes.

Stage 1: Drafting the Initial Prompt Framework

An initial prompt framework was created, structured around three essential components: identifying the problem solution, analyzing representations and reasoning, and suggesting directions for improvement. Pedagogical principles inspired by inquiry-oriented instruction, such as promoting critical thinking, mathematical communication, and representational diversity, were embedded within the structure. Initial experiments with student solutions led to the removal of non-measurable expressions (e.g., “the student understands that a straight line is most efficient”), focusing the language exclusively on observable evidence (mathematical objects and descriptions of actions) within the student’s written work.

Stage 2: Establishing a Stable Evaluation Format

A consistent output format was developed to ensure comparability across responses. This format, based on the rubric-assessment of the Balance-Assessment format, included a checklist addressing problem requirements, an analysis along four dimensions (representations, quantitative reasoning, inference, and communication), a performance table (complete/partial/missing), identification of strengths and areas for improvement, and a teacher-oriented summary. This structure provided the foundation for all subsequent evaluations.

Stage 3: Alignment with Inquiry-Based Assessment Principles

The evaluation criteria were revised to emphasize the processes of exploration, justification, and reasoning demonstrated in students' responses. Additional components encouraged the identification of alternative approaches and boundary cases. A distinction was introduced between formulation (what the student wrote or did not write) and execution (what was actually done), supporting deeper analysis of students' inquiry trajectories.

Stage 4: Refining Language and Logical Assumptions

All prompt statements were carefully revised to exclude unverifiable assumptions about students' intentions or understanding. The LLM was instructed to base evaluations strictly on evidence present in the solution, to avoid attributing implicit knowledge, and to maintain a descriptive, professional tone rather than a judgmental one. For example, The initial feedback was subjective and judgmental: "poor work that is lacking reasoning". The revised version stated: "The student provided a graphic representation of three routes but did not include written reasoning. The drawing shows a single integrated layout rather than separate routes for each condition." This phrasing remains factual and professional, describing what appears in the student's work without inferring intent or ability. This refinement ensured that the assessment language remained both pedagogically responsible and analytically precise.

Stage 5: Iterative Testing and Final Consolidation

Multiple versions of the prompt were tested on authentic student solutions to examine its ability to distinguish levels of reasoning, identify atypical responses, and yield pedagogically relevant insights. Through repeated refinement, a final version was consolidated, integrating accumulated improvements into a stable, inquiry-aligned structure. A corresponding teacher evaluation form was then developed, maintaining the same four analytic dimensions while adapting the format for practical classroom and research use.

Stage 6: Validation

The refined prompt was applied to authentic student solutions to examine its validity and pedagogical utility. The prompt demonstrated sensitivity to diverse reasoning levels, the ability to identify cases where students deviated from the task, and the capacity to produce informed teacher recommendations for supporting inquiry-based learning and assessment practices.

Stage 7: Documentation and Implementation

The entire process was documented in a formal record that distinguished between the training phase (prompt development and refinement) and the research phase (application to authentic data). Only evaluations produced during the research phase were included in the study’s analysis. This distinction ensured methodological transparency and analytical rigor.

Progressive refinements shifted the focus to observable mathematical reasoning and evidence-based description. Several pivotal developments emerged: (1) The introduction of a category for responses unrelated to the task. (2) The creation of an explicit mathematical assessment section addressing formal objects, operations, models, accuracy, and inference. (3) The reframing of accuracy as a mathematical rather than graphical attribute. (4) The inclusion of explicit references to exploratory strategies, as described in Table 1.

Table 1: Inquiry-based learning categories assessed

Category	Description
----------	-------------

“What if not?” reasoning	Testing alternative or negated conditions to explore the robustness of a solution.
Extreme cases	Considering boundary or limiting scenarios to test the generality of a conclusion.
Generalization	Extending from a specific solution to a broader principle or rule.
Quantitative criteria	Defining and applying measurable optimization metrics (e.g., sum of distances, regression line).
Representation shifts	Moving between different representations, such as diagrams, verbal explanations, or algebraic expressions.
Mathematical modeling	Translating the problem into formal mathematical objects (e.g., lines, circles, coordinates) and applying relevant formulas or functions.
Justification and reasoning	Using argumentation, examples, or calculations to substantiate mathematical claims.

The final version of the prompt guided the evaluation process through six structured steps: (1) Checklist of task requirements: Did the student address each sub-task (constrained straight line, unconstrained straight line, circular route, fifth city)?; (2) Qualitative assessment: Examination of representations, quantitative criteria, reasoning, and communication; (3) Problem-solving rubric: Levels (incomplete / partial / advanced / full) across domains.; (4) Mathematical problem solving rubric: Explicit identification of mathematical objects (e.g., line, circle, perpendicular bisector, coordinate system) and operations (e.g.,

slope calculation, distance formula, regression, centroid formula); (5) Strengths and directions for improvement: Based on previous steps; and (6) Teacher’s summary: What the student’s solution reveals about their mathematical practices and how ensuing instruction can build on it.

1.4.5 Data Sources and Collection

The data set consisted of 12 student solutions. The students were asked to solve the open-ended geometry/optimization problem that required them to propose bus routes, as described in the previous section. The answers were written on paper. Each answer was scanned and anonymized. Next, the answers were uploaded to the LLM, which then provided a detailed analysis report on each of the answers. The reports were stored with student answers.

1.4.6 Data Analysis

The twelve student solutions were collected and evaluated using the final version of the prompt. Each response was analyzed both by the language model and by the research team. The analysis proceeded in several stages. First, all AI-generated reports were reviewed in full. Each assessment was then examined across several analytic dimensions: (a) mathematical content, (b) correctness and validity of reasoning, (c) representations, and (d) problem-solving characteristics. For each dimension, two criteria were applied: (1) whether the language model identified the relevant elements present in the student’s work (yes/no), and (2) whether its interpretation and description aligned with the expert evaluators’ assessment and terminology. Representative examples were selected for each dimension to illustrate the main patterns observed in the AI’s analyses. These examples highlight recurring phenomena such as accurate identification of mathematical structures or occasional inaccuracies, among other phenomena relevant to the incorporation of LLM’s assisting teachers in inquiry-based learning.

1.5 Results

In order to describe how an LLM analyzes students' written responses to a rich mathematical problem according to IBL principles, we designed the results section in accordance with the different stages of the AI generated reports, as specified above. As we believe that the actual work of the students must be viewed

in order for the reader to understand the challenge of complex work on rich tasks, we present authentic figures (sometimes handwritten in Hebrew) followed by a translation. We begin each section with a representative student's answer and the corresponding analysis (Student 12). Then we present different examples of other student's work and their corresponding analysis to highlight the observed patterns in the LLM's analysis.

1.5.1 Meeting the requirements of the task

The comments by the LLM about meeting the different requirements of the task were generally accurate when analyzing visual representation. Figure 2 shows Student 12's answer (with a translation below of the Hebrew writing), and Table 2 shows the LLM's analysis of meeting the task's requirements.

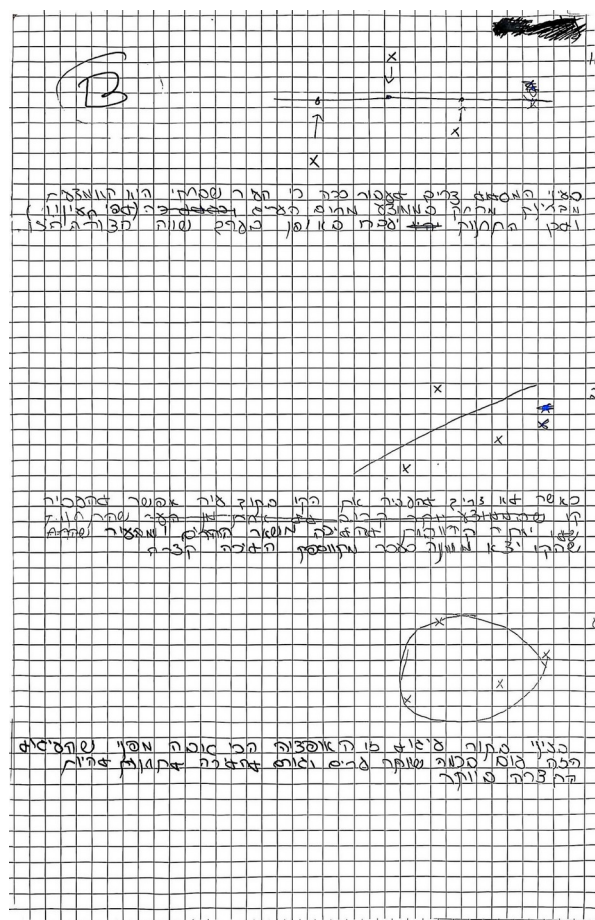


Figure 2: Student 12's answer to the task

Translation:

1A. In my view, the route should go like this – as I sketched – because the city I chose is the middle one in terms of average distance from most of the cities – by eye – and therefore the stops will be approximately equal.

1B. When the line does not need to pass through one of the cities, it is possible to draw a line whose stops are close in terms of walking distance from the other cities and from the city it started from. Compared to before (sub task 1A), a short walk is added.

1C. In my view, a circular route should be created that surrounds all the cities, because this is the best option as this circle passes through as many cities as possible and causes the walking distance to the stations to be the shortest.

Table 2: Rubric assessment on meeting the requirements for student 12

Criterion	✓ / ✗	Comment
Sub-task 1A	✓	A straight line with the reasoning “visually appears to be the middle .”
Sub-task1B	✓	Another straight line with an argument based on walking distances.
Sub-task 1C	✓	A circular route with reasoning of coverage.
Task 2	✗	Not answered.

Several cases in which the LLM misanalyzed a student’s answer were cases in which the student described the answer verbally, but did not draw it in a way the LLM could recognize it. For example, Figure 3 shows Student 2’s answer to task 2. There is no indication that it is the answer for task 2 (no number), and the marking of the fifth city is not that distinct as there is also an erased one.

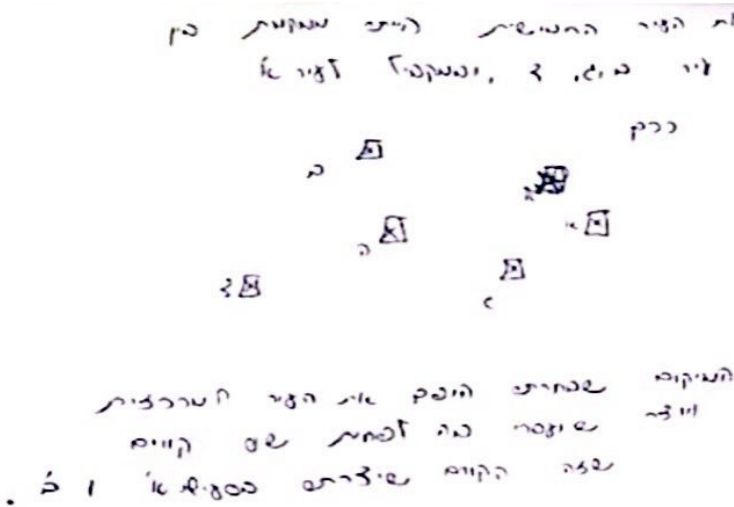


Figure 3: Student 3's answer to task 2

Translation:

I would place the fifth city between city B, C and D, and parallel to city A – the location I chose makes the city central and creates a situation in which at least two routes would pass through it, which are the straight lines I created in sections 1A and 1B in the first task.

Another notable phenomenon is that the accompanying comments were different in wording, even when sharing a common meaning. For example, Student 3 and Student 7 both submitted sketches of straight lines in 1A (Figure 4).

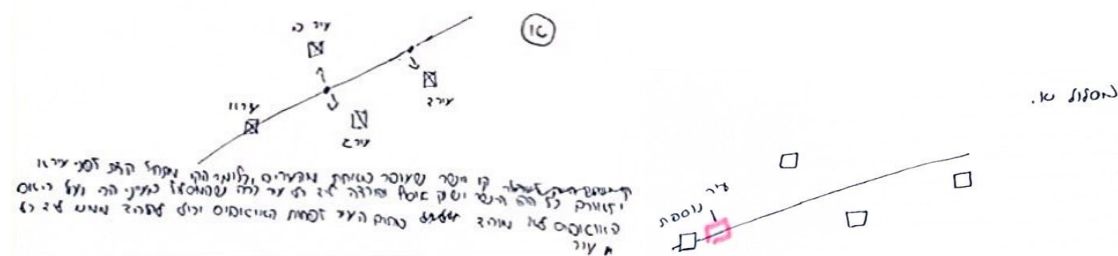


Figure 4: Student 3 and Student 7's answers to task 1A

The LLM described student 3's answer as "A straight line (even if drawn freehand)", while student 7's answer was described as "A straight line drawn freehand but considered a straight line."

Finally, The LLM did not analyze fragmented answers. When Student 5 specified in her answer to task 2 which route was the best one, the LLM dismissed it stating that "The answer dealt with choosing an optimal

route – not a requirement of the task.” Although this was part of the assignment and specified in the first part - namely task 1.

One notable point was that when the LLM misanalyzed the student answer, this obviously made the following parts of the analysis irrelevant. The example of the report regarding student’s 5 work on task 2, which the submitted visual answer shown in Figure 5 clearly shows a fifth city marked, which was wrongly assessed.

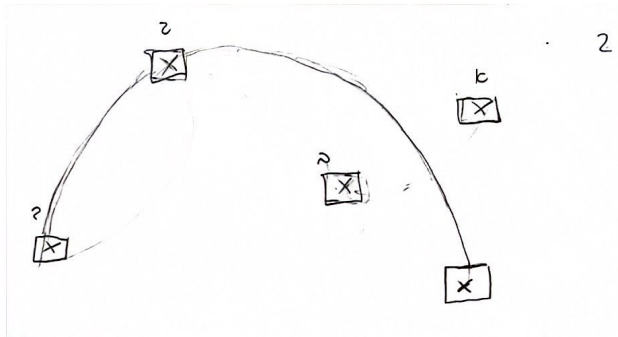


Figure 5: Student 5’s answer to task 2

1.5.2 Descriptive assessment of general problem solving characteristics

The LLM succeeded in general in describing the PS (problem solving) characteristics in student answers.

Describing Student 12’s answer (Figure 2), the LLM stated as described in Table 3.

Table 3: Descriptive assessment of general problem solving characteristics for student 12

Criterion	Comments
Representations and Action	<ul style="list-style-type: none"> • Drawings + verbal formulations. • Transition from a straight line to a circle.
Quantitative Consideration / Measure	<ul style="list-style-type: none"> • Use of intuition (“visual average”). • No formal calculation.
Reasoning and Conclusions	<ul style="list-style-type: none"> • Conceptual explanations for each section. • No boundary check or formula.

Breaking down the categories, in the first one, representation and actions, the LLM analyzed the sketches of the lines, circles, and curves correctly. In addition, it recognized verbal and symbolic representations when they appeared. In some cases there are some nuanced interpretations of moving between representations as when more than one representation was used. This was attributed as switching (going back and forth) between representations, although the use of graphical representation seems like a unidirectional description, and the meaning of a “switch” is not evident in the student’s work. See Figure 6 (Student 1), for example, when a verbal description is used in addition to the visual representation.

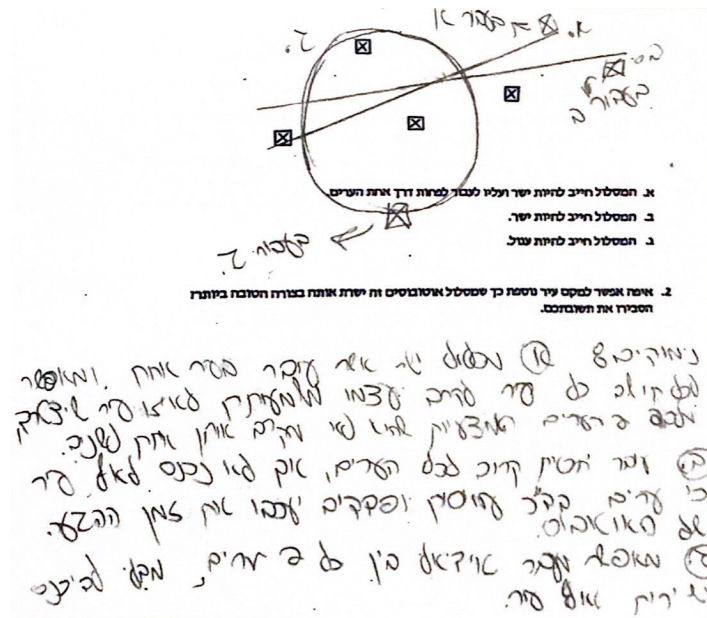


Figure 6: Student 1's answer to the task

Translation:

1A. A straight route that passes through one city, and allows every resident in any other city to bring themselves significantly closer to whichever city they need, except for the two middle cities which the route does not bring residents closer to each other.

1B. The route passes relatively close to all the cities but does not enter into any city, because cities are usually crowded and traffic jams would delay the bus's arrival time.

1C. The circular route allows residents of the cities travelling between any two cities without entering directly into any city.

In the third category, Reasoning and conclusions, the LLM analyzed the argumentation for each of the different suggested routes. When there was no explanation beyond the sketch, the LLM referred to it as "intuitive conclusion" (Student 3), or "intuition and visual alternatives" (Student 1, Figure 6). It was looking for explicit mentioning of a measure, and employing "what if not" strategy as was analyzed for the solution of Student 4, appearing in Figure 7, which was described as: (a) Reasoning based on intuition and visual alternatives; (b) No proof or formula; and (c) No evidence of a "what if not?" strategy. The long answer, which includes both graphic representations and long texts, might be characterized as a bit blurry, but it includes quantitative measure (time), and also reasoning about traffic and also directionality. Yet it also contradicts the statements about avoiding the traffic suggesting it goes through the city. There is no evidence for checking visual alternatives - on the contrary, the ideas are conceptional and general.



pass through as many cities as possible, and I quickly reached the conclusion that we would definitely have to sacrifice one city. I came up with several options. For example, the green and orange circles are smaller and therefore require fewer resources and construction maintenance of the route compared to the yellow one. However, one can see a certain advantage in the yellow route, since in this route the added city is located at the center of the path, which provides a sense of unity among all the cities.

Then I moved to requirement 1B:

Requirement 1B: If the requirement is that the route must be a straight line, then it is more correct and fair to make a route through the center of all the cities. This is not really possible with a continuous straight line, so I tried to create a kind of trend line. I measured the distances between the cities with a ruler (both in terms of the x-axis and the y-axis) and created a coordinate system with points representing the cities. To make the trend line accurate, I entered the points into Excel and obtained the slope and the y-intercept of the trend line, and then roughly sketched it on the page.

Then I moved to requirement 1A

[Second handwritten page] If the route must be straight and must pass through at least one city, as in requirement 1C, the most efficient option would be to give up on one city. Although here as well it would be possible to create a trend line that we would 'force' to pass through one of the cities, its efficiency greatly decreases when the route goes through fewer cities. The moment it passes through two cities, that is the optimal route.

So according to which requirement or criterion should we create the best route? First of all, what makes a route efficient? The number of cities the route actually passes through – very important, because if public transport does not pass directly through a city, it means people must take both public transport and non-public vehicles to reach public transport, which in most cases defeats the purpose of public transport. The length of the route – is it just unnecessarily long? This also requires more resources and maintenance and also wastes passengers' time. And does the route serve all the cities in both directions?

According to these measures we can rule out the route from requirement 1B. It does not pass through any of the cities so it will not really solve the cars on the road problem and it will not be comfortable to reach it's stations. Now we need to decide between requirement 1A and 1C in which we pass through a city. If I understood the task correctly then when the route is straight it is not bidirectional so according to requirement A it will be simple and comfortable to get from one city to another but what will you do on the way back? You will go back to personal transportation... Therefore I would go with the route in Requirement 1C, since it is [Third handwritten page] circular you can get to any city on the route. On the second hand if the routes are bidirectional and the problem created by the one way straight route disappears, the route from Requirement 1A wins since it is the shortest one so its efficiency measure wins the efficiency of the circular route which goes through unnecessary distance without a reason.

2. I will answer according to the two approaches I mentioned above. First approach – where the routes are one-way: now we can place an additional city in one of the 'dead' sections of the circular route, which will give it a reason to pass there and increase its efficiency. According to the second approach I mentioned, where the routes are two-way, we will place another city on the line itself. The specific location does not matter, but it is better that it be between the existing cities, since then there is no need to build and pave additional roads without an existing need.

This shows that an abundance of unstructured ideas presented verbally, might prove difficult for an LLM to process according to specific rubrics. The flow of thought presented by the student in a very elaborate and associative form was not properly analyzed by the LLM.

Mathematical communication

The LLM accurately analyzed the communication modalities used (Words, visual, mathematical). It also added an assessment of the level of readability and clarity, as well as richness. In cases that were missing elaboration such as student 3's work (Figure 8) the LLM noted that as well, stating that:(a) Legible at a basic level, and (b) Lacking verbal elaboration.

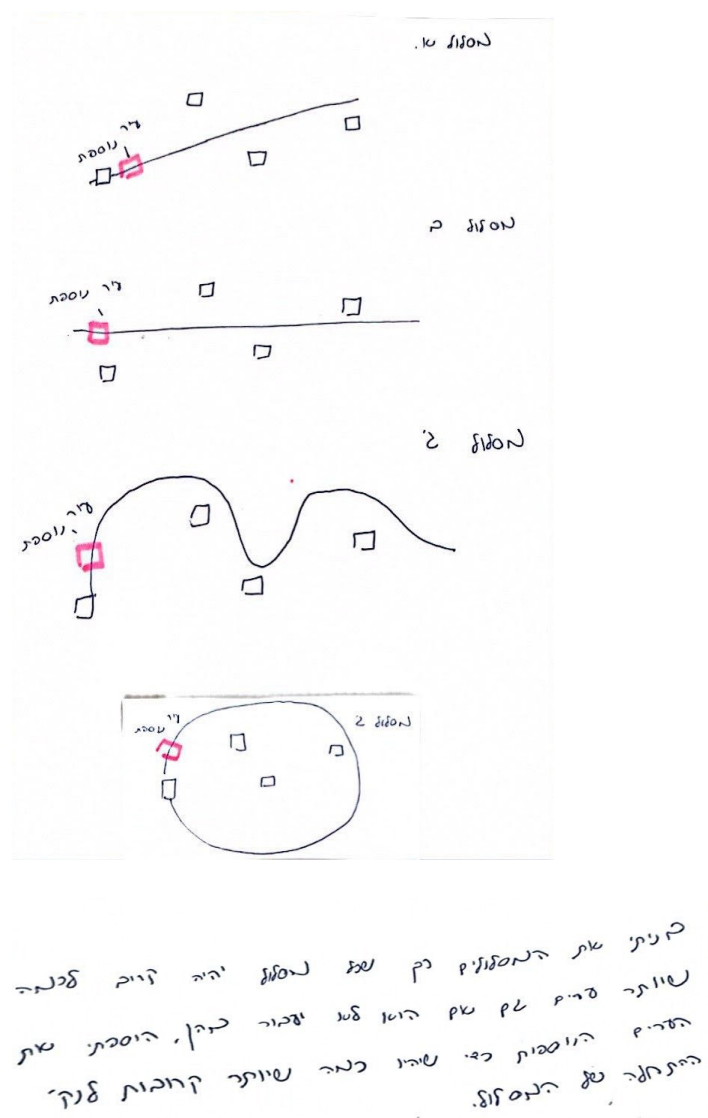


Figure 8: Student 3's answer to the task

Translation:

I built the routes so that each route would be close to as many cities as possible even if it does not pass through them. The additional cities; I added a fifth city in each case so that it would be as close as possible to the starting point of the route for each requirement in the task.

In cases the communication was off topic, for example student 10’s submitted answer: “I did not understand the exercise. Even after an explanation, I really sat and tried, but I did not understand. Isn't there some kind of mathematical calculation or something? It is really hard for me to understand what to do because this is a kind of fictional exercise.”

The LLM classified it as “Verbal communication on the difficulty only”.

1.5.3 Rubric-assessment of general and mathematical problem solving characteristics

Analysis according to specific rubrics proved to be a task that the LLM executed with a high level of accuracy. Below (Table 4) we provide an example of the LLM’s rubric analysis of student 12’s work, as it appears in Figure 2.

Table 4: Rubric assessment of general problem solving characteristics

Domain	Performance Level	What Contributed	Improvement
Representations and Action	Full	Drawings + text	Add measurement
Quantitative Consideration / Measure	Partial	Intuition	Convert to numbers
Reasoning and Conclusions	Partial	Intuitive reasoning	Additional checks

Mathematical	Full	Clear	Increase mathematical
Communication		formulations	precision

For the general problem solving rubrics (Table 4), the LLM analyzed the text and drawings to determine there was a full performance level in terms of representations and actions. The LLM also determined that in terms of quantitative consideration or measurement the student's answer lacked a numerical representation. One notable result is the LLM's phrasing, using "intuitive" also in the rubrics elaboration, in the cases that were analyzed as such in the qualitative stage of the analysis.

Table 5: Rubric assessment of mathematical problem solving characteristics

Domain	Level	What Was Observed
Objects	Full	Lines, circle
Actions	Partial	Intuition only
Models	Partial	Conceptual
Accuracy	Partial	No formulas
Inference	Partial	Only local reasoning

For the mathematical characteristics rubrics (Table 5), the analysis detected the mathematical objects in the answer, and categorized the actions as “intuition only”, i. e. not demonstrating any formal mathematical actions apart from the verbal explanation. In terms of the models, the LLM analyzed that ideas about the measure exist in the student's answer, yet they were on a conceptual level only. The lack of formulas resulted in partial accuracy, and the inference was categorized as local, reflecting only the given state without any generalization, which would require a formal measure description.

1.5.4 Strength / for improvement/summary for teacher

In the final part of the assessment, the LLM provided a focused analysis of the strengths, elements requiring improvement, and a summarizing statement for the teacher. We return to student 12's solution analysis, presented in Table 6.

Table 6: Summary assessment of strengths, improvement, and summary

Strengths	Improvement	Teacher's Summary
<ul style="list-style-type: none"> • Rich response with drawings and words. • Intuitive explanations. 	<ul style="list-style-type: none"> • Turn intuition into measurable reasoning. • Complete the answer to task 2. 	<p>The solution describes ideas well but remains at a conceptual level. There is room for guidance on how to move from intuition to calculation.</p>

While this summary accurately reflects qualitative analysis, in this case the LLM uses conceptual and intuitive as carrying almost similar meaning. This also is apparent in switching between formulation and calculation. These subtleties might be overlooked by the LLM, but carry different meanings when discussed in mathematics education terms.

Analysis of the summary also showed cases in which the LLM used undefined terms, which have a potential for misinterpretation. Among the examples are “lacking full mathematization”, on student 1's work, probably related to the answer not containing symbolic formulas for measure. “Requires significant strengthening in the mathematical aspect” - on student 9's summary, related to lack of formal mathematical formulas and calculations. Obviously, “mathematical aspect” is not a well defined term and could be misinterpreted. Another example is the use of the word “convert” in some of the reports. In student 4's work, where under improvement, the LLM suggests “convert some of the statements to a full calculation”. The verb “convert” carries a mathematical meaning not necessarily fitting the idea behind the comment, which probably was to move beyond showing mathematical ideas and following them through with full calculations. “Convert” was also used in student's 7 report, in the summary part: “...lead the student to convert the understanding into mathematical calculations”, demonstrating once more a potential misunderstanding of the verb.

In one occasion, summarizing the report on student 3's answer (Figure 8), the LLM suggested "Add a detailed verbal explanation and integrate mathematical formulas." for improvement, and "There is a promising beginning of cognitive flexibility here, but it still lacks mathematical grounding or a clear verbal explanation." in the teacher's summary. Yet the analysis of the presented idea would probably lead a teacher to conduct a process of development of the idea into a quantifiable measure, rather than focus on the verbal explanation. This might be an occurrence in which the LLM strives to achieve the goal of the task, yet is not able to provide meaningful scaffolding on how to reach it.

Two of the students (8 and 10) did not answer the task, as can be seen in student 10's answer quoted above. In these cases, the LLM summarized that "no strengths were identified", and for improvement, that "Guide the student on how to begin from the problem requirements themselves.", leading to the teacher's summary: "This is not an outlier case but rather a situation indicating difficulty in approaching the problem. It is important to view this as a sign of the challenges that open inquiry problems pose for students."

Finally, when assessing the requirements inaccurately, the summary is skewed. This appears in the summary of student 5's work (presented in Figure 5). One of the parts was not accurately assessed, leading the summary to state that "This solution constitutes an outlier - the student answered a different task" - which is not the case, leading to a wrong analysis.

1.6 Summary and discussion

Recent literature calls for the development of mathematics-specific approaches to automated assessment that remain rooted in the authentic mathematical work of students (Olsher et al., 2024; Stacey and Wiliam, 2013). While the use of large language models may easily follow broader technological trends in education through tools such as ChatGPT, Moodle, or Google Classroom, this study demonstrates that their potential can be refined and customized to address challenges that are particular to mathematics education. This responds directly to the call by Olsher et al. (2024) to move beyond generic applications of technology and toward discipline-sensitive integration.

Research on AI in education further highlights the dual nature of generative AI. On the one hand it offers new forms of analysis, and on the other hand it may generate unexpected or disruptive feedback (Selwyn, 2022). As with any educational technology, productive integration requires schemes of use that are developed by teachers and embedded in practice (Trouche, 2005). Introducing such tools inevitably creates moments of noise, understood as small diversions from a planned learning trajectory (Cusi and Contell, in press), but these moments can be pedagogically meaningful when interpreted carefully. The present study illustrates how a clearly defined assessment framework and a principled task design can channel some of this noise into insight and help educators make purposeful use of LLM-generated feedback.

1.6.1 Acknowledging contextual challenges

A central challenge that emerged from this study concerns the contextual nature of the terminology used by the LLM. ChatGPT frequently drew on terms such as “intuition” or “conceptual explanation” when characterizing the reasoning in students’ responses. These terms were often used in ways that reflected everyday experience or informal sense making rather than disciplinary notions of intuition, and in other cases referred to abstract relationships without clear mathematical grounding. Such terminology is deeply context dependent and shaped by local classroom discourse, textbook traditions, and teachers’ professional language. Similar issues are well documented in the literature on mathematical terminology as a boundary object that shifts meaning across contexts and communities of practice (Akkerman & Bakker, 2011). Establishing a common interpretive vocabulary with the LLM therefore requires deliberate prompt design, yet formulating and stabilizing this language is itself a non-trivial task for teacher professional development (Cooper et al., 2020).

These contextual challenges were also evident when the LLM attempted to translate informal reasoning into more formal computation. For example, in cases where the student’s response blended intuitive reasoning with partial procedures, ChatGPT recognized the general intention but used imprecise descriptions when suggesting how the reasoning might be strengthened. This indicates that while the model

can often identify promising mathematical directions, its ability to distinguish between productive informal reasoning and incomplete mathematical argumentation remains limited. Teachers engaging with LLM-generated assessments therefore need to be aware of these shifts in terminology and interpretation, and to consider how professional judgment can complement and refine the model's descriptions.

1.6.2 The challenge of treating answers to rich tasks

Rich mathematical problems often consist of multiple interconnected sub-tasks, and this structure poses difficulties for LLM-based assessment. A notable challenge arises when students present their work as a single comparative narrative rather than as discrete answers to separate prompts. While teachers often welcome such holistic responses because they demonstrate a unified understanding of the situation, ChatGPT struggled to extract the specific responses to each individual item. In one example, the model judged a fully coherent comparative explanation as unacceptable, simply because it did not mirror the expected item-by-item format. This difficulty suggests that LLMs need more support in interpreting narrative reasoning and segmenting it into meaningful components without penalizing integrated mathematical thinking.

At the same time, the model had difficulty synthesizing related sub-tasks across the first stages of the problem. Its summaries were frequently mild or generic, lacking the constructive detail that teachers rely on when diagnosing mathematical progress. This limitation raises important design questions about whether an assessment prompt should explicitly separate the stages of inquiry or instead preserve their interconnected nature. Introducing more structure in the prompt may reduce ambiguity, but it may also diminish the essence of inquiry by isolating what students naturally integrate. An alternative approach may involve multi-agent architectures where different agents handle different dimensions of the task, allowing the assessment to reflect the multi-directional character of inquiry, as suggested in mathematical task design by Schorcht et al. (2025).

The interpretation of inscriptions also emerged as a significant issue. For example, a curved sketch was misread as a geometric property of the solution rather than as an informal drawing, and dashed lines were interpreted as violating the task requirements. These misreadings indicate the need for task designs that help students clearly label or specify the intention of their diagrams. They also connect to broader challenges documented in environments such as the STEP platform, where student-generated representations often require contextual cues in order to be interpreted accurately (Yerushalmy et al., 2017).

A further challenge lies in determining whether a response is relevant to the mathematical intent of the task. Some submissions were labeled unusual or incorrect because they pursued alternative interpretations of the situation. A couple of students, for example, interpreted the scenario in terms of real-world mobility rather than the intended geometric model. Rich tasks are designed to allow such divergent pathways, and teachers typically engage in normative judgment to determine whether an unexpected direction still reflects meaningful mathematical reasoning. ChatGPT did not consistently distinguish between productive divergence and misinterpretation, which highlights the complexity of identifying relevance in open mathematical work.

These challenges collectively raise an important question about how LLMs might support teachers in responding to students who express uncertainty or claim not to know how to begin, as in the case of student 10. One possible direction is to design agents that identify the strengths of each submission with respect to specific assessment categories and present these insights to the teacher. A complementary agent could be trained to convey similar forms of guidance to students, phrased in ways that preserve inquiry, encourage exploration, or validate creativity. Such differentiation between teacher-directed and student-directed feedback may be essential for integrating LLM assessment into inquiry-based learning in a pedagogically responsible manner.

1.7 References

Akkerman, S. F., & Bakker, A. (2011). Boundary crossing and boundary objects. *Review of educational research*, 81(2), 132-169.

- Applebaum, M. (2025). Thinking with Machines: How AI Solves Math Puzzles and What Educators Can Learn. *International Journal of Advanced Research in Science, Communication and Technology*, 5(13), 355-375. <https://doi.org/10.48175/IJARSCT-26055>
- Chazan, D. (2000). *Beyond formulas in mathematics and teaching: Dynamics of the high school algebra classroom*. Teachers College Press.
- Chazan, D., & Ball, D. (1999). Beyond being told not to tell. *For the learning of mathematics*, 19(2), 2-10.
- Chazan, D., & Schnepf, M. (2002). Methods, goals, beliefs, commitments, and manner in teaching: Dialogue against a calculus backdrop. *Advances in Research on Teaching*, 9, 171-196.
- Cooper, J., Olsher, S., & Yerushalmy, M. (2020). Didactic metadata informing teachers' selection of learning resources – Boundary-crossing in professional development. *Journal of Mathematics Teacher Education*, 23, 363–384. <https://doi.org/10.1007/s10857-019-09428-1>
- Cusi, A., & Contel, F. (in press). A reconceptualisation of the instrumental genesis process in the context of human-LLM interactions: exploring students' perspective on their use of GPT-4o as a tool to support self-assessment. In E. Geraniou, C. Crisan & M. Mavrikis (Eds.), *Digital Technology and Artificial Intelligence in Mathematics Education Assessment* (pp. 129-149). Routledge.
- Huda, N. Anwar, K., Nofverma, & Kurniawan, W. (2025). ChatGPT scaffolding in supporting metacognition for limit concepts in guided inquiry mathematics learning. *Journal of Information Technology Education: Research*, 24, Article 38. <https://doi.org/10.28945/5645>
- Jin, B., & Suh, J. (2024). Exploring automatic scoring of mathematical descriptive assessment using prompt engineering with the GPT-4 model: Focused on permutations and combinations. *The Mathematical Education*, 63(2), 187–207. <https://doi.org/10.7468/mathedu.2024.63.2.187>
- Lampert, M. (2001). *Teaching problems and the problems of teaching*. Yale University Press.
- Olsher, S., Chazan, D., Drijvers, P., Sangwin, C., & Yerushalmy, M. (2023). Digital Assessment and the “Machine”. In: Pepin, B., Gueudet, G., Choppin, J. (eds) *Handbook of Digital Resources in Mathematics Education*. Springer International Handbooks of Education. Springer, Cham. https://doi.org/10.1007/978-3-030-95060-6_44-1
- Pepin, B., Buchholtz, N. & Salinas-Hernández, U. (2025). A Scoping Survey of ChatGPT in Mathematics Education. *Digit Exp Math Educ* 11, 9–41. <https://doi.org/10.1007/s40751-025-00172-1>
- Rigaud Téllez, N., Rayón Villela, P., & Blanco Bautista, R. (2024). Evaluating ChatGPT-generated linear algebra formative assessments. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8(5), 75–82. <https://doi.org/10.9781/ijimai.2024.02.004>
- Schorcht, S., Peters, F. & Kriegel, J. (2025). Communicative AI Agents in Mathematical Task Design: A Qualitative Study of GPT Network Acting as a Multi-professional Team. *Digit Exp Math Educ* 11, 77–113. <https://doi.org/10.1007/s40751-024-00161-w>
- Schwartz, J. L., & Kenney, J. M. (1995). Assessing mathematical understanding and skills effectively. An Interim Report of the Harvard Group Balanced Assessment in Mathematics Project, September. *Educational Technology Center, Harvard Graduate School of Education*.
- Selwyn, N. (2022). The future of AI and education: Some cautionary notes. *European Journal of Education*, 57(4), 620-631. <https://doi.org/10.1111/ejed.12532>

Stacey, K., & Wiliam, D. (2012). Technology and assessment in mathematics. *Third international handbook of mathematics education*, 721-751.

Teledahl, A., Kilhamn, C., Ahl, L. M., et al. (2025). *Defining and measuring quality in students' mathematical writing: A systematic literature review*. *Mathematics Education Research Journal*, 37, 557–576. <https://doi.org/10.1007/s13394-024-00501-4>

Trouche, L. (2005). Instrumental genesis, individual and social aspects. In *The didactical challenge of symbolic calculators: Turning a computational device into a mathematical instrument* (pp. 197-230). Boston, MA: Springer US.

U.S. Department of Education, Office of Educational Technology (2023). Artificial intelligence and the future of teaching and learning: insights and recommendations.

Van den Heuvel-Panhuizen, M., & Drijvers, P. (2020). Realistic mathematics education. In *Encyclopedia of mathematics education* (pp. 713-717). Cham: Springer International Publishing.

Yerushalmy, M., Nagari Haddif, G., & Olsher, S. (2017). How Might the Design of Tasks for Online Formative Assessment Support our Understanding of Students' Conceptions? *ZDM - The International Journal on Mathematics Education*. 49(5), 701-716. <https://doi.org/10.1007/s11858-017-0871-7>