Using a large language model to analyse problem solving and simulate student solutions in lower secondary mathematics

Jiří Přibyl[1,*], Michaela Tichá[1]

[1] Faculty of Science, Jan Evangelista Purkyně University, Pasteurova 3632/15, 400 96 Ústí nad Labem, Czech Republic; jiri.pribyl@ujep.cz

This study examines how a large language model (LLM) can support lower secondary mathematics education, not by measuring performance, but by analysing how complete and didactically useful its solution processes are. The focus is on whether the model explicitly verifies results, which is an important but often neglected phase of mathematical problem solving. Two exploratory experiments were conducted with Gemini 2.5 Pro using 24 lower secondary mathematics tasks presented in Czech. In Experiment 1, the model solved all tasks in a baseline condition and then again using a structured prompt inspired by Pólya's four phases of problem solving. In Experiment 2, the model was asked to simulate solution attempts of three contrasting student profiles. The results show that verification is a fragile step. In the baseline condition, the model often produced correct answers but did not check them. When guided by the structured prompt, verification appeared in every task. In student simulation, the model produced plausible mistakes and omissions in routine tasks, but often became unrealistically advanced in non-routine and construction problems, reducing profile fidelity. Overall, LLM outputs can support a priori didactic analysis, but they require careful interpretation, especially when used to simulate student thinking in demanding tasks.

Key words: mathematical problem solving, result verification, a priori didactic analysis, large language models, lower secondary education.

Received 2/2026
Revised 4/2026
Accepted 4/2026

1 Introduction

1.1 Mathematical problem solving, verification, and *a priori* didactic analysis

Mathematical problem solving is a core component of lower secondary mathematics education and contributes to the development of mathematical thinking (Rendl et al., 2013). It is typically conceptualised as a multi-phase process that includes reflection and verification. Although the development of verification skills is critical to improving problem-solving ability, prior research in geometry problem solving suggests that learners often omit explicit verification, and when verification occurs it may remain cursory (Papadopoulos & Dagdilelis, 2008).

Understanding how problem solving processes unfold, and where typical breakdowns occur, is therefore an important concern for both researchers and teachers in lower secondary mathematics education.

The need for verification becomes particularly visible in tasks whose structure does not uniquely determine a single interpretation or solution pathway. Jonassen (1997) distinguishes well-structured problems from ill-structured problems, the latter requiring interpretative and evaluative judgements in addition to procedural work. In such tasks, a correct-looking sequence of steps may still lead to an inadequate solution if key assumptions or constraints remain unchecked. For teachers, this creates an instructional challenge: it is necessary not only to anticipate typical procedural errors, but also to foresee how students may interpret a task and where verification is likely to be missing. This anticipatory challenge connects directly to the practice of *a priori* didactic analysis in instructional planning.

Research on mathematical cognition shows that successful problem solving itself involves anticipatory thinking, such as foreseeing intermediate outcomes, control steps, and potential breakdowns during the solution process (Carlson & Bloom, 2005). From an instructional perspective, teachers need to anticipate these processes at the level of student thinking. This anticipatory dimension is captured by *a priori* task analysis, a concept rooted in the Theory of Didactical Situations (Brousseau, 1997) and operationalised for instructional planning by Nováková (2013), in which teachers consider likely student strategies, interpretations, and errors before classroom implementation.

However, systematic *a priori* analysis is cognitively demanding and often depends on teachers' experience and intuition, particularly in tasks requiring non-routine reasoning or explicit verification. This motivates the search for tools that could support teachers in anticipating student thinking during instructional planning. These open questions motivate the present study, which examines LLM outputs not in terms of correctness but with respect to process completeness and didactic relevance.

1.2 Large language models in mathematics education: opportunities and open questions

Recent literature highlights both opportunities and risks associated with the use of large language models (LLMs) in education. LLMs may support instruction by providing feedback, generating instructional materials, and assisting with individualisation (Kasneci et al., 2023). Reviews in mathematics education similarly suggest that tools such as ChatGPT can generate examples, explanations, and assessment items for both teachers and learners, while also raising concerns about superficial learning, academic integrity, and overreliance on AI-generated solutions (Almarashdi et al., 2024).

From a didactic perspective, however, important questions remain regarding LLMs as problem solvers and as generators of student-like solution attempts. Although LLMs can often produce correct final answers, it is less clear how they enact the underlying problem solving process and whether they spontaneously engage in essential phases such as verification. McGalliard and Otten (2025) examined AI responses to challenging mathematical problems and found that while models could handle routine tasks, their performance degraded on non-routine problems requiring insight or heuristic reasoning. This finding motivates empirical studies that examine LLM outputs not only in terms of correctness, but also with respect to process characteristics and didactic relevance.

1.3 Aim and research questions

The aim of this study is to explore the didactic potential of a large language model (LLM) for analysing mathematical problem solving processes and for supporting *a priori* didactic analysis in lower secondary mathematics education. Specifically, the study investigates how an LLM enacts key phases of mathematical problem solving under different prompting conditions and to what extent the same model can be used to generate student-like solution attempts that are meaningful for instructional planning.

The study is guided by three research questions that build on each other progressively. The first question establishes a baseline by examining the model's spontaneous problem solving behaviour. The second question tests whether structured prompting can improve the completeness of the process. The third question extends the investigation to the model's capacity for generating student-like solution attempts.

- *RQ1:* To what extent does an LLM, acting as an independent problem solver, enact the key phases of the mathematical problem solving process when solving lower secondary mathematics tasks?
- *RQ2:* How does the quality and completeness of the LLM's problem solving process change when the model is guided by a structured Pólya-inspired prompting framework?
- *RQ3:* To what extent can an LLM generate solution attempts that plausibly reflect different lower secondary student profiles in terms of ability and effort, and how useful are such outputs for *a priori* didactic analysis of typical errors, incomplete strategies, and instructional needs?

2 Theoretical background

2.1 Problem solving phases and verification as result checking

Problem solving in mathematics is commonly described as a process comprising several interrelated phases, including understanding the problem situation, devising a solution plan, carrying out this plan, and subsequently examining the obtained result. Although different authors use varying terminology, there is broad agreement that successful problem solving involves not only executing procedures but also assessing whether the produced result adequately satisfies the task conditions.

In his seminal work, Pólya (1945) conceptualises problem solving as a four-step process—understanding the problem, devising a plan, carrying out the plan, and looking back—where the final phase explicitly involves examining the result and checking the correctness of both the outcome and the procedure. Pólya's framework was adopted in this study for three reasons. First, it is the most widely recognised model of mathematical problem solving and provides a common reference point across the literature. Second, it explicitly identifies verification (looking back) as a distinct phase, which makes it well suited for operationalising verification as a coding criterion. Third, its four-step structure is naturally amenable to translation into a structured prompt for a language model. The use of this framework in the present study is pragmatic rather than theoretical: it serves as a scaffolding tool for the LLM, not as a theory of cognition.

Empirical studies indicate that the final phase is frequently omitted by novice solvers. In contrast to experts, who regularly monitor and verify their reasoning, beginners tend to apply familiar procedures quickly and uncritically, often without checking the adequacy of their solution (Schoenfeld, 1985). Kontorovich (2019) argues that students' failure to check solutions is not merely a matter of carelessness but reflects deeper epistemological dispositions about what constitutes a finished mathematical task.

In this study, the term *verification* is used in a narrow, operational sense to denote explicit checking of the correctness and plausibility of an obtained result. Such checking may include, for example, substituting a solution back into the original conditions, examining whether all constraints of the task have been respected, or providing a brief justification of why the result is valid. Verification in this sense does not necessarily involve extended metacognitive reflection but rather a concrete test of the solution's validity.

From a cognitive perspective, the frequent omission of verification can be interpreted through the lens of dual-process theories of reasoning. According to these theories, human thinking involves the interaction of two qualitatively different modes: a fast, intuitive, and heuristic mode (often referred to as System 1) and a slower, more deliberate, and analytical mode (System 2). While System 1 enables rapid generation of plausible solution steps, System 2 is typically required for careful monitoring, error detection, and explicit verification of results (Liu et al., 2025). In the absence of deliberate engagement of analytical control, problem solving may terminate prematurely once a seemingly acceptable result has been obtained. Liu et al. (2025) argues that contemporary LLMs tend to operate as black-box generators of fluent solution sequences, producing responses that resemble fast heuristic reasoning unless explicit mechanisms for verification are invoked. Their proposed *Invoke–Verify–Inject* framework highlights that verification does not emerge spontaneously but must be deliberately activated as a separate control process. This perspective provides theoretical support for interpreting verification as a fragile phase of problem solving that is easily omitted unless explicitly prompted.

At the same time, research on mathematical problem solving emphasises that verification should not be understood solely as a final, terminal step. Studies of open-ended and exploratory problem solving show that the process is inherently recursive rather than linear. Cifarelli and Cai (2005), for example, demonstrate that solvers continuously reformulate goals, pose subproblems, and adjust strategies while working on a task. Such recursive activity requires ongoing monitoring of intermediate results and assumptions, effectively embedding verification throughout the solution process rather than relegating it to the end.

From an educational perspective, this recursive view further explains why verification poses difficulties for novice solvers. Continuous monitoring and revision place high demands on cognitive control and metacognitive awareness, which are often underdeveloped in lower secondary students. As a result, verification may be reduced to an optional or entirely omitted activity, especially in the absence of explicit instructional emphasis. Treating verification as a distinct and observable component of problem solving therefore makes it possible to analyse whether this essential control mechanism is present or absent in a given solution, independently of the correctness of the final result.

By focusing on verification as explicit result checking, the present study adopts an operational perspective that is directly aligned with the evaluation criteria used in the empirical analysis. This approach enables a transparent examination of the completeness of problem solving processes while remaining consistent with the scope and aims of the study.

2.2 Ill-structured tasks and the need for interpretation

Mathematical tasks differ in the extent to which their structure constrains interpretation and solution pathways. This distinction is commonly captured through the concept of *ill-structured problems*, which are characterised by ambiguous conditions, incomplete specification of constraints, multiple reasonable assumptions, and the absence of a single uniquely defined solution (Jonassen, 1997). In contrast to well-structured tasks, ill-structured problems require solvers to make interpretative decisions about what the task situation means and which assumptions are admissible.

In mathematics education, ill-structuredness is often realised in classroom practice through *open-ended tasks*. Such tasks deliberately allow for multiple solution strategies, representations, or outcomes. While the terms *ill-structured* and *open-ended* are not strictly synonymous, they are closely related: open-ended tasks can be understood as a didactic instantiation of ill-structured problems, designed to foreground interpretation and evaluation as integral parts of mathematical activity.

In ill-structured and open-ended tasks, producing a procedurally correct result does not automatically guarantee that the task has been solved appropriately. The validity of a solution depends on whether the adopted interpretation, assumptions, and strategy are consistent with the task conditions. Explicit verification therefore becomes a central control mechanism, as it enables the solver to evaluate whether the obtained result is acceptable within the intended constraints and whether alternative interpretations would lead to different conclusions.

From an educational standpoint, open-ended tasks are also linked to the development of mathematical creativity. Empirical work suggests that open-ended problems can foster key components of creative mathematical thinking, including fluency, flexibility, and originality (Suyitno et al., 2018). At the same time, these tasks place substantially higher demands on self-monitoring and evaluation than routine problems. Without explicit instructional support, students may prioritise producing a plausible solution over critically examining whether it satisfies the conditions and assumptions embedded in the problem situation.

The importance of validation mechanisms in ill-structured tasks is highlighted within the Theory of Didactical Situations. Brousseau and Gibel (2005) argues that learning emerges through interaction between the learner and the problem environment (milieu), which provides feedback on the adequacy of proposed actions and interpretations. If the milieu does not offer meaningful feedback, learners lack resources for validating their implicit models and may remain unaware of inconsistencies in their reasoning. Under such conditions, verification is less likely to occur spontaneously because the need for it is not made visible by the task environment.

This perspective is directly relevant for analysing large language models in the role of problem solvers. When an LLM is presented with an ill-structured or open-ended task without external feedback and without explicit prompts for reflection, it operates in an environment analogous to a learner facing a non-responsive milieu. Under these conditions, the model—much like a student—may generate a fluent and procedurally coherent solution without recognising the need for explicit verification. This makes ill-structured and open-ended tasks a particularly sensitive context for examining verification as a fragile but didactically significant phase of the problem solving process.

2.3 *A priori* didactic analysis and anticipatory teacher knowledge

In mathematics education, *a priori* didactic analysis refers to systematic work with a task before it is implemented in the classroom, focusing on how students are likely to interpret the task, which strategies they may employ, and which errors or incomplete approaches may arise (Brousseau, 1997). In this sense, *a priori* analysis represents a form of anticipatory teacher knowledge that supports instructional planning by identifying critical points of reasoning, likely breakdowns, and opportunities for targeted intervention.

The ability to work productively with student errors is increasingly conceptualised as part of teachers' professional competence. Research on teacher noticing describes this competence as a process involving perception, interpretation, and decision-making. In the PID model proposed by Hoth et al. (2022), teachers must first notice a student's error, then interpret its underlying cause, and finally decide on an instructional response. While the PID model primarily describes in-the-moment classroom reasoning, *a priori* analysis can be seen as directly supporting its first two components by strengthening teachers' sensitivity to typical errors and their interpretative readiness.

Importantly, anticipatory accuracy is not fixed. Stannard and Foster (2025) provide evidence that secondary mathematics teachers can significantly improve their ability to predict student errors through repeated practice and comparison with empirical student data. However, such anticipatory work is time-demanding, especially for complex or non-routine tasks. Empirical findings suggest that expert teachers adjust their anticipation time based on task complexity; they devote substantially more time to anticipating typical student errors than novice teachers as task complexity increases, whereas they identify errors in simpler tasks more rapidly (Pankow et al., 2016). This highlights a practical tension: *a priori* analysis is didactically valuable, but often constrained by time and cognitive load.

From this perspective, large language models may serve as supportive tools for *a priori* didactic analysis by rapidly generating multiple plausible solution attempts, including incomplete reasoning, typical procedural errors, and missing verification steps. In the present study, such outputs are not interpreted as predictions of individual student behaviour, but as illustrative artefacts that may stimulate teachers' anticipatory thinking and support the interpretation phase of professional noticing.

At the same time, the use of AI-generated outputs in lesson preparation requires a critical stance. Recent work emphasises the need for *critical AI literacy* among educators, understood as the ability to evaluate AI outputs with awareness of their limitations, biases, and potential hallucinations (Ocak et al., 2025). When LLMs are used to simulate student thinking, their responses may appear pedagogically plausible while still misrepresenting actual student reasoning or overstating mathematical sophistication. Consequently, AI-generated solution attempts should be treated as material for reflection rather than as authoritative representations of students' cognitive processes.

2.4 Using large language models as simulated learners to support instructional planning

Recent research increasingly explores the use of artificial intelligence not only as a source of solutions or feedback, but also as a means of simulating learner behaviour for the purpose of teacher education and instructional planning. In this context, large language models have been proposed as *simulated learners*-artificial agents that generate plausible student-like solution attempts, including typical errors, misconceptions, and incomplete reasoning.

A key strength of LLM-based simulated learners is that they can generate student-like solution attempts that include not only procedural errors but also conceptually flawed interpretations of the task. In AI-driven role-play environments, teachers can therefore work with a range of plausible yet imperfect student responses and rehearse diagnostic reasoning and instructional moves in a low-risk setting (Zhuang & Zhang, 2025). Such simulations support a *decomposition of practice* by allowing teachers to focus on specific aspects of teaching (e.g., noticing and interpreting errors) without simultaneously managing the full complexity of classroom interaction (Zhuang & Zhang, 2025).

Zhuang and Zhang (2025) introduced *Student GPT*, a system designed to simulate erroneous mathematical reasoning in the domain of ratios, enabling pre-service teachers to practise diagnosing student misconceptions and selecting appropriate responses. Their findings suggest that repeated work with simulated student solutions allows teachers to practise and reflect on situations that are difficult to capture and revisit in real classrooms.

At the same time, empirical work cautions against assuming full fidelity of LLM-based student simulations. One limitation concerns the instability of role adherence: even when prompted to act as a struggling student, LLMs may shift into an expert or teacher-like register, which reduces the authenticity of the simulated learner and may distort the diagnostic task for the teacher (Zhuang & Zhang, 2025). More broadly, Guerra et al. (2025) argue that interactions with large language models can create an *illusion of understanding*, in which AI-generated responses appear coherent and pedagogically plausible while lacking grounding in human cognitive processes. From this perspective, simulated student solutions should not be interpreted as representations or predictions of actual student thinking, but rather as constructed artefacts whose plausibility depends on task type, prompting, and model characteristics.

Beyond student simulation, large language models have also been shown to support instructional preparation in the form of scaffolding and material design. Malik et al. (2025) demonstrate that, when appropriately prompted, LLMs can generate scaffolding sequences and preparatory tasks for middle school mathematics that are comparable in quality to those produced by expert educators. Such findings support the view that LLMs may contribute to *a priori* planning by helping teachers explore alternative task formulations, anticipate learning trajectories, and design targeted supports.

Taken together, these strands of research suggest that the educational value of large language models lies not in replacing teacher judgement, but in augmenting reflective instructional planning. When used as simulated learners or generators of preparatory materials, and when embedded within a framework of *a priori* analysis and critical interpretation, LLMs can support teachers in anticipating student thinking and identifying potential breakdowns in problem solving processes. From a cognitive perspective, these affordances and limitations are consistent with accounts that distinguish between rapid generation of plausible responses and slower processes of verification and integration, suggesting that AI-generated simulations should be treated as heuristic artefacts rather than models of fully grounded student cognition.

3 Methodology

3.1 Design overview

The study adopts a qualitative demonstration design aimed at exploring the problem solving behaviour of a large language model (LLM) and its potential didactic use in lower secondary mathematics education. Rather than evaluating model performance in terms of efficiency or optimisation, the focus is on analysing the structure and completeness of problem solving processes and on examining the didactic relevance of model-generated solution attempts.

The methodology comprises two complementary experiments. Experiment 1 investigates the behaviour of the LLM when solving lower secondary mathematics tasks under two different prompting conditions. In the baseline condition, the model acts as an independent problem solver without explicit methodological guidance. In the second condition, the model is guided by a structured Pólya-inspired framework that explicitly prompts different phases of problem solving. This experiment addresses Research Questions RQ1 and RQ2 by comparing the presence and quality of problem solving phases across conditions.

Experiment 2 explores the extent to which the same LLM can be used to generate student-like solution attempts by simulating different lower secondary student profiles. The aim of this experiment is not to predict individual student behaviour but to examine whether the generated outputs can plausibly reflect typical strategies and errors and whether they are useful for *a priori* didactic analysis. Experiment 2 addresses Research Question RQ3.

Across both experiments, the analysis is based on qualitative evaluation using explicitly defined criteria aligned with the theoretical framework outlined above. The methodological choices are designed to ensure transparency and consistency between the theoretical concepts, the experimental procedures, and the subsequent interpretation of results.

3.2 Model and experimental setting

All experiments were conducted using the Gemini 2.5 Pro large language model on 22 September 2025. The model was accessed in a standard interactive setting without any task-specific fine-tuning or additional training.

For each experimental condition, a single interaction thread was used. In Experiment 1, the full sequence of 24 tasks was presented within one interaction in the baseline condition and within a separate interaction in the Pólya-guided condition. In Experiment 2, separate interaction threads were used for each simulated student profile. No feedback was provided between tasks, and the model was not allowed to revise or correct its previous responses.

All tasks were presented to the model in Czech, and all responses were generated in Czech. The tasks were designed or selected to correspond to the mathematical level and forms of expression expected at the lower secondary level. Czech was used consistently as the language of interaction throughout the study. As Czech represents a less-resourced language in the training data of contemporary large language models, the observed problem solving behaviour may be influenced not only by task characteristics and prompting conditions but also by language-related factors.

In the baseline condition of Experiment 1, the model received no instructions regarding how to approach or present its solutions; the input consisted solely of the task statements. In the Pólya-guided condition and in Experiment 2, the model's behaviour was influenced only by the explicit prompts described in the corresponding methodological sections. No additional guidance regarding solution structure, verification, or expected length was provided beyond these prompts.

The experiments were conducted over a limited time period to reduce the potential impact of model updates. While the exact internal parameters of the model are not accessible, procedural transparency is ensured by clearly documenting the prompts, task set, and evaluation criteria used in the analysis.

3.3 Task set and task classification

The study is based on a set of 24 mathematical tasks selected for the purposes of the present analysis. The task set combines tasks designed or adapted by the authors with non-routine tasks. The non-routine tasks included in Appendix B were taken from publicly available problem sets of the Czech Mathematical Olympiad (MO). The tasks are reproduced for research and academic purposes with full source attribution. The original official wording and task context (year, round, category) are credited in the references. All tasks were chosen to be appropriate for the lower secondary level in terms of mathematical content, required procedures, and expected forms of reasoning.

To support a differentiated analysis of problem solving behaviour, the tasks were organised into four categories according to their structural characteristics and cognitive demands. The first category consisted of well-structured algebraic tasks, in which the problem conditions were explicitly stated and a standard solution procedure could be applied. These tasks served as a reference point for analysing basic procedural correctness and explicit result checking.

The second category comprised ill-structured or open-ended word problems designed or adapted by the authors. These tasks allowed for ambiguity in interpretation or admitted multiple reasonable solution approaches and required solvers to make interpretative decisions. They provided a suitable context for examining whether explicit verification was used to check the adequacy of the adopted interpretation.

The third category included non-routine problems taken directly from the Czech Mathematical Olympiad, category Z9, which targets mathematically advanced ninth-grade students (aged 14–15). They typically require insight, heuristic reasoning, or a non-standard strategy and are substantially more demanding than routine classroom tasks. Research has shown that AI models may struggle with such non-routine problems even when performing well on standard tasks (McGalliard & Otten, 2025). The inclusion of these tasks therefore provides a stringent test of the model's problem solving behaviour in mathematically demanding situations.

The fourth category consisted of geometric construction tasks designed by the authors. These tasks required solvers to propose and justify a construction procedure and to consider the validity of the resulting construction.

Each category contained six tasks, resulting in a total of 24 tasks. The same task set was used in both prompting conditions in Experiment 1. In Experiment 2, a subset of tasks was selected from each category in order to reduce the overall number of simulated solutions while preserving the diversity of task types.

3.4 Experiment 1: LLM as problem solver

Experiment 1 examined how the large language model responded to lower secondary mathematics tasks under two different conditions: (a) without any instructional prompting beyond the task statements themselves, and (b) with an explicit, structured prompt inspired by Pólya's problem solving framework.

3.4.1 Procedure and conditions

In both prompting conditions, the model interacted with the full set of 24 tasks within a single interaction thread. The tasks were presented sequentially, one after another, without feedback or correction between tasks.

In the baseline condition, the model received no instructions regarding how to approach or present its solutions. The input consisted solely of the task statements provided in Czech. No guidance was given concerning solution structure, explanation of reasoning, or result checking. This condition was intended to capture the model's spontaneous responses to a sequence of mathematical tasks.

In the Pólya-guided condition, a new interaction thread was initiated. At the beginning of the interaction, the model received a structured introductory prompt inspired by Pólya's problem solving framework. This prompt instructed the model to address four phases of problem solving: understanding the problem, devising a plan, carrying out the plan, and looking back, with explicit emphasis on result checking in the final phase. The full wording of this prompt, in its original Czech form, is provided in Appendix A. The same 24 tasks were then presented sequentially within the same interaction thread. After each task statement, a brief reminder was included prompting the model to adhere to the problem solving framework introduced at the beginning of the interaction. No feedback was provided between tasks, and the model was not allowed to revise or correct its previous responses. The full wording of the 24 tasks (in Czech) is provided in Appendix B. The complete transcripts of all model interactions, including all task statements and verbatim model-generated responses for both experiments (approx. 200 pages), are provided as Supplementary Material and are available at: https://https://osf.io/k2vpw/.

3.4.2 Evaluation criteria for Experiment 1

The solutions generated in Experiment 1 were analysed using a set of qualitative evaluation criteria focusing on the presence and completeness of key phases of the problem solving process. The analysis did not aim to assess mathematical elegance or efficiency but rather to examine whether essential components of a complete problem solving process were present in the model's responses.

Each solution was evaluated independently with respect to four criteria: (1) understanding of the task, (2) solution process, (3) correctness of the result, and (4) verification as result checking. The criteria were applied consistently across both prompting conditions.

Understanding of the task This criterion captured whether the model's response demonstrated an appropriate interpretation of the task. A solution was considered to show understanding if the subsequent reasoning addressed the relevant task conditions and goals. Misinterpretations of the task statement or solutions addressing a different problem were coded as lacking task understanding, regardless of the correctness of intermediate steps.

Solution process The solution process criterion focused on whether the model provided a coherent sequence of steps leading from the interpretation of the task to a proposed result. Solutions were coded positively if they included a logically connected procedure or line of reasoning. Fragmented, incomplete, or internally inconsistent solution attempts were coded negatively, even if a numerical result was present.

Correctness of the result Correctness referred to whether the final result produced by the model was mathematically correct with respect to the adopted interpretation of the task. This criterion was evaluated independently of the presence or absence of explicit verification. A solution could therefore be correct without being verified, or verified without being correct.

Verification as result checking Verification was operationalised as the explicit checking of the obtained result with respect to the task conditions. This included, for example, substituting the result back into the original conditions, explicitly confirming that all constraints had been satisfied, or providing a brief justification of why the result was valid. Implicit confidence statements or unexamined final answers were not considered evidence of verification.

For each criterion, solutions were coded using a three-level scale (0–2), where 0 indicated absence, 1 partial or ambiguous presence, and 2 clear and explicit presence of the respective feature. The coding scheme was applied to all solutions generated in Experiment 1 and formed the basis for the subsequent qualitative comparison between the baseline and the Pólya-guided conditions.

3.5 Experiment 2: LLM as simulated learner

The experiment was conducted using the same model and interaction setting as in Experiment 1. All prompts and task statements were provided in Czech, and all responses were generated in Czech.

3.5.1 Student profiles

Three student profiles were defined to reflect differences commonly observed among lower secondary students with respect to mathematical proficiency, persistence, and attention to result checking. The profiles were informed by typologies of student performance and typical difficulties described in the literature, in particular by the classification of student solution types discussed by Rendl et al. (2013).

Each profile was described in a short but explicit narrative specifying typical characteristics of the student's approach to mathematical problem solving. The profiles were intentionally formulated in a non-diagnostic manner and were not intended to represent real individual students. Their purpose was to provide a didactically meaningful contrast between different types of solution attempts that teachers may anticipate when planning instruction. The full wording of the student profile prompts, in their original Czech form, is provided in Appendix C.

3.5.2 Procedure and task selection

From each of the four task categories described earlier, two tasks were selected, resulting in a total of eight tasks used in Experiment 2. This selection was made to preserve task diversity while limiting the overall number of generated solutions.

For each student profile, a separate interaction thread was initiated. At the beginning of each thread, the model received a detailed description of the corresponding student profile. The selected tasks were then presented sequentially within the same thread. This design was chosen to maintain consistency in the simulated student behaviour across multiple tasks.

As in Experiment 1, no feedback was provided between tasks, and the model was not instructed to revise or correct its solutions. In contrast to the baseline condition of Experiment 1, the prompts in Experiment 2 explicitly framed the model's role as that of a student rather than an independent problem solver.

3.5.3 Evaluation criteria for Experiment 2

The generated solution attempts were analysed qualitatively with respect to role fidelity and didactic usefulness.

Role fidelity Role fidelity referred to the extent to which the generated solution attempts were consistent with the specified student profile across tasks. Solutions were considered to show high role fidelity if the level of mathematical reasoning, types of errors, degree of persistence, and presence or absence of result checking aligned with the profile description throughout the interaction thread.

Didactic usefulness Didactic usefulness captured whether the generated solution attempts could support teachers in *a priori* didactic analysis. Solutions were considered didactically useful if they illustrated typical student errors, incomplete reasoning, missing or incorrect verification steps, or partially successful strategies that could reasonably occur in classroom practice. Such solution attempts may help teachers anticipate critical moments in instruction, prepare targeted questions, or plan follow-up activities. Fully polished expert-like solutions or implausible student responses were considered less useful from a didactic perspective.

The analysis in Experiment 2 was exploratory and qualitative in nature and aimed to identify illustrative patterns rather than to quantify frequencies or make predictive claims.

3.6 Trustworthiness of the design

Several design choices support the trustworthiness of the study. The experimental procedures, prompts, and task set are documented in detail, including the original Czech wording of all prompts and tasks (Appendices A–C). The use of the same task set across conditions in Experiment 1 enables a focused comparison of problem solving behaviour with and without structured guidance. Finally, the explicit evaluation criteria provide a transparent basis for the qualitative coding.

4 Results

4.1 Experiment 1: Completeness of problem solving phases and result verification

4.1.1 Overall comparison of baseline and Pólya-guided conditions

Table 1 reports the number of tasks in which selected problem solving components were present in each experimental condition. A component was considered present if it was coded as partially or clearly present (codes 1 or 2).

As shown in the table, the model produced a response that included explicit task understanding, a coherent solution process, and a final result for all tasks in both conditions. These components were therefore present across the entire task set.

In contrast, clear differences between conditions were observed with respect to explicit result verification. In the baseline condition, explicit verification was observed in only 7 out of 24 tasks, whereas in the Pólya-guided condition verification was present in all 24 tasks.

Table 1: Occurrence of key problem solving components across tasks in Experiment 1 ($n = 24$)

Problem solving component	*Baseline*	*Pólya-guided*
Explicit task understanding	24	24
Coherent solution process	24	24
Correct final result	24	24
Explicit result verification	7	24

4.1.2 Verification across task categories in the baseline condition

Table 2 shows the occurrence of explicit result verification across different task categories in the baseline condition. Verification was unevenly distributed across task types.

Table 2: Occurrence of explicit result verification across task categories in the baseline condition (Experiment 1, $n = 24$)

Task category	*Tasks with verification*	*Total tasks*
Category A: Well-structured algebraic tasks	4	6
Category B: Ill-structured word problems	1	6
Category C: Non-routine (olympiad) problems	2	6
Category D: Geometric construction tasks	0	6

Figure 1 presents these results graphically, making the contrast between the two conditions immediately visible across all task categories.

Explicit verification occurred most frequently in well-structured algebraic tasks, where it was present in four out of six tasks. In ill-structured word problems, verification was observed only once. In non-routine olympiad problems, explicit verification occurred in two tasks. No explicit verification was observed in any of the geometric construction tasks.

These results indicate that, in the absence of explicit guidance, the occurrence of verification depended on task characteristics.

In the Pólya-guided condition, explicit result verification was present in all tasks across all task categories.

4.1.3 Illustrative examples of verification behaviour

To provide a richer picture of the observed patterns, we present selected examples illustrating how verification was enacted or omitted across task categories.

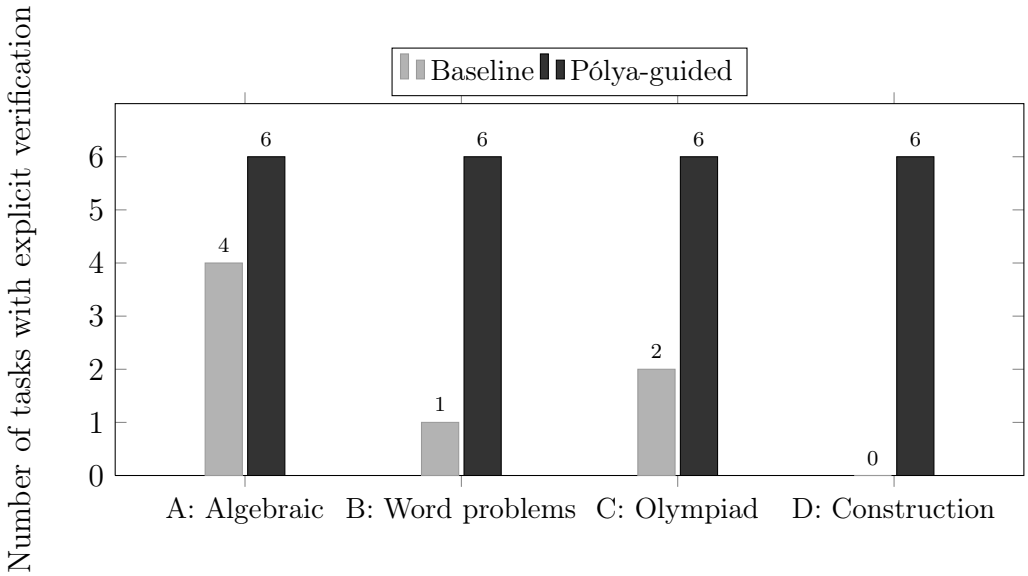

Fig. 1: Occurrence of explicit result verification across task categories in the baseline and Pólya-guided conditions (Experiment 1, $n = 6$ tasks per category)

Verification present (Task A1, baseline) For the equation $3x - 7 = 2x + 5$, the model produced a clean solution with all intermediate steps and concluded with an explicit substitution check: it substituted $x = 12$ back into both sides of the equation and confirmed equality. This exemplifies verification as explicit result checking in its most straightforward form.

Verification absent (Task B1, baseline) For the word problem about a car travelling at 60 km/h for 2 hours including a break, the model correctly identified the ambiguity (the break duration is unknown, so the distance cannot be uniquely determined) and presented alternative cases. However, no verification was performed: the model did not check whether its proposed cases were consistent with the problem conditions, nor did it reflect on whether additional interpretations existed.

Verification absent (Task D4, baseline) For the construction of a rhombus with given diagonals, the model described a correct and elegant construction procedure using the perpendicular bisector. However, no final verification was provided—the model did not check that the constructed quadrilateral indeed possessed all required properties of a rhombus (equal sides, perpendicular bisecting diagonals).

Verification present (Task D4, Pólya-guided) The same task, when solved under the Pólya-guided condition, included an explicit verification step. After completing the construction, the model confirmed that the diagonals bisect each other perpendicularly and that all four resulting right triangles are congruent, thereby establishing that the figure is indeed a rhombus.

Detailed task-level coding results and qualitative notes supporting the evaluation for both experimental conditions are provided in Appendices D and E.

4.2 Experiment 2: Simulation of student solution profiles

Experiment 2 examined whether the large language model could generate solution attempts that plausibly correspond to different lower secondary student profiles and whether such solutions are didactically useful. The evaluation did not focus on mathematical correctness but on (a) alignment with the intended student profile and (b) didactic usefulness of the generated solutions.

Three student profiles were considered: a diligent but low-achieving student, a mathematically able but unmotivated student, and a student with neither strong ability nor motivation. For each profile, a subset of eight tasks was analysed. Figures 2–4 summarise the results for the three profiles using a three-level scale (0–2). Values of 0, 1, and 2 indicate low, partial, and high presence of the respective criterion (role alignment or didactic usefulness).

Across the three profiles, the generated solution attempts differed in both role alignment and didactic usefulness. Figures 2–4 present these results graphically. In routine lower secondary tasks from categories A and B, the model often produced solution attempts that matched the intended profiles and contained

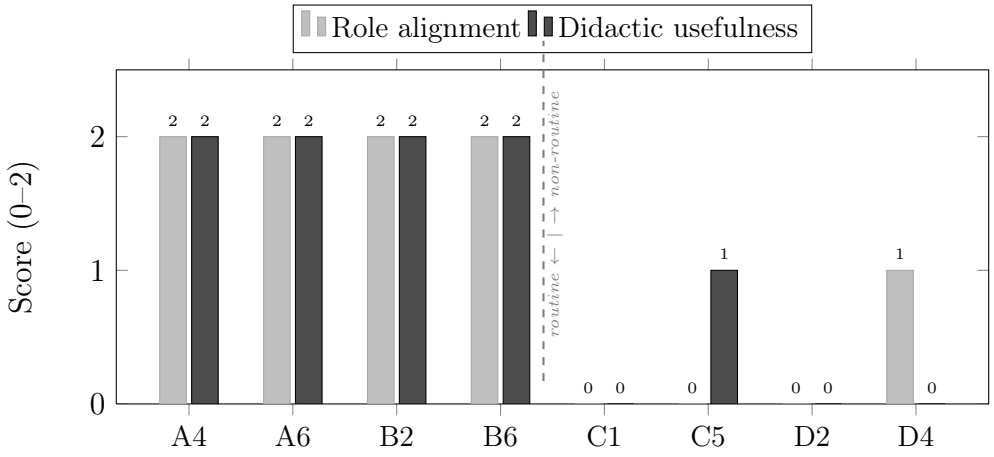

Fig. 2: Experiment 2 — Diligent but low-achieving student: role alignment and didactic usefulness across tasks (0–2 scale). The dashed line separates routine tasks (Categories A, B) from non-routine and construction tasks (Categories C, D)

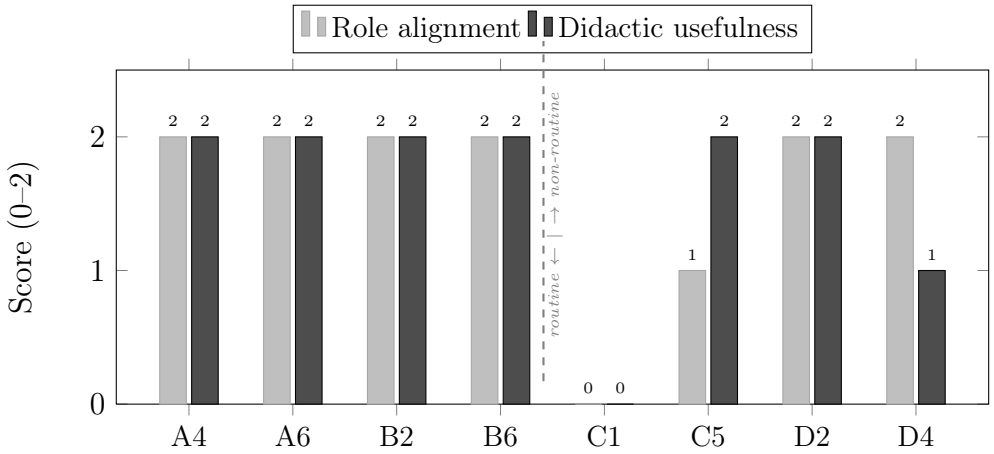

Fig. 3: Experiment 2 — Able but unmotivated student: role alignment and didactic usefulness across tasks (0–2 scale). The dashed line separates routine tasks from non-routine and construction tasks

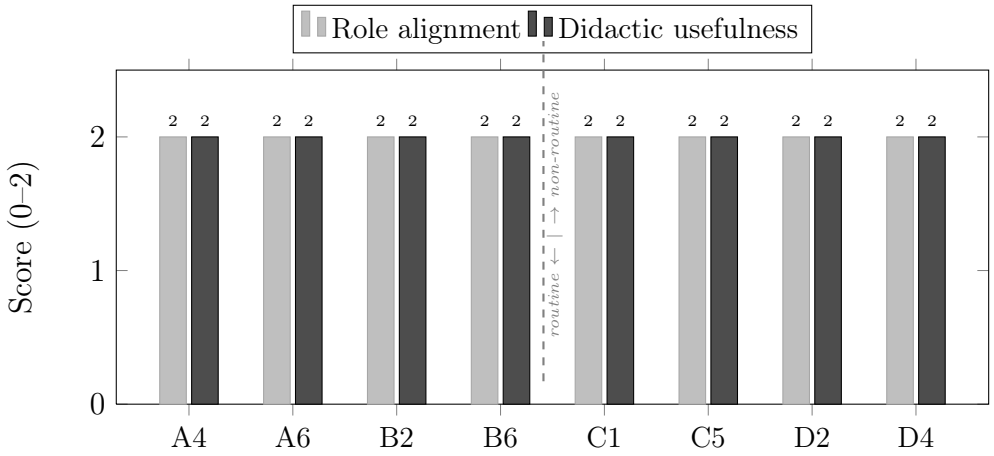

Fig. 4: Experiment 2 — Struggling student: role alignment and didactic usefulness across tasks (0–2 scale). Unlike the other two profiles, both criteria remain at the maximum level across all tasks, including non-routine and construction tasks

plausible student-like errors or omissions. Such errors were treated as didactically valuable, as they can serve as material for *a priori* analysis (e.g., anticipating typical pitfalls and planning teacher prompts).

However, the results also show clear limits of the simulation approach. For the diligent but low-achieving profile, role alignment deteriorated in several non-routine and construction tasks (categories C

and D), where the model sometimes produced solutions that were too advanced or too systematically argued to be plausible for the intended student type. A similar tendency was observed for the able-but-unmotivated profile in the most demanding tasks, although some responses still reflected the intended profile through shortcuts, missing checks, or inconsistent reasoning.

In contrast, the struggling student profile showed consistently high role alignment across all selected tasks, with both criteria reaching the maximum score of 2 for every task including the non-routine and construction categories. The corresponding solution attempts frequently included incomplete reasoning, breakdowns in procedure, or missing result checking, which were considered didactically useful as representations of typical difficulties that teachers may encounter. This uniform pattern stands in marked contrast to the other two profiles and confirms that the model more reliably simulates lower levels of performance than higher-but-constrained levels.

Detailed qualitative analyses of individual solution attempts for all three profiles are provided in Appendix F.

4.2.1 Qualitative illustrations of role alignment and didactic usefulness

The following vignettes illustrate both successful and unsuccessful student simulations across profiles and task types.

Successful simulation—diligent but low-achieving student, Task A4 The model produced a solution attempt in which the student correctly identified a common denominator but made a typical sign error when removing parentheses (failing to distribute the minus sign). The condition $x \neq 1$ was omitted entirely. The language was diligent but mechanical, consistent with the profile. This type of error—a minus sign in front of parentheses—is a well-documented difficulty at the lower secondary level, making this output directly useful for a priori analysis.

Successful simulation—able but unmotivated student, Task B2 The model simulated a student who quickly added the two discounts ($20\% + 10\% = 30\%$) and treated the discounted price as 70% of the original, ignoring the fact that percentage discounts are applied sequentially and multiplicatively. The resulting answer was incorrect, but the error pattern—additive composition of percentages—is a frequent and well-documented student misconception. The output clearly matched the intended profile (fast, intuitive, no verification) and provided useful diagnostic material.

Failed simulation—diligent but low-achieving student, Task C1 For this olympiad problem requiring prime factorisation and comparison of exponents, the model produced a systematic, complete, and error-free solution that would be unrealistic even for mathematically gifted ninth-grade students, let alone for the intended profile of a diligent but untalented student. The solution employed parametrisations and systematic minimisation procedures characteristic of university-level work. This output entirely missed the intended student profile and offered no didactic value for anticipating the difficulties of the targeted student type.

Consistently successful—struggling student, all tasks The struggling student profile showed the highest role alignment across all tasks. For Task D4 (rhombus construction), the model produced a superficial attempt without analysis or verification, with a key error of confusing the diagonal's length with its half. For Task C1 (olympiad problem), the model responded with a simple numerical substitution attempt followed by resignation. These outputs plausibly represent the kinds of difficulties teachers encounter with struggling students and can support anticipatory planning.

5 Discussion

The present study examined the didactic potential of a large language model for analysing mathematical problem solving processes and supporting *a priori* didactic analysis in lower secondary mathematics education. The focus was deliberately not on mathematical performance or optimisation, but exclusively on the structure, completeness, and didactic relevance of generated solution attempts.

The present findings are consistent with previous studies emphasising the supportive yet limited role of LLMs in mathematics education. Similar to the observations of Pepin et al. (2025), our results suggest that LLMs can assist with instructional planning and reflection. In line with Kasneci et al. (2023), the study also reinforces concerns that uncritical reliance on AI-generated solutions may reduce engagement in verification and deeper mathematical reasoning.

5.1 Verification as a fragile phase of problem solving

Results from Experiment 1 show that explicit result verification represents a particularly fragile phase of the problem solving process. In the baseline condition, verification was frequently omitted despite the presence of correct interpretations, coherent solution procedures, and correct final results. This finding is consistent with long-standing observations in mathematics education that students often stop once a result is obtained and do not spontaneously engage in checking or reflection (Kontorovich, 2019).

The absence of verification was not evenly distributed across task types. It occurred most frequently in ill-structured word problems and geometric construction tasks, where correctness cannot be reduced to a single numerical outcome. These task types require solvers to reflect on assumptions, constraints, and the adequacy of the chosen interpretation. The results suggest that, without explicit prompting, even a highly capable language model tends to prioritise producing a plausible solution over critically examining its validity.

When guided by a structured Pólya-inspired framework, explicit verification was present in all tasks across all categories. We acknowledge that this finding may appear predictable: if the prompt explicitly requests verification, it is not surprising that the model produces it. However, the empirical demonstration of this contrast is valuable for two reasons. First, it quantifies the gap between prompted and unprompted conditions across different task types, which has not been documented for lower secondary mathematics tasks in Czech. Second, it has direct practical implications for teachers: structured prompting is not merely a theoretical recommendation but a necessary condition for obtaining didactically complete outputs.

This sensitivity to prompting is consistent with recent findings on how users elicit higher-quality reasoning from LLMs. Urban et al. (2025) report that successful users of ChatGPT often include explicit evaluation criteria in their prompts. Our results suggest a similar mechanism: the Pólya-inspired prompt did not change the model's mathematical ability, but it consistently elicited verification, which was frequently missing in the baseline condition. This indicates that structured prompting can serve as a practical scaffolding tool for activating checking behaviour.

5.2 LLMs as tools for analysing problem solving processes

The findings of Experiment 1 suggest that large language models can be used as analytical tools for examining problem solving processes. The baseline condition revealed a systematic pattern: the model consistently prioritised procedural completion over explicit reflection or result checking. As the illustrative examples in Section 4.1.3 show, even when the model identified ambiguity (as in Task B1) or produced elegant constructions (as in Task D4), it did not spontaneously engage in verification.

At the same time, the results underscore the importance of prompt design. Structured prompts did not merely change the form of the output but fundamentally altered the completeness of the problem solving process. This sensitivity to prompting highlights both the potential and the limitation of LLMs: they can support didactic analysis, but they do not autonomously prioritise pedagogically desirable behaviours such as verification or reflection.

5.3 Simulating student thinking: Didactic potential and limits

Results from Experiment 2 show that, for routine and moderately demanding tasks (Categories A and B), the model often produced profile-consistent solution attempts containing plausible student-like errors. The diligent student's sign error in Task A4, the gifted student's additive percentage misconception in Task B2, and the struggling student's procedural breakdown in Task B6 all represent documented patterns that teachers may encounter in classroom practice. From a didactic perspective, such imperfect solutions are valuable because they make typical breakdowns in problem solving processes visible and can therefore serve as concrete material for anticipatory instructional planning.

This aligns with research on LLM-based simulated learners, which suggests that simulated student responses can support practice-based work with student errors and misconceptions (Zhuang & Zhang, 2025). The present study extends this perspective by showing that the plausibility of such simulations is strongly task-dependent.

However, the results also demonstrate clear limits. For non-routine and construction tasks, the model frequently produced solutions that were too advanced for the intended profiles. This was most striking for the diligent but low-achieving profile: in Tasks C1 and D2, the model generated solutions at a level of sophistication that would be unrealistic even for mathematically gifted students (see Section 4.2.1). This deterioration of role fidelity occurred precisely in the task categories where anticipatory teacher knowledge is most needed, consistent with the finding of McGalliard and Otten (2025) that AI performance patterns differ between routine and non-routine problems.

The struggling student profile was the exception: it showed consistently high role alignment across all tasks, suggesting that the model more reliably simulates lower levels of performance than higher-but-constrained levels. This asymmetry has practical implications for the use of LLM-generated solutions in a priori analysis.

These findings are consistent with broader limitations identified in the literature. Zhuang and Zhang (2025) observe that simulated learners often adapt unrealistically quickly to instructional input and may respond in a compliant and passive manner, without the hesitation, partial attempts, or follow-up questions that characterise genuine student behaviour. Such limitations restrict the extent to which LLM-based simulations can approximate the dialogical dynamics of classroom diagnosis.

Role fidelity is also a matter of linguistic register. Even when instructed to adopt the voice of a lower secondary student, LLMs may produce language that is overly formal or "teacher-like" (Zhuang & Zhang, 2025). Pando and León (2025) similarly found that achieving age-appropriate language required iterative prompting with explicit output constraints, rather than role prompts alone. This is consistent with our observation that simulated profiles sometimes drifted toward expert-like discourse.

More broadly, LLM-generated responses may exhibit what Guerra et al. (2025) term an *illusion of understanding*: outputs that appear coherent while lacking deeper conceptual grounding. In the context of student simulation, this may lead to overestimating the cognitive realism of generated solutions. One contributing factor is that student profiles are specified only narratively, leaving room for the model to improvise competence. Zhang et al. (2025) suggest that structured representations of learner proficiency (e.g., a "Skill-Tree") can help constrain such drift, which may be a promising direction for improving simulation fidelity in non-routine tasks.

Finally, logical inconsistency across tasks remains a concern. In our data, the diligent but low-achieving profile alternated between elementary errors in routine tasks and unexpectedly sophisticated reasoning in olympiad problems (compare Tasks A4 and C1 in Figure 2). While some degree of inconsistency may resemble authentic student behaviour, the abrupt shifts observed here exceeded what would be plausible for a single student and require cautious interpretation by the teacher.

5.4 Implications for *a priori* didactic analysis

Taken together, the results of both experiments suggest a specific and circumscribed didactic role for large language models in mathematics education: supporting *a priori* didactic analysis by making potential solution paths, typical errors, and missing problem solving phases visible before instruction. In this role, the value of LLM-generated solutions lies not in their correctness or realism, but in their capacity to stimulate teachers' anticipatory thinking.

Building on research showing that error anticipation can be developed through practice (Stannard & Foster, 2025), the present findings suggest that LLM-generated solutions may serve as a readily available source of varied examples for such anticipatory work. In this sense, LLMs can reduce the initial time and effort required to generate and compare alternative student-like solution attempts during lesson preparation. At the same time, the limitations of simulation highlight the need for critical AI literacy: teachers should treat AI-generated outputs as prompts for professional judgement rather than as proxies for real student thinking.

The present findings further underscore the instructional importance of explicit result verification. If even a language model with strong procedural capabilities omits verification unless explicitly prompted, this reinforces the need to foreground result checking as an explicit instructional goal, particularly in ill-structured and open-ended tasks (Schoenfeld, 1985).

These implications also reinforce the importance of critical AI literacy among teachers. In their systematic review of ChatGPT in school mathematics education, Turmuzi et al. (2026) emphasise that AI integration requires teachers to remain the primary facilitators who critically evaluate and contextualise AI outputs. Our results provide a concrete illustration of why this is necessary: the model not only drifted into unrealistic student simulations in cognitively demanding tasks (e.g., producing expert-like constructions for a low-achieving profile), but also tended to omit explicit result verification unless it was systematically prompted. Uncritical adoption of such outputs could therefore lead teachers to anticipate incorrect difficulties, overlook missing checking behaviour, and design misguided instructional responses. Consequently, LLM-generated solution attempts should be treated as reflective artefacts for professional judgement rather than as reliable proxies for authentic student thinking.

6 Methodological limitations

The present study was designed as a qualitative demonstration study, and its findings should be interpreted with this scope in mind. The aim was not to establish generalisable claims about the performance

of large language models but to provide a transparent and analytically grounded examination of their problem solving behaviour under specific conditions relevant to mathematics education.

Several limitations of the design need to be acknowledged. First, the experiments were conducted using a single large language model and a single model version. The observed behaviour may therefore reflect characteristics of this particular model rather than properties shared by all contemporary LLMs.

Second, all interactions were conducted in Czech, which represents a less-resourced language in the training data of most LLMs. Language-related factors may thus have influenced the completeness and structure of the generated solutions.

Third, the analysis relied on qualitative coding performed by the authors. While explicit evaluation criteria were used to support consistency, no claims are made regarding inter-rater reliability. The purpose of the analysis was exploratory and illustrative rather than statistical. Similarly, the student profiles used in Experiment 2 were intentionally simplified and do not correspond to diagnostic categories or individual learners.

Fourth, the empirical basis consists of outputs from a single model and a limited number of generated responses. While LLM outputs could in principle be produced in larger quantities, the qualitative demonstration design required detailed multi-criteria analysis of each solution, which limited the feasible sample size. Additionally, simulations conducted through text-only interfaces necessarily omit non-verbal and multimodal information that teachers routinely use when interpreting student difficulties (Guerra et al., 2025).

Fifth, the development of AI technologies is extremely rapid, and the findings reflect the behaviour of a specific model version (Gemini 2.5 Pro, September 2025). Future model updates may alter the observed patterns. However, the methodological framework and evaluation criteria developed in this study remain applicable to future models, which is one of the intended contributions of the paper.

7 Future research

Future research could extend this work by systematically comparing different models and languages, involving independent coders, and examining how teachers interpret and use LLM-generated solutions in authentic lesson planning contexts. Multi-model comparative studies using larger samples with streamlined coding procedures would help establish the generalisability of the observed patterns. A particularly important direction is the development of multimodal and interactionally richer simulated learners that better approximate classroom diagnosis beyond text-only outputs. Future work could also explore whether structured learner models (e.g., Skill-Tree representations) improve the stability of student simulations, particularly in non-routine and construction tasks. Another promising direction is to test iterative prompting strategies that explicitly constrain linguistic register and mathematical tool use to match a targeted age group.

8 Conclusion

This study explored the didactic potential of a large language model as a tool for analysing mathematical problem solving processes and supporting *a priori* didactic analysis in lower secondary mathematics education. The focus was deliberately placed on the structure and completeness of solution processes and on their didactic relevance, rather than on mathematical performance or optimisation.

Experiment 1 showed that explicit result verification represents a fragile phase of problem solving that is often omitted unless it is explicitly prompted. The contrast between the baseline and the Pólya-guided conditions suggests that the absence of verification should not be interpreted as a lack of capability but rather as a tendency not to activate this phase spontaneously. From a didactic perspective, this finding reinforces the importance of explicitly foregrounding verification and reflection in instructional design, particularly in ill-structured tasks and construction problems.

Experiment 2 demonstrated that a large language model can generate solution attempts that plausibly reflect certain student-like behaviours, including typical errors, omissions, and incomplete strategies. At the same time, clear limits emerged. For non-routine and olympiad-type problems, the model tended to produce highly sophisticated or expert-level solutions even when prompted to simulate an average or low-achieving student. As a result, profile alignment deteriorated precisely in tasks where anticipatory teacher knowledge is most needed. These findings suggest that LLM-based student simulation is strongly task-sensitive: it can support anticipatory work for routine tasks, but its use for mathematically demanding problems requires particular caution.

Overall, the study suggests that large language models may serve as reflective tools for didactic analysis when used critically and transparently. Their value lies not in replacing teacher expertise or

evaluating student performance, but in making aspects of problem solving visible and open to reflection. Further research is needed to examine how such tools can be integrated into instructional planning in a responsible and pedagogically meaningful way.

Declarations

Funding: The authors received no specific funding for this work.
Conflict of interest: The authors declare no conflict of interest.
Generative AI use: The authors used ChatGPT to support English-language editing and improve clarity of phrasing. All scientific content, study design, analysis, interpretation, and final wording were reviewed and approved by the authors.

References

Almarashdi, H. S., Jarrah, A. M., Abu Khurma, O., & Gningue, S. M. (2024). Unveiling the potential: A systematic review of ChatGPT in transforming mathematics teaching and learning. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(12), em2555. https://doi.org/10.29333/ejmste/15739

Brousseau, G. (1997). *Theory of Didactical Situations in Mathematics*. N. Balacheff, M. Cooper, R. Sutherland, & V. Warfield (Eds.) Kluwer Academic Publishers.

Brousseau, G., & Gibel, P. (2005). Didactical handling of students' reasoning processes in problem solving situations. *Educational Studies in Mathematics*, *59*(1–3), 13–58. https://doi.org/10.1007/s10649-005-2532-y

Carlson, M. P., & Bloom, I. (2005). The cyclic nature of problem solving: An emergent multidimensional problem-solving framework. *Educational Studies in Mathematics*, *58*(1), 45–75. https://doi.org/10.1007/s10649-005-0808-x

Cifarelli, V. V., & Cai, J. (2005). The evolution of mathematical explorations in open-ended problem-solving situations. *Journal of Mathematical Behavior*, *24*(3–4), 302–324. https://doi.org/10.1016/j.jmathb.2005.09.007

Guerra, E., Peña, M., & Araya, R. (2025). The inevitable and unpredictable role of large language models in education: A commentary on Huettig and Christiansen (2024). *Cognitive Science*, *49*, e70105. https://doi.org/10.1111/cogs.70105

Hoth, J., Larrain, M., & Kaiser, G. (2022). Identifying and dealing with student errors in the mathematics classroom: Cognitive and motivational requirements. *Frontiers in Psychology*, *13*, 1057730. https://doi.org/10.3389/fpsyg.2022.1057730

Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development*, *45*(1), 65–94. https://doi.org/10.1007/BF02299613

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

Kontorovich, I. (2019). Why do students not check their solutions to mathematical problems? A field-based hypothesis on epistemological status. *International Journal of Mathematical Education in Science and Technology*, *50*(7), 1050–1062. https://doi.org/10.1080/0020739X.2019.1650304

Liu, J., Huang, Z., Liu, Q., Ma, Z., Zhai, C., & Chen, E. (2025). Knowledge-centered dual-process reasoning for math word problems with large language models. *IEEE Transactions on Knowledge and Data Engineering*, *37*(6), 3457–3471. https://doi.org/10.1109/TKDE.2025.3556367

Malik, R., Abdi, D., Wang, R., & Demszky, D. (2025). Scaffolding middle school mathematics curricula with large language models. *British Journal of Educational Technology*, *56*(3), 999–1027. https://doi.org/10.1111/bjet.13571

Matematická olympiáda. (2021). Czech Mathematical Olympiad: 71st year, category Z9, Round I (problem set). Retrieved February 8, 2026, from https://www.matematickaolympiada.cz/media/3492117/z71i-9.pdf

Matematická olympiáda. (2022a). Czech Mathematical Olympiad: 72nd year, category Z9, Round I (problem set). Retrieved February 8, 2026, from https://www.matematickaolympiada.cz/media/3494370/z72i-9.pdf

Matematická olympiáda. (2022b). Czech Mathematical Olympiad: 72nd year, category Z9, Round II (problem set). Retrieved February 8, 2026, from https://www.matematickaolympiada.cz/media/3516462/z72ii-9-zr.pdf

Matematická olympiáda. (2022c). Czech Mathematical Olympiad: 72nd year, category Z9, Round III (problem set). Retrieved February 8, 2026, from https://www.matematickaolympiada.cz/media/3516464/z72iii-9-zr.pdf

Matematická olympiáda. (2024a). Czech Mathematical Olympiad: 74th year, category Z9, Round II (problem set). Retrieved February 8, 2026, from https://www.matematickaolympiada.cz/media/3828323/z74ii-zadani.pdf

Matematická olympiáda. (2024b). Czech Mathematical Olympiad: 74th year, category Z9, Round III (problem set). Retrieved February 8, 2026, from https://www.matematickaolympiada.cz/media/3824271/z74iii-9.pdf

McGalliard, W., & Otten, S. (2025). AI Responses to Challenging Problems and Educator Responses to AI Availability. *Digital Experiences in Mathematics Education*, *11*, 319–332. https://doi.org/10.1007/s40751-024-00167-4

Nováková, H. (2013). Analýza a priori jako součást přípravy učitele na výuku. *Scientia in educatione*, *4*(2), 20–51. https://doi.org/10.14712/18047106.70

Ocak, C., Kopcha, T. J., Hodges, C. B., Sadik, O., & Ozogul, G. (2025). How artificial intelligence will reshape education: Conversations with the educational technology community. *TechTrends*, *70*, 330–345. https://doi.org/10.1007/s11528-025-01149-y

Pando, M., & León, M. (2025). Mathematics disciplinary literacy: A case study of a bilingual teacher's interaction with ChatGPT. *Language and Education*, *40*(4), 980–999. https://doi.org/10.1080/09500782.2025.2601055

Pankow, L., Kaiser, G., Busse, A., König, J., Blömeke, S., Hoth, J., & Döhrmann, M. (2016). Early career teachers' ability to focus on typical students' errors in relation to the complexity of a mathematical topic. *ZDM Mathematics Education*, *48*, 55–67, https://doi.org/10.1007/s11858-016-0763-2

Papadopoulos, I., & Dagdilelis, V. (2008). Students' use of technological tools for verification purposes in geometry problem solving. *The Journal of Mathematical Behavior*, *27*(4), 311–325. https://doi.org/10.1016/j.jmathb.2008.11.001

Pepin, B., Buchholtz, N., & Salinas-Hernández, U. (2025). A scoping survey of ChatGPT in mathematics education. *Digital Experiences in Mathematics Education*, *11*, 9–41. https://doi.org/10.1007/s40751-025-00172-1

Pólya, G. (1945). *How to solve it*. Princeton University Press.

Rendl, M., Vondrová, N., Hříbková, L., Jirotková, D., Kloboučková, J., Kvasz, L., Páchová, A., Pavelková, I., Smetáčková, I., Tauchmanová, E., & Žalská, J. (2013). *Kritická místa matematiky na základní škole očima učitelů*. Univerzita Karlova, Pedagogická fakulta.

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Academic Press.

Stannard, A., & Foster, C. (2025). Secondary school mathematics teachers' accuracy at predicting student errors. *School Science and Mathematics*. https://doi.org/10.1111/ssm.18404

Suyitno, A., Suyitno, H., Rochmad, & Dwijanto. (2018). Use of open-ended problems as the basis for the mathematical creativity growth disclosure of student. *Journal of Physics: Conference Series*, *983*(1), 012110. https://doi.org/10.1088/1742-6596/983/1/012110

Turmuzi, M., Azmi, S., & Kertiyani, N. M. I. (2026). ChatGPT in school mathematics education: A systematic review of opportunities, challenges, and pedagogical implications. *Teaching and Teacher Education*, *170*. https://doi.org/10.1016/j.tate.2025.105286

Urban, M., Lukavský, J., Brom, C., Hein, V., Svacha, F., Děchtěrenko, F., & Urban, K. (2025). Prompting for creative problem-solving: A process-mining study. *Learning and Instruction*, *99*. https://doi.org/10.1016/j.learninstruc.2025.102156

Zhang, X., Zhang, C., Sun, J., & Xiao, J. (2025). EduPlanner: LLM-based multi-agent systems for customized and intelligent instructional design. *IEEE Transactions on Learning Technologies*, 1–11. https://doi.org/10.1109/TLT.2025.3561332

Zhuang, Y., & Zhang, S. (2025). Integrating ChatGPT in mathematics teacher education: AI-based simulation role-playing to support practice-based teaching. *International Journal of Artificial Intelligence in Education*, *35*, 3873–3895. https://doi.org/10.1007/s40593-025-00519-0

Appendix A: Pólya-inspired structured prompt (original Czech wording)

This appendix provides the original Czech wording of the structured prompt used in Experiment 1 to guide the model's problem solving process according to a Pólya-inspired framework.

```
Předložím ti 24 matematických úloh. Po celou dobu se řiď tím, že řešení má mít čtyři
části:

1) Porozumění zadání: uveď, co je neznámé, jaká jsou data a podmínky, a jaké
   předpoklady je nutné přijmout.
2) Plán: navrhni jasnou strategii řešení a krátce vysvětli, proč je vhodná.
3) Provedení: kroky očísluj; vyhýbej se skokům; dbej na ekvivalentní úpravy.
4) Pohled zpět: proveď zkoušku výsledku nebo alternativní kontrolu; u konstrukčních
   úloh ověř, proč postup splňuje požadované vlastnosti.

Řešení vždy napiš v češtině. Drž se těchto čtyř částí u každé úlohy, bez ohledu na
její typ. Nepřidávej jiné nadpisy ani vysvětlení, pouze jednotlivé části v uvedeném
pořadí.
```

English translation:

```
I will present you with 24 mathematical tasks. Throughout the entire session, follow
the rule that each solution must consist of four parts:

1) Understanding the problem: state what is unknown, what data and conditions are
   given, and what assumptions need to be made.
2) Plan: propose a clear solution strategy and briefly explain why it is
   appropriate.
3) Execution: number the steps; avoid jumps; ensure equivalent transformations.
4) Looking back: verify the result or perform an alternative check; for construction
   tasks, verify why the procedure satisfies the required properties.

Always write the solution in Czech. Follow these four parts for every task,
regardless of its type. Do not add any other headings or explanations, only the
individual parts in the order given.
```

Appendix B: Task set

This appendix provides the full wording of all 24 mathematical tasks used in the study, in their original Czech form.

B.1 Category 1: Well-structured algebraic tasks

A1 Vyřeš rovnici $3x - 7 = 2x + 5$.

A2 Vyřeš soustavu $2x + 3y = 7$; $-x + y = 4$.

A3 Vyřeš rovnici $x^2 - 5x + 6 = 0$.

A4 Zjednoduš výraz $\dfrac{x+2}{x-1} - \dfrac{x-3}{2}$.

A5 Vyřeš nerovnici $-2(3 - x) \geq 4x - 10$.

A6 Vyřeš rovnici $\sqrt{x+5} = x - 1$.

B.2 Category 2: Ill-structured word problems

B1 Auto jelo rychlostí 60 km/h celkem 2 hodiny včetně přestávky. Kolik kilometrů ujelo?

B2 Na vývěsce je: „Sleva 20 % – nyní 1 600 Kč.“ Na účtence je navíc položka „věrnostní sleva 10 %“. Jaká mohla být původní cena?

B3 Smícháme roztok s 10% koncentrací a roztok s 26% koncentrací a vznikne 500 ml roztoku s 18% koncentrací. Kolik je v něm z každého?

B4 Školní výlet se má uskutečnit autobusem. Autobus má 50 míst, ale není jisté, kolik učitelů pojede. Víme jen, že celkový počet účastníků je 46 žáků a několik učitelů. Cena pronájmu autobusu je 9 000 Kč. Navrhni dvě rozumné interpretace, jak se může cena rozpočítat mezi účastníky.

B5 Válcová nádrž má průměr cca 3,2 m a výšku 2,5 m. Kolik vody pojme?

B6 Dvě čerpadla naplní nádrž společně za 6 h. První čerpadlo samo by to zvládlo asi dvojnásobně dlouho. Urči časy obou čerpadel.

B.3 Category 3: Olympiad problems

These tasks originate from publicly accessible materials intended for educational use.

C1 Najděte nejmenší kladná celá čísla a a b, pro která platí $7a^3 = 11b^5$.

(MO 72nd year, Z9-I-4; 2022a)

C2 Jana si vymyslela 2022místné číslo a jeho ciferný součet pošeptala Petrovi. Petr vypočítal ciferný součet čísla, které mu sdělila Jana, a výsledek pošeptal Zuzce. Zuzka též vypočítala ciferný součet čísla, které dostala od Petra, a výsledek, jímž bylo dvojmístné číslo, pošeptala Adamovi. Adam provedl totéž s číslem od Zuzky a vyšel mu ciferný součet 1. Která čísla mohl šeptat Petr Zuzce? Určete všechny možnosti.

(MO 71st year, Z9-I-2; 2021)

C3 Najděte všechna dvojmístná přirozená čísla, která mají následující vlastnost: Když před číslo připíšeme součin jeho první číslice a jeho první číslice zvětšené o 1, dostaneme druhou mocninu původního čísla.

(MO 74th year, Z9-II-1; 2024a)

C4 Najděte všechna čtyřmístná čísla, která mají přesně pět čtyřmístných a přesně devět jednomístných dělitelů.

(MO 72nd year, Z9-II-1; 2022b)

C5 Je dán bod B a rovnostranný trojúhelník se stranami délky 1 cm. Vzdálenosti bodu B od dvou vrcholů tohoto trojúhelníku jsou 2 cm a 3 cm. Vypočítejte vzdálenost bodu B od třetího vrcholu trojúhelníku.

(MO 72nd year, Z9-III-2; 2022c)

C6 Najděte všechny dvojice přirozených čísel a a b, pro které platí

$$7a + 4b + 74 = a \cdot b.$$

(MO 74th year, Z9-III-3; 2024b)

B.4 Category 4: Geometric construction tasks

D1 Sestrojte trojúhelník ABC, jestliže $b = |AC|$, $c = |AB|$ a výška v_a na stranu $a = |BC|$.

D2 Sestrojte trojúhelník ABC, je-li dána strana $a = |BC|$, vnitřní úhel $\beta = \angle ABC$ a těžnice t_b na stranu $b = |AC|$.

D3 Sestrojte všechny tečny z bodu P ke kružnici $k(S, r)$ a body dotyku označte T_1, T_2.

D4 Sestrojte kosočtverec $ABCD$, jehož úhlopříčky mají délky $|AC| = e$ a $|BD| = f$.

D5 Sestrojte lichoběžník $ABCD$, má-li základny $|AB| = a$, $|CD| = c$, rameno $|AD| = d$ a výšku v.

D6 Sestrojte pravidelný šestiúhelník $A_1A_2A_3A_4A_5A_6$ vepsaný do kružnice $k(S, r)$.

English translation:

B.5 Category 1: Well-structured algebraic tasks

A1 Solve the equation $3x - 7 = 2x + 5$.

A2 Solve the system $2x + 3y = 7$; $-x + y = 4$.

A3 Solve the equation $x^2 - 5x + 6 = 0$.

A4 Simplify the expression $\dfrac{x+2}{x-1} - \dfrac{x-3}{2}$.

A5 Solve the inequality $-2(3 - x) \geq 4x - 10$.

A6 Solve the equation $\sqrt{x+5} = x - 1$.

B.6 Category 2: Ill-structured word problems

B1 A car travelled at a speed of 60 km/h for a total of 2 hours including a break. How many kilometres did it cover?

B2 A sign reads: "20% off – now 1 600 Kč." The receipt also shows an item "loyalty discount 10%". What could the original price have been?

B3 A solution with 10% concentration is mixed with a solution with 26% concentration to produce 500 ml of a solution with 18% concentration. How much of each is in the mixture?

B4 A school trip is to be taken by bus. The bus has 50 seats, but it is not certain how many teachers will go. We only know that the total number of participants is 46 students and several teachers. The bus rental costs 9,000 CZK. Propose two reasonable interpretations of how the cost could be divided among the participants.

B5 A cylindrical tank has a diameter of approximately 3.2 m and a height of 2.5 m. How much water can it hold?

B6 Two pumps fill a tank together in 6 hours. The first pump alone would take approximately twice as long. Determine the times for both pumps.

B.7 Category 3: Olympiad problems

C1 Find the smallest positive integers a and b such that $7a^3 = 11b^5$.

(MO 72nd year, Z9-I-4; 2022a)

C2 Jana invented a 2022-digit number and whispered its digit sum to Petr. Petr calculated the digit sum of the number Jana told him and whispered the result to Zuzka. Zuzka also calculated the digit sum of the number she received from Petr, and the result, which was a two-digit number, she whispered to Adam. Adam did the same with Zuzka's number and obtained a digit sum of 1. Which numbers could Petr have whispered to Zuzka? Determine all possibilities.

(MO 71st year, Z9-I-2; 2021)

C3 Find all two-digit natural numbers with the following property: When the product of its first digit and its first digit increased by 1 is written in front of the number, the result is the square of the original number.

(MO 74th year, Z9-II-1; 2024a)

C4 Find all four-digit numbers that have exactly five four-digit divisors and exactly nine one-digit divisors.

(MO 72nd year, Z9-II-1; 2022b)

C5 A point B and an equilateral triangle with sides of length 1 cm are given. The distances from point B to two vertices of this triangle are 2 cm and 3 cm. Calculate the distance from point B to the third vertex of the triangle.

(MO 72nd year, Z9-III-2; 2022c)

C6 Find all pairs of natural numbers a and b for which

$$7a + 4b + 74 = a \cdot b.$$

(MO 74th year, Z9-III-3; 2024b)

B.8 Category 4: Geometric construction tasks

D1 Construct triangle ABC, given $b = |AC|$, $c = |AB|$ and the altitude v_a to side $a = |BC|$.

D2 Construct triangle ABC given side $a = |BC|$, interior angle $\beta = \angle ABC$ and the median t_b to side $b = |AC|$.

D3 Construct all tangent lines from point P to the circle $k(S, r)$ and label the points of tangency T_1, T_2.

D4 Construct a rhombus $ABCD$ whose diagonals have lengths $|AC| = e$ and $|BD| = f$.

D5 Construct a trapezoid $ABCD$ given bases $|AB| = a$, $|CD| = c$, leg $|AD| = d$ and altitude v.

D6 Construct a regular hexagon $A_1A_2A_3A_4A_5A_6$ inscribed in the circle $k(S, r)$.

Appendix C: Student profile prompts (original Czech wording)

This appendix provides the original Czech wording of the prompts used in Experiment 2 to specify the simulated student profiles.

C.1 Profile A: Diligent But Untalented Student

```
Předložím ti 8 matematických úloh. Po celou dobu se drž role 15letého žáka 9. třídy,
který pracuje pilně, ale čistě rutinně. Má snahu, ale nemá nadání. Opírá se o naučené
postupy, bez hlubšího porozumění. Piš 2--4 věty „přemýšlím nahlas" a pak řešení krok
za krokem. Pokud se ti omylem povede zajímavý nápad, zahoď ho a vrať se k mechanice.
Typické chyby, které udržuj: chybějící znaménko, nevhodné krácení zlomků, opomenutí
podmínek, přehazování členů rovnice bez vysvětlení. Bude následovat 8 úloh jedna po
druhé.
```

English translation:

```
I will present you with 8 mathematical tasks. Throughout the entire session, stay in
the role of a 15-year-old 9th-grade student who works diligently but in a purely
routine manner. This student tries hard but lacks talent. They rely on memorised
procedures without deeper understanding. Write 2--4 sentences of ‘‘thinking aloud’’
and then solve step by step. If you accidentally come up with an interesting idea,
discard it and return to mechanical work. Typical errors to maintain: missing sign,
inappropriate cancellation of fractions, overlooking conditions, rearranging equation
terms without explanation. 8 tasks will follow one after another.
```

C.2 Profile B: Gifted But Lazy Student

```
Předložím ti 8 matematických úloh. Po celou dobu se drž role 15letého žáka 9. třídy,
který řeší úlohy podle prvního nápadu. Má nadání, ale nemá snahu. Rozhoduje se
rychle, spoléhá na intuici a analogie. Piš 2--4 věty „přemýšlím nahlas", pak řešení
krok za krokem tak, jak by ho žák řešil. Nepřidávej přesné důkazy ani podrobné
výpočty navíc. Typické chyby, které udržuj: chybějící kontrola výsledku, ukvapené
dosazení, analogie „jako minule" i tam, kde neplatí, odhady „od oka" vedoucí
k nesprávnému výsledku. Bude následovat 8 úloh jedna po druhé.
```

English translation:

```
I will present you with 8 mathematical tasks. Throughout the entire session, stay in
the role of a 15-year-old 9th-grade student who solves tasks based on their first
idea. This student has talent but lacks effort. They decide quickly, relying on
intuition and analogies. Write 2--4 sentences of „thinking aloud", then solve step by
step as the student would. Do not add precise proofs or detailed extra calculations.
Typical errors to maintain: missing result verification, hasty substitution, „same as
last time" analogies applied where they do not hold, rough estimates leading to
incorrect results. 8 tasks will follow one after another.
```

C.3 Profile C: Unmotivated Struggling Student

```
Předložím ti 8 matematických úloh. Po celou dobu se drž role 15letého žáka 9. třídy,
který má dlouhodobé potíže v matematice. Nemá ukotvené základy z předchozích tříd,
nemá nadání ani snahu. Piš 2--4 věty „přemýšlím nahlas", pak stručný pokus o řešení.
Řešení nemusíš vždy dokončit. Typické chyby, které udržuj: nepochopení zadání,
vynechání údajů, přerušení výpočtu, záměna základních postupů (zlomky, desetinná
čísla), chybné pořadí kroků u konstrukcí. Pokud se ti podaří začít správně, projeví
se nejistota a nepokračuješ správně. Bude následovat 8 úloh jedna po druhé.
```

English translation:

```
I will present you with 8 mathematical tasks. Throughout the entire session, stay in
the role of a 15-year-old 9th-grade student who has long-term difficulties in
mathematics. This student lacks solid foundations from previous grades and has
neither talent nor effort. Write 2--4 sentences of „thinking aloud", then a brief
attempt at a solution. You do not always need to finish the solution. Typical errors
to maintain: misunderstanding the problem statement, omitting data, interrupting the
computation, confusing basic procedures (fractions, decimals), incorrect order of
steps in constructions. If you manage to start correctly, show uncertainty and do not
continue correctly. 8 tasks will follow one after another.
```

Appendix D: Detailed task-level coding for Experiment 1

Detailed task-level coding results for Experiment 1 are reported in Table 3 (baseline condition) and Table 4 (Pólya-guided condition).

Table 3: Detailed task-level coding for Experiment 1 (baseline condition; 0–2 scale)

Task	*Correctness*	*Task understanding*	*Solution process*	*Result verification*	*Total score*
A1	2	2	2	2	8
A2	2	2	2	2	8
A3	2	2	2	1	7
A4	2	2	2	0	6
A5	2	2	2	0	6
A6	2	2	2	2	8
B1	2	2	2	0	6
B2	2	2	2	0	6
B3	2	2	2	0	6
B4	2	2	2	0	6
B5	2	2	2	0	6
B6	2	2	2	2	8
C1	2	2	2	0	6
C2	2	2	2	0	6
C3	2	2	2	2	8
C4	2	2	2	2	8
C5	2	2	2	0	6
C6	2	2	2	0	6
D1	2	2	2	0	6
D2	2	2	2	0	6
D3	2	2	2	0	6
D4	2	2	2	0	6
D5	2	2	2	0	6
D6	2	2	2	0	6

Table 4: Detailed task-level coding for Experiment 1 (Pólya-guided condition; 0–2 scale)

Task	*Correctness*	*Task understanding*	*Solution process*	*Result verification*	*Total score*
A1	2	2	2	2	8
A2	2	2	2	2	8
A3	2	2	2	2	8
A4	2	2	2	2	8
A5	2	2	2	2	8
A6	2	2	2	2	8
B1	2	2	2	2	8
B2	2	2	2	2	8
B3	2	2	2	2	8
B4	2	2	2	2	8
B5	2	2	2	2	8
B6	2	2	2	2	8
C1	2	2	2	2	8
C2	2	2	2	2	8
C3	2	2	2	2	8
C4	2	2	2	2	8
C5	2	2	2	2	8
C6	2	2	2	2	8
D1	2	2	2	2	8
D2	2	2	2	2	8
D3	2	2	2	2	8
D4	2	2	2	2	8
D5	2	2	2	2	8
D6	2	2	2	2	8

Appendix E: Qualitative notes supporting the coding decisions in Experiment 1

This appendix provides brief qualitative notes supporting the coding decisions for both the baseline and the Pólya-guided conditions in Experiment 1. The notes complement the quantitative coding reported in Appendix D by illustrating how the criteria were applied across individual tasks (Table 5 and Table 6).

Table 5: Comments on problems – Experiment 1 (baseline condition)

Problem	*Comment*
A1	An exemplary, clean solution including all steps and a verification; an ideal answer for the basic equation.
A2	The solution is not only correct but also didactically exemplary—using a clear substitution method, verifying the result, and providing sound justification throughout.
A3	A very high-quality answer that clearly presents two different algorithms; for completeness, it would suffice to conclude with a standard check by substitution.
A4	The procedure is very well explained and the result is correct, but verification is entirely missing – the solver merely states the result without demonstrating any check.
A5	The procedure is correct and clearly presented, but an elementary verification is missing.
A6	A very careful solution with clearly stated conditions and verification of the extraneous root.
B1	The answer correctly emphasizes that the problem cannot be solved uniquely and presents alternative cases; however, verification is entirely missing.
B2	The ambiguity is correctly identified and the alternative cases are described clearly; however, a final retrospective check, including verification and a precise specification of the conditions, is missing.
B3	The problem is solved using two methods and both yield the same result, but an explicit check by substitution back into the original problem is missing.
B4	Two meaningful interpretations with a clearly presented computation; however, a final verification and a clear final statement of the conditions are missing.
B5	A clear and correctly executed computation, but verification or a backward check of the result is missing.
B6	A very clean and complete solution with a clear verification that confirms the result.
C1	Elegant use of prime factorization and comparison of exponents, but a final verification that the original equation is indeed satisfied after substituting the obtained values is missing.
C2	The solution is logically structured and leads to the correct possibilities, but a final check by substitution back through the entire chain of the problem is missing.
C3	A clear and well-structured solution, concluded with a clear verification that confirms the validity of the result.
C4	A systematic solution with verification of all candidates, leading to a clearly justified and verified conclusion.
C5	A clever observation involving collinearity and a clean completion using the law of cosines; however, no separate verification is provided.
C6	A very nicely executed factorization into a product and a systematic search for all pairs; however, a final check by substitution back into the original equation is missing.
D1	The construction is described clearly and correctly, but a final retrospective check is missing—namely, verification that a triangle with the required properties has indeed been constructed.
D2	The construction is described convincingly and with attention to possible variants, but a final verification that the constructed triangle indeed satisfies all given conditions is missing.
D3	The construction is described clearly and supplemented by a case analysis based on the position of the point, but a backward check that the resulting segments indeed satisfy the tangency condition is missing.
D4	The solution is correct and elegant, but a final verification that the constructed quadrilateral indeed has all the properties of a rhombus is missing.
D5	The construction is described clearly and includes a discussion of the number of solutions, but a final verification that a trapezoid with all the given parameters has indeed been constructed is missing.
D6	The procedure is clear and didactically sound, but a final verification that the resulting hexagon is truly regular is missing (for example, by checking equality of sides or angles).

Table 6: Comments on problems – Experiment 1 (Pólya-guided condition)

Problem	*Comment*
A1	Model solution following the four-phase framework exactly; clear, logical, and concluded with a thorough final check.
A2	Worked out in an exemplary manner according to the four phases, with a clear plan, clean computation, and a thorough verification.
A3	Precisely developed solution with a well-explained plan, clear computation, and additionally a double verification of the result (by substitution and by Viète's relations).
A4	Very carefully developed solution with clear justification of the conditions and an exemplary verification using a specific numerical value.
A5	The problem is solved cleanly and systematically, including a thorough check using values both inside and outside the interval.
A6	The solution is very careful: the conditions are correctly established, the quadratic equation is solved, and the check reveals an extraneous root. An exemplary four-step procedure.
B1	The problem is solved systematically, with clear explanation and exemplary verification using specific values.
B2	Both interpretations are presented clearly, with a transparent plan and verification; clean and complete.
B3	The problem is solved cleanly and systematically, including a thorough check of both conditions.
B4	Two reasonable variants with a clear plan and verification, concisely and practically concluded.
B5	Clearly structured computation with verification by estimation; the result is well justified.
B6	Exemplarily formulated joint-work problem, with a clear system of equations and careful verification.
C1	Elegant use of prime factorization, clear selection of the smallest exponents, and correct verification.
C2	Very well-guided backward reasoning, a complete enumeration of possibilities, and a clear consistency check.
C3	Clear algebraic formulation, systematic exploration of the possibilities, and a clear verification; the solution is complete.
C4	Logical narrowing using the least common multiple and an interval, with correct verification of both conditions; the solution is unambiguous.
C5	Collinearity is correctly identified and the result obtained; verification is performed.
C6	Clean completion to a product, a complete enumeration of divisors, and a brief verification; the solution is complete.
D1	The solution is correct; the existence conditions and the number of solutions are discussed clearly.
D2	Clearly guided plan, correct determination of the center and the median, clear verification, and discussion of the number of solutions.
D3	Clean standard procedure using Thales' circle, with clear steps and verification of all possible positions of the point.
D4	Clear standard procedure via the perpendicular bisector; verification of the relevant properties is concise and complete.
D5	A clear band of parallel lines, determination of the vertex using a circle, clean verification, and discussion of the number of solutions.
D6	Elegant procedure involving transferring the radius along the circle, with clear steps and verification that the construction closes correctly.

Appendix F: Qualitative analysis of simulated student solutions in Experiment 2

This appendix provides detailed qualitative comments on individual solution attempts generated in Experiment 2. The comments illustrate how role alignment and didactic usefulness were assessed for each student profile.

F.1 Comments on Individual Problems – diligent but low-achieving student

A4 Both the language and the procedure appear diligent but mechanical. The student correctly chose the common denominator $2(x-1)$, but, in line with the given profile, made a typical routine sign error when removing parentheses:

$$2x + 4 - (x^2 - 4x + 3) \to 2x + 4 - x^2 - 4x + 3$$

instead of the correct $2x + 4 - x^2 + 4x + 3$. There is also no mention of the condition $x \neq 1$. This makes the solution an excellent counterexample for practising the issue of a "minus sign in front of parentheses" and for reinforcing awareness of the domain of definition.

A6 The procedure is diligent and mechanical. A typical error occurs when expanding $(x-1)^2$ as x^2-1 instead of x^2-2x+1; the student also fails to address the condition $x \geq 1$. A check is performed, but purely mechanically, and based on the incorrect computation the student concludes that no solution exists. From an instructional perspective, this solution serves well as a counterexample illustrating correct expansion and the necessity of respecting conditions.

B2 The procedure is mechanical and hasty. The student works only with adding 20% back to the discounted price, which is a typical student error (confusing the base of a percentage calculation). The possibility that the 10% discount is applied after the first discount is completely ignored, and no consideration of alternative cases is given. The result of 1920 CZK is incorrect, but for the teacher the response is valuable—it can be used to demonstrate why percentages cannot simply be added back and why it is crucial to determine correctly the quantity from which a discount is calculated.

B6 The procedure is diligent and template-based (the student correctly writes the equation $\frac{1}{t_1}+\frac{1}{t_2}=\frac{1}{6}$ and the relation $t_1 = 2t_2$), but a typical mechanical error occurs when adding fractions: from $\frac{1}{2t_2}+\frac{1}{t_2}$ the result should be $\frac{3}{2t_2}$, not $\frac{3}{4t_2}$. This is followed by an incorrect computation, yielding $t_2 = 4.5$ h instead of the correct values $t_2 = 9$ h and $t_1 = 18$ h. The student also ignores the word "approximately" in the problem statement (accepting equality without comment) and does not verify the combined time. The solution is a useful counterexample for working with expressions of the form $\frac{1}{t}$ and for checking results by substitution.

C1 The problem originates from a mathematical olympiad and is unsolvable for a ninth-grade student without mathematical aptitude. The submitted solution corresponds more closely to the work of a university student—it is systematic, complete, and error-free. Such a performance cannot be expected even from mathematically gifted pupils. The text therefore entirely misses the intended student profile and provides no didactic value for instruction.

C5 The problem also originates from a mathematical olympiad and is unsolvable for a ninth-grade student without mathematical aptitude. The submitted text contains a complex and highly structured solution using the law of cosines, which is entirely unrealistic for such a student. Thus, there is no alignment with the intended profile. On the other hand, the solution contains an error in determining the angle $\angle BAD$, where an angle of 60° is taken instead of the correct 120°. This type of angle confusion is a typical mistake when applying the law of cosines, and from this perspective the solution can be partially didactically useful as an illustration of incorrect angle interpretation.

D2 The text is procedurally precise and systematic (analysis, construction, discussion) and employs a construction strategy based on the midline: choosing S_a on BC, determining the line $p \parallel AB$ through S_a, using the circle $k(B, t_b)$ to select S_b, and determining A as the intersection of CS_b with the ray of angle β. The construction is mathematically correct (the converse use of the midline parallel to AB is valid; after placing A, the point S_b lies on AC and $\overline{S_aS_b} \parallel AB$ implies that S_b is the midpoint of AC). However, this is a typical textbook solution to a construction problem employing auxiliary points and a guided correctness argument, and is unrealistic for the "diligent without aptitude" profile (ninth grade). It contains no student-like deviations or errors and therefore offers no didactic value within this profile.

D4 The solution is precise and mathematically correct, without a single error. For this profile, however, it is unrealistic – such a student would typically hesitate, might forget one of the defining properties of a rhombus, or carry out the construction incorrectly. From a didactic perspective, there is no moment that could be used to demonstrate a typical mistake or misunderstanding. Because the solution contains neither error nor hesitation, it provides no usable instructional value.

F.2 Comments on Problems – mathematically gifted student with low effort

A4 From the perspective of research question RQ3, the simulation of this profile exhibits the expected characteristics of a student solution: minor errors, inefficient strategies, and unrecognized blind spots that commonly occur in authentic student work. The style is fast and intuitive (with a reference to cross-multiplication); the computation is carried out without verification and without mentioning the condition $x \neq 1$. The algebraic steps themselves are correct and lead to the result $\frac{-x^2+6x+1}{2x-2}$; however, the student makes no attempt at verification or at discussing possible simplification. From a didactic perspective, it is appropriate to add that the denominator can be written as $2(x-1)$ and that the numerator cannot be cancelled by the factor $x-1$ (e.g., since $f(1)=6$). The response matches the intended profile and is useful for diagnosing neglect of the domain of definition.

A6 The procedure is fast and lacks verification. After squaring, the equation $x^2 - 3x - 4 = 0$ is correctly obtained and factored as $(x-4)(x+1) = 0$. However, the conditions and verification are missing ($x \geq 1$, respectively $x - 1 \geq 0$); the root $x = -1$ is invalid, and only $x = 4$ is admissible, since $\sqrt{4+5} = 3$ and $4 - 1 = 3$. This response is didactically suitable for demonstrating the emergence of extraneous roots after squaring and the necessity of verification.

B2 A quick idea without verification: the student adds the discounts $20\% + 10\% = 30\%$ and interprets 1600 as 70% of the original price. The fact that discounts are applied sequentially and multiplicatively is ignored. There are two meaningful interpretations: either $1600 = 0.8P \Rightarrow P = 2000$, or $1600 = 0.8 \cdot 0.9P = 0.72P \Rightarrow P \approx 2222.22$. The computed value of 2286 CZK results from an additive conception of percentages combined with the absence of verification. The response matches the profile and is didactically suitable for diagnosing a frequent error in problems involving compound discounts.

B6 A fast, intuitive decision made without attention to the nuances of the problem statement. The phrase "approximately twice as long" is interpreted relative to the total time (12 h), rather than to the time of the other pump, which is the standard interpretation of the phrase (i.e., $t_1 \approx 2t_2$). The subsequent computation leads to $t_2 = 12$ and to the conclusion that the pumps have equal performance, without any verification of plausibility. A correct (and ideally exact) interpretation would be $t_1 = 2t_2$ and

$$\frac{1}{t_1} + \frac{1}{t_2} = \frac{1}{6} \Rightarrow \frac{1}{2t_2} + \frac{1}{t_2} = \frac{1}{6} \Rightarrow \frac{3}{2t_2} = \frac{1}{6} \Rightarrow t_2 = 9 \text{ h},\ t_1 = 18 \text{ h}.$$

The response clearly illustrates a typical error of this profile: the hasty adoption of an initial interpretation and the absence of verification.

C1 The problem originates from a mathematical olympiad and is intended for highly proficient solvers. The submitted solution is systematic, employing the parametrizations $a = 7^x 11^y$, $b = 7^z 11^w$, comparison of exponents $1+3x = 5z$, $3y = 1+5w$, and selection of minimal solutions. This approach reflects the deliberate work of an experienced solver rather than that of a student profile relying on first ideas and analogies without engaging in structured search. Although the final numerical evaluation and a check of minimality are missing, the overall conception is too advanced and does not align with the intended role. Didactically, the solution is uninformative for this profile: it is not a realistic student performance and does not reveal typical manifestations of a fast but careless approach.

C5 The problem is olympiad-level and demanding for ninth grade. The response partially corresponds to the profile through a rapid, unjustified idea (assuming a right angle at D, applying the Pythagorean theorem, without a sketch or verification), but it completely ignores a key piece of information, $BA = 2$, and fails to account for the collinearity of A, B, C implied by the problem statement. For this reason, alignment with the role is only partial: a quick idea and lack of verification are evident, but the overall reaction to the difficulty of the problem is unrealistically confident. Didactically, the response is useful as an illustration of inappropriate fixation on a right angle without support from the data, and as motivation for systematic consistency checks against all given information; the correct result is $BD = \sqrt{7}$.

D2 The response corresponds to the profile. A quick idea involving completion to a parallelogram is present, but the work is unfocused and lacks verification. Inconsistencies appear in both the data and the notation: the final procedure begins with a segment AD of length a, although $a = |BC|$ (not $|AD|$); further, the angle $\angle DAY = 180° - \beta$ is introduced without a clear connection to the given conditions, and a circle with centre D is used without being properly defined. In earlier steps, three incompatible approaches alternate (midline construction, parallelogram $ABCD$, an attempt via S_b), with parts explicitly abandoned ("I don't know this either", "I give up"). Didactically, the text is valuable: it clearly illustrates typical difficulties of a fast solver—confusion of given elements, inconsistent notation, switching between methods, and erroneous fixation on an auxiliary construction. It can be used to show how to formalize the parallelogram strategy correctly (extending BS_b to length $2t_b$, verifying that $\overline{AC}$ and $\overline{BD}$ bisect each other) and why continuous consistency checks with the given data $a = |BC|$, $\beta = \angle ABC$, $t_b = |BS_b|$ are necessary.

D4 The style matches the profile: a quick idea based on properties of diagonals, a concise procedure without proof, verification, or boundary considerations. The construction itself is mathematically correct (taking $AC = e$, midpoint S, the perpendicular through S, and points B, D at distance

$f/2$), but verification that a rhombus is indeed obtained is missing (e.g., via congruence of the four right triangles at S), as well as a discussion of conditions on $e, f > 0$, uniqueness (the solution is determined only up to congruence), and degeneracy. Didactically, the solution is partially usable: it can illustrate the need for brief justification and explicit formulation of conditions, even when the construction itself is straightforward.

F.3 Comments on Problems – student with neither effort nor aptitude

A4 The style is hesitant and uncertain (with a mention of cross-multiplication). The procedure begins correctly by seeking a common denominator, but is followed by an incorrect expansion of the expression $(x-3)(x-1)$ and errors in handling negative signs. The resulting fraction $\frac{-2x+7}{2x-2}$ is incorrect, and the student neither completes the solution nor verifies it. From a didactic perspective, these features are precisely what make the response valuable: typical errors such as incorrect multiplication of binomials, omission of signs, and failure to recognize that the denominator cannot be cancelled with the numerator. The response matches the profile and is well suited as an illustration of an uncertain and unfinished solution.

A6 The style is uncertain; the student opts for mechanical squaring and then does not proceed further. Typical errors of a weak student appear: incorrect expansion of $(x-1)^2$ (correctly x^2-2x+1, the student writes x^2-1), omission of the conditions $x \geq 1$ and $x \geq -5$, and absence of verification of possible roots. From the quadratic form $x^2-x-6=0$, the solution does not continue, even though this would be a standard next step. The output is useful as a diagnostic example of errors related to squaring, neglect of the domain, and omission of checks for extraneous roots.

B2 The style is indecisive. The student mechanically adds the discounts and treats the advertised price of 1600 CZK as the price after both discounts. A typical confusion occurs: discounts are added instead of being applied sequentially by multiplication, and the base from which the second discount is calculated is not distinguished. The unit-percentage approach leads to an unrealistic result of 2285.7 CZK, and there is no check of the problem statement or consideration of alternative interpretations. Didactically, the response is useful as an illustration of adding percentage discounts, confusing the base, and neglecting the fact that the price of 1600 CZK typically represents 80% of the original price (original price 2000 CZK), while the 10% loyalty discount is subsequently applied to this 1600 CZK.

B6 The procedure is mechanical and incorrect from the outset. Instead of working with rates, times are added (the equation $x+y=6$), indicating a confusion of quantities in a joint-work problem, and the linear relation $x=2y$ is not justified with respect to the problem statement. Verification of the result is missing; times of 2 h and 4 h would lead to a combined time of 4/3 h, not 6 h. The imprecision conveyed by the word "approximately", indicating that the value is not exact and that the problem requires work with rates or at least acknowledgement of the vagueness of the data, is also ignored. Didactically, the output is useful as an illustration of typical errors: adding times instead of rates, incorrect formulation of equations, and absence of verification. For contrast, one may mention the correct scheme $\frac{1}{x}+\frac{1}{y}=\frac{1}{6}$ and that "twice as long" refers to the time of 6 h, i.e. approximately $x \approx 12$ h, which leads to $y \approx 12$ h; a more precise solution, however, requires an unambiguous problem statement.

C1 This is an olympiad problem that goes beyond the standard curriculum. The student responds in a typical way: not knowing how to begin, attempting a simple numerical substitution, and giving up after the first failure. The approach is not systematic, and there is no work with prime factorization or powers, which corresponds to the assumed profile of a weak student. Didactically, the output can be used to illustrate how more demanding problems may lead to immediate blocking and premature resignation, a strategy that is common among some students.

C5 This is again an olympiad problem. The response corresponds to the profile of a weak student: uncertainty, an incorrect assumption of a right angle, and omission of basic checks (the triangle inequality for sides 1, 2, and 3 leads to degeneracy). There is no consideration of the fixed 60° angle between the sides of an equilateral triangle, nor of an appropriate choice of tools (such as the law of cosines or coordinates). The result $\sqrt{5}$ is incorrect; once it is recognized that point B must lie on the line AC (the law of cosines yields $\cos\angle ABC = 1$), one obtains $|BD| = \sqrt{7}$. The output is useful as an illustration of typical blocking in a demanding geometry problem and of replacing argument with guesswork.

D2 The procedure is fragmentary and lacks analysis. The median t_b is confused with the length BA; a circle with centre B and radius t_b is used to determine point A, even though it determines possible midpoints M of side AC for which $BM = t_b$. A crucial connection is missing, namely that M is the midpoint of segment AC and that A is the image of point C under reflection across M. There is no guarantee that point A lies on the ray determining angle β at point B; no verification or discussion of the number of solutions and existence conditions is provided. Didactically, the output is useful as an illustration of typical errors in constructions: omission of analysis and a sketch, confusion of the meaning of given quantities, and neglect of the relationships between the conditions.

D4 The procedure is superficial, without analysis and without a final verification. The key error occurs in the step involving diagonal BD: on the perpendicular through point S, one must mark the distance $\frac{f}{2}$, not f. A circle with centre S and radius $\frac{f}{2}$ determines points B and D; the student uses the full length f, thereby invalidating the construction. The properties of a rhombus are not stated (the diagonals bisect each other and are perpendicular), and discussion of the number of solutions and existence conditions is missing. Didactically, the output is useful as an illustration of typical confusions in constructions involving diagonals: confusing a diagonal's length with its half, omitting the locus, and neglecting final verification that a rhombus with the prescribed diagonal lengths has indeed been constructed.