AI in the Classroom: Didactical Misalignments in Geometry Between Czech and Anglo-Saxon Contexts

Magdalena Krátká, Jiří Přibyl and Michaela Tichá
J. E. Purkynje University in Ústí nad Labem, Czechia

magdalena.kratka@ujep.cz
jiri.pribyl@ujep.cz
michaela.ticha@ujep.cz

Abstract: As artificial intelligence (AI) tools—such as chatbots and large language models—become increasingly accessible in educational settings, both teachers and students are relying on them more during the learning process. These tools provide various pedagogical benefits. However, their integration also introduces didactical risks, particularly when their outputs reflect implicit assumptions and educational paradigms that diverge from those in specific national curricula. This paper explores such risks in the context of lower secondary mathematics education (ages 11–15), focusing on geometry instruction in Czechia. The study builds on the differing conceptualizations of square and rectangle in Czech and Anglo-Saxon didactics. In the Anglo-Saxon tradition, a square is typically regarded as a special type of rectangle, emphasizing hierarchical classification. In contrast, Czech didactics treats these shapes as categorically distinct. This difference reflects broader didactical orientations: Czech mathematics education often emphasizes analytical decomposition and local precision, while Anglo-Saxon approaches favor structural generalization and class inclusion. These contrasting tendencies are mirrored in curricular goals, instructional strategies, and classroom expectations. This divergence becomes especially problematic when AI models—trained largely on English-language data—produce responses that implicitly reflect Anglo-Saxon conventions, which may conflict with the Czech didactical contract. The study uses a comparative, non-experimental methodology to analyze responses from multiple AI systems, including ChatGPT, Gemini, Claude, Copilot, and Mistral. Prompts were administered in both Czech and English to assess the consistency and contextual adaptability of the models. Findings suggest that AI tools may inadvertently reinforce foreign conceptual frameworks, creating tension in cross-cultural educational contexts. The paper highlights the importance of contextual sensitivity, critical digital literacy, and pedagogical oversight in the integration of AI into mathematics instruction. By revealing how culturally embedded definitions in geometry—shaped by language and curriculum—can clash with AI-generated content, this paper offers a relevant perspective for educators facing misalignment between linguistic training data and local didactical norms.

Keywords: Artificial Intelligence in Education, Didactical Contract, Geometric Concepts, Language Models and Educational Contexts, Cross-cultural Misalignment

1. Introduction and Theoretical Background

In recent years, large language models (LLMs) have been increasingly integrated into mathematics classrooms, both globally (Pepin , Buchholtz, and Salinas-Hernández , 2025; Stefanova and Georgiev, 2024; Korkmaz Guler et al., 2024) and in national contexts such as Czechia (E-Bezpečí, 2023). These tools offer benefits like instant explanations and personalized feedback, yet also pose didactical risks when their outputs reflect educational assumptions embedded in one tradition but are deployed in another. This paper explores such tensions through the lens of geometry instruction in Czech lower-secondary education.

This paper focuses on a specific terminological and conceptual misalignment concerning the classification of rectangles. Unlike English, Czech possesses a distinct term—obdélník—which refers to a right-angled quadrilateral with unequal adjacent sides. In this framework, obdélníky form the complement of squares within the broader set of right-angled quadrilaterals. However, in practice, the term obdélník is often equated with pravoúhelník (rectangle in the broad sense), a conflation influenced not only by Anglo-Saxon terminology but also by everyday usage and instructional simplification (Vízek, Samková and Star, 2023). This divergence reflects broader curricular and linguistic differences and may lead to misalignments when students interact with AI systems trained predominantly on English-language data.

To analyze these issues, we draw on the concept of the didactical contract (Brousseau, 2002), which refers to the implicit norms and expectations governing classroom interactions. This contract is culturally embedded, shaping how knowledge is presented and interpreted. In Czech classrooms, students are generally taught that a square and rectangle are mutually exclusive categories, aligned with the local terminological system distinguishing *čtverec* and *obdélník*. Conversely, most LLMs default to inclusive definitions based on hierarchical classification, risking conceptual conflict when applied in Czech settings.

Differences in shape classification also connect to theoretical models like the van Hiele model (van Hiele, 1986), which highlights how students understand geometry in stages, initially recognizing shapes by visual features. At

these early stages, treating squares and rectangles as distinct is developmentally appropriate. This is further supported by Fischbein's theory of prototypicality (Fischbein, 1993), which explains students' tendency to form mental representations based on typical visual features—a square appears more symmetric and less elongated than a rectangle. Czech didactics supports this perceptual distinction through distinct terms and definitions, whereas English-language AI tools often override it.

These tensions are not merely linguistic. Studies have shown that students often reject the inclusive definition *a square is a rectangle* due to its conflict with visual intuition (Tsamir, Tirosh and Levenson, 2008). The Czech system accommodates this by emphasizing partitive classification, often found in resources like Mikulčák (2025), which explicitly exclude squares from the definition of rectangles.

Existing literature on AI in mathematics education reflects similar concerns. While LLMs can support problem-solving and conceptual clarity (Ergene and Ergene, 2025; Awang, Yusop and Danaee, 2025), they may also reinforce misconceptions if not critically mediated by teachers (Daher and Gierdien, 2024; Dilling and Herrmann, 2024). In cross-cultural settings, the risk intensifies: Yan et al. (2024) and Pepin, Buchholtz and Salinas-Hernández (2025) emphasize that language models reflect the conceptual defaults of their training data.

In summary, this study applies the concept of the didactical contract to assess the risk of definitional mismatches introduced by AI systems in the specific context of classifying squares and rectangles within Czech geometry education. It situates this issue within broader theoretical, linguistic, and empirical research on both geometry instruction and the educational implications of LLMs.

2. Methodology

This section describes the design of the study, including the structure of prompts, rationale for their formulation, the selection of AI models, and the analytical focus. The aim was to explore how current large language models respond to culturally dependent geometry questions—especially the classification of squares and rectangles.

2.1 Structure and Content of Prompts

To explore how AI systems handle culturally contingent definitions in mathematics, we designed a structured three-question prompt sequence simulating typical student–AI interactions. Each prompt pair was submitted in a fresh session to avoid contextual contamination.

The first question—our main focus—had four phrasings (V1–V4), all reflecting ways a Czech lower-secondary student might ask whether a square is a type of rectangle. Though similar in tone and wording, they differed slightly to capture natural variation. Each was followed by two identical prompts: one asking for similarities and differences between the shapes, and a final one referencing a Czech teacher to test contextual awareness and whether the model would adjust its earlier answer to match Czech teaching practices.

All three-question sequences were posed in both Czech and English, resulting in 8 unique language–prompt combinations (4 Czech + 4 English). These 8 combinations were each submitted to 7 different AI models (described in section 2.3), yielding a total of 56 AI interactions for analysis, see Table 1.

Table 1: Prompt structure

Prompt Variant	Czech	English
Initial Q1 – V1	Je čtverec obdélník?	Does a square count as a rectangle?
Initial Q1 – V2	Může mít obdélník všechny strany stejně dlouhé?	Can a rectangle have all its sides the same length?
Initial Q1 – V3	Může mi vyjít čtverec, když se v úloze ptají na obdélník?	Can I get a square if the question is about rectangles?
Initial Q1 – V4	Můžeš mi vysvětlit, jestli je čtverec taky obdélník nebo ne?	Can you explain if a square is also a rectangle or not?
Follow-up Q2	Dobře. Co ještě mají čtverec a obdélník společného a co mají různé?	OK. What's the same and what's different between a square and a rectangle?
Follow-up Q3	Souhlasil by s tím učitel matematiky v České republice?	Would a math teacher in the Czech Republic agree with that?

Note: Full transcripts of the AI model responses (all 56 interactions) are available from the authors upon request.

2.2 Justification of Prompt Design

Our prompt sequences were designed to test how AI systems interpret seemingly simple math questions with culturally specific implications. Rather than addressing student misconceptions, the prompts focused on how models handle genuine definitional differences—particularly the contrasting treatment of *square* and *rectangle* in Czech vs. Anglo-American curricula.

The first prompt in each sequence plays a central role. It simulates a spontaneous question that a Czech lower-secondary student might pose when working independently or with AI assistance. We created four natural-sounding phrasings to capture variations in tone and formulation, thus testing whether differences in wording or language affect the model's initial classification. This setup allowed us to observe whether the AI defaulted to an Anglo-Saxon interpretation—typically classifying a square as a special case of a rectangle—or whether it demonstrated sensitivity to the Czech curricular convention, where the two shapes are defined as mutually exclusive.

The second prompt was included to explore the internal consistency of the model's reasoning. By asking about shared and differing properties between the two shapes, we encouraged the system to articulate the geometric basis of its classification.

The third prompt added a cross-cultural layer by referencing a Czech teacher, inviting the model to reconsider whether its earlier answer aligns with local educational standards.

2.3 Models Used

We tested seven large language model configurations from five AI systems: ChatGPT 3.5, ChatGPT o3 (GPT-4), Gemini, Claude, Copilot (Quick Nuance and Deep Thinker), and Mistral. These models reflect tools accessible—or likely accessible—to upper-secondary students in Czechia. All are free or integrated into common platforms, except GPT-4, which was included for performance comparison.

ChatGPT was evaluated in its free-tier 3.5 and premium o3 (GPT-4) versions. Both are trained primarily on English-language data and follow Anglo-American educational norms, potentially limiting alignment with Czech curricula.

Gemini (Google DeepMind), tested in its Gemini 2.0 Flash version, supports multiple languages but often reflects international textbook logic. Claude 3 (Anthropic) offers clear, pedagogical responses but also mirrors Western didactic assumptions.

Microsoft Copilot, based on GPT models, was tested in Quick Nuance (fast) and Deep Thinker (detailed) variants. Both produce fluent English-language responses, with varying depth, but lack strong adaptation to local curricula.

Mistral Le Chat, based on the Mistral Large model, was included for contrast. As a European open-weight model with strong multilingual capabilities, it shows potential for better adaptation to less dominant languages like Czech.

The AI systems tested vary in their handling of Czech, depending on multilingual training data and sensitivity to linguistic variation. ChatGPT o3 (GPT-4), both Copilot variants, and Mistral Large perform well in Czech, generating fluent responses with few translation artefacts. Mistral, developed in a European context, shows strong results in less dominant languages, though no formal Czech benchmarks confirm this yet. Claude 3 also handles Czech well but was primarily trained on English data, which may shape its structure and style. Gemini 2.0 supports multiple languages but often reflects English syntax, especially in complex prompts—suggesting its internal models are influenced by dominant training languages. ChatGPT 3.5 performs less consistently in Czech, occasionally producing phrasing resembling literal English translations, likely due to limited Czech training data. These differences align with multilingual benchmarks such as HELM (Liang et al., 2023), MMLU (Hendrycks, Mazeika and Woodside, 2023), and MT-Bench (Zheng et al., 2024), which show variation in language-specific performance across LLMs.

2.4 Analytical Focus

The analysis focused on three dimensions, each corresponding to one of the three prompts. The first examined whether models classified squares as a subset of rectangles (inclusive framing) or treated them as distinct (exclusive framing), and whether their reasoning was logically consistent. The second explored how many and

which geometric features the models identified when comparing squares and rectangles. The third assessed whether models demonstrated awareness of Czech curricular norms when asked whether a local teacher would agree with their explanations.

While several types of problematic output phenomena were noted during data collection (e.g., terminological inconsistencies, factual errors), these were not analyzed systematically in this study and are reserved for future work. This includes deeper content analysis of Q3 responses, which revealed a richness of interpretation and implicit assumptions about teacher reasoning that warrant separate investigation.

3. Results

This section presents how large language models (LLMs) responded to culturally shaped geometry questions. The analysis followed the three-question structure outlined in the methodology: (Q1) classification of square vs. rectangle, (Q2) similarities and differences between the shapes, and (Q3) recognition of Czech curricular norms. Each question served to assess different aspects: conceptual alignment, explanatory breadth and curricular awareness.

3.1 Consistency and Framing in Responses to Q1: "Is a Square a Rectangle?"

The first question examined whether AI models adhered to the inclusive (Anglo-American) or exclusive (Czech) interpretation of the square–rectangle relationship, and whether their answers were logically consistent.

3.1.1 Conceptual Framing: Inclusive Versus Exclusive Definitions

All 56 responses to Q1 consistently reflected the Anglo-American inclusive view—that a square is a type of rectangle with four equal sides. This was true across all models, languages (Czech/English), and prompt phrasings. Only GPT-4 (ChatGPT o3) acknowledged, briefly and only in Czech, the existence of an alternative (exclusive) definition, but framed it as outdated or informal. No model explicitly recognized this exclusive definition as the Czech standard (see Section 2.2), underscoring a major didactical misalignment when students receive AI-generated definitions.

3.1.2 Logical Contradictions in Tesponses to "Can a Rectangle Have all Sides the Same Length?"

A second finding concerns logical inconsistencies within individual responses, especially those triggered by Q1 variant V2: "Can a rectangle have all its sides the same length?"

Here, several models displayed a pattern of internal contradiction. They began by stating: "A rectangle cannot have all its sides the same length and still be classified only as a rectangle…," (ChatGPT-3.5, EN) implicitly assuming the exclusive definition. Yet later in the same response, they added: "A square is a special type of rectangle…" thereby invoking the inclusive definition. This contradiction undermines the internal coherence of the explanation and may confuse learners—especially when the distinction between generic and special cases is left unstated.

Notably, this inconsistency was more frequent in Czech-language responses. Claude 3, Copilot Quick Nuance, and Mistral exhibited the contradiction only in Czech, while their English responses remained logically consistent. ChatGPT 3.5, by contrast, showed the inconsistency in both Czech and English, making it the only model to do so systematically across both languages. This pattern may reflect weaker internal reasoning alignment in Czech, or broader variability in how different models apply inclusion logic depending on the language context.

3.2 Feature Richness and Conceptual Scope in Responses to Q2: "What's the Same and What's Different?"

This question served to assess how well models identified geometric features. We analyzed the number and nature of features mentioned, and how responses varied by language and prompt variant.

3.2.1 Clustering Language Models by Geometric Feature Mentions

To visualize model responses to Q2, we created a clustermap (Figure 1) showing which geometric features each language model mentioned. This helps us compare how models conceptualize *square vs. rectangle* and what explanations they prioritize or omit.

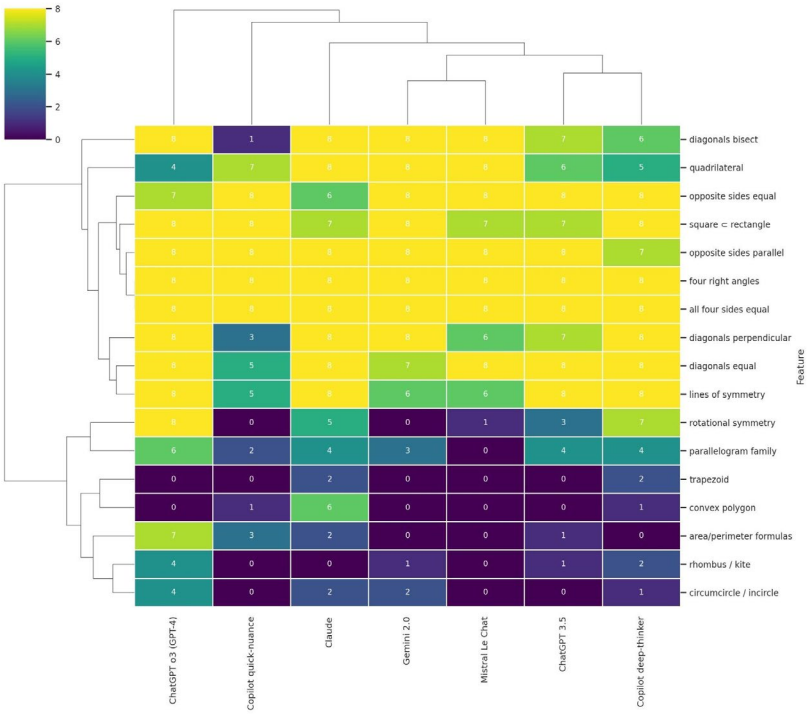

Figure 1: Clustermap of geometric feature mentions across language models

Models are listed on the X-axis, features on the Y-axis. Each cell is color-coded: yellow indicates frequently mentioned features, blue-green indicates rarely or never mentioned ones.

Across all models, several core concepts are consistently cited: *four right angles*, *opposite sides parallel/equal*, *square as a subset of rectangle*, *equal diagonals*, and *all four sides equal*. These align with geometry curricula for ages 11–15, showing that LLMs have internalized textbook-level definitions.

Less commonly mentioned terms (e.g., *parallelogram family*, *trapezoid*, *rhombus/kite*, *circumcircle/incircle*) appear mainly in responses from Claude 3, GPT-4 (o3), and Copilot Deep Thinker. Models like Gemini, Mistral, and Quick Nuance typically omit these, offering a simpler, more age-appropriate response. This has didactic implications: some models risk overwhelming younger learners, while others stick to essential content. Educators should choose models or apply filters accordingly.

The dendrogram shows GPT-4 and Copilot Quick Nuance clustering together despite differences in verbosity. Both omit advanced features similarly, suggesting shared norms on age-appropriateness but different expression styles—useful for researchers studying instruction tuning.

Feature clustering reflects curricular logic: basic properties group together, symmetry terms form a second cluster, and broader classifications like quadrilaterals appear on a separate branch. This structure mirrors didactic models and hints that LLM behavior may help in curriculum mapping.

The clustermap shows that large language models reproduce core textbook facts and reflect elements of curricular structure. This can aid quick reviews or scaffolded explanations but also underscores the need to align the model's output with the learner's developmental stage. Some models keep explanations age-appropriate, while others use advanced terms too early. For educators, this means choosing an AI model is not just technical but also a didactical decision, as model choice, prompt design, and context all impact the quality of explanations students receive.

3.2.2 Effect of Language on Feature Richness in Model Responses

Most language models gave more detailed responses in Czech than in English. Six of seven systems included more geometric features in Czech, omitting two to nine features in English. Mistral was the only exception, performing slightly better in English. Copilot Quick Nuance showed the largest drop, skipping nine features. Basic geometric features were stable across languages, but advanced or less common terms were more often left out in English, suggesting reduced precision or depth.

Figure 2 shows a dumbbell plot comparing each model's Czech and English feature counts. Most lines slope downward, showing consistent information loss in English; only Mistral slopes upward.

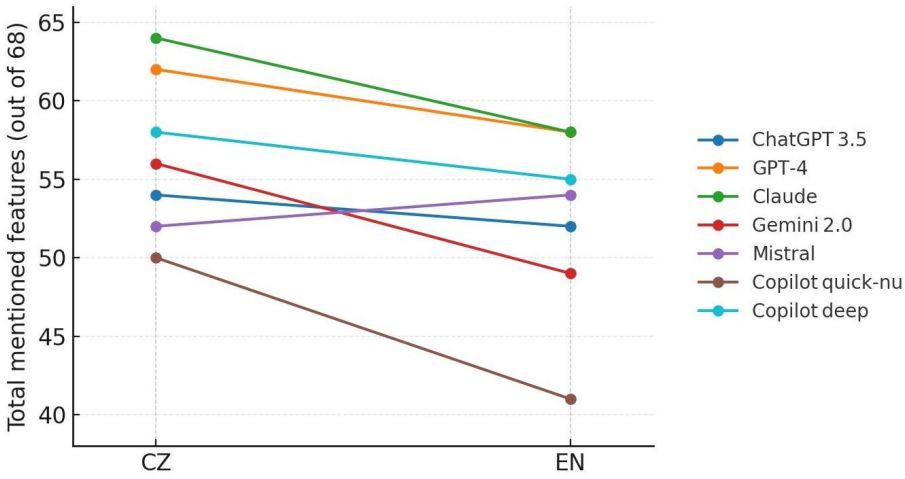

Figure 2: Dumbbell plot comparing feature counts in Czech vs. English responses across models

To test the trend's significance, a two-sided Wilcoxon signed-rank test was applied to paired feature count differences ($n = 7$). The test evaluates whether models tend to behave consistently in favoring one language over the other. At a significance level of $\alpha = 0.05$, the computed statistic ($W = 1.5$) fell below the critical value ($W = 2.1$), allowing us to reject the null hypothesis. We therefore conclude that Czech responses contain significantly more geometric features than English ones.

This analysis captures overall language trends, not individual model differences. Exploring model-specific effects would require a different approach and remains for future work.

Although English prompts were direct translations, Czech responses were often more elaborate. This likely reflects not just style, but how models interpret prompts across languages. Czech may trigger overcompensation, with models adding detail to ensure clarity in a lower-resource language. Also, Czech lacks many idiomatic shortcuts common in English, leading to more explicit wording. These factors help explain the higher detail in Czech outputs.

3.2.3 Prompt-wording Variants and Feature Count

The four wording variants of the initial question (Q1) led to measurable differences in the richness of responses to the follow-up question (Q2) about similarities and differences between squares and rectangles. Although Q2 was identical across sequences, the phrasing of Q1 influenced how many geometric features the models included.

Figure 3 illustrates this with a boxplot comparing feature counts across Q1 variants. Variant V3 elicited the richest responses (highest median and upper quartile), while V4 produced the leanest. V1 and V2 fell in between. The trend suggests that prompt phrasing affects conceptual scope.

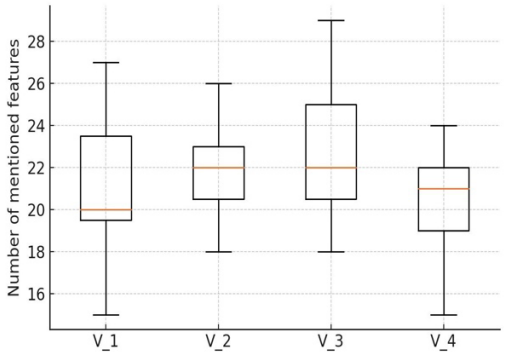

Figure 3: Boxplot showing impact of Q1 wording on number of features mentioned in Q2

A Friedman test was applied to test statistical significance, treating each model as a block ($n = 7$). The result, $\chi^2(3) = 8.91$, exceeds the critical value of 7.82 at $\alpha = 0.05$, allowing us to reject the null hypothesis: prompt wording significantly affects geometric detail in responses.

This may stem from the cognitive framing of each Q1. V3 sets up a classroom-like scenario and hints at a conceptual edge case, prompting broader elaboration in Q2. In contrast, V4 asks a direct classification question, leading to more concise answers.

In sum, subtle changes in prompt wording can influence not only accuracy but also the depth of AI explanations—highlighting the importance of thoughtful prompt design in educational contexts.

3.3 Context Awareness in Responses to Q3: "Would a Czech Math Teacher Agree?"

The third question in each sequence of prompts explores how the model "thinks" about the thinking of Czech teachers. We analyzed the types and frequencies of responses produced by different models in both language versions. Particular attention was paid to the depth and breadth of the responses to this open-ended question.

Our analysis showed no significant differences between the Czech and English outputs, with one exception: the GPT-4 model. Its responses stood out in terms of both richness and caution.

All responses began with a similar phrase: "Yes, a math teacher in the Czech Republic would agree with this explanation." However, they varied in detail and contextualization. In general, the English responses tended to be more elaborate than their Czech counterparts.

The content analysis revealed a notable richness of ideas. This inspired the formulation of several research questions and hypotheses that reflect our expectations. However, given the scope of this article, we do not address these findings in detail here.

4. Discussion

As large language models become more accessible in education, their use in math teaching presents complex challenges. While they offer benefits like quick definitions and explanations, they risk didactical misalignment—especially across cultural or curricular lines. This study highlights three key tensions: the demands of Czech math instruction, risks of unguided AI use, and the challenge of applying global tools in localized educational contexts.

4.1 Implications for Mathematics Education in the Czech Context

Our results show that language models almost universally apply the inclusive definition of a rectangle, treating a square as a special case. While this definition reflects formal mathematics and is common in international discourse, it differs from how the relationship between squares and rectangles is typically taught in the Czech educational system—including primary, lower-secondary, and often also upper-secondary levels—where these shapes are usually presented as distinct and mutually exclusive categories.

This discrepancy can lead to mismatches between what students learn through AI tools and what is expected in schools. If students encounter AI-generated explanations that conflict with textbook definitions, classroom language, or assessment criteria, the result may be not only confusion, but a weakening of curricular coherence. This places new demands on teachers, who must increasingly act as interpreters between digital content and local curriculum goals. In practice, this means helping students recognize that some mathematical terms are context-dependent, and that definitions can vary across educational traditions—even when they are mathematically consistent.

To meet this challenge, teacher education should include not only basic AI literacy, but also training in didactical mediation: how to recognize when an AI-generated explanation might introduce conflicting conceptual structures, and how to respond to such moments constructively within the classroom context.

4.2 Risks of Student Reliance on AI Without Teacher Mediation

Challenges grow when students use AI tools without teacher guidance. While LLMs can offer useful information, they may also provide oversimplified, incomplete, or contradictory explanations—especially with ambiguous prompts or language switches. Younger students may lack the skills to evaluate such output critically.

Without teacher mediation, students risk internalizing inconsistent or curriculum-incompatible definitions. In geometry, where precision is vital, these misconceptions can persist and hinder future learning. Viewing AI as always correct may further entrench misunderstandings, making later correction more difficult for teachers.

4.3 Global Tools Versus Culturally Situated Education

These findings reveal a key tension: globally trained AI tools often conflict with nationally grounded curricula. Though LLMs support multiple languages, their reasoning reflects dominant cultures—mainly English-language math education.

This raises a critical question: should curricula adapt to AI, or should AI align with national pedagogy? Didactically, preserving cultural and curricular specificity is vital. Students need tools that match their educational context.

This calls for a more intentional approach to digital tool integration—one that respects local curricular norms while leveraging the flexibility of AI. It also suggests that developing students' and teachers' critical digital literacy will be just as important as technical access. AI can enrich learning, but only when used with awareness of the cultural and curricular frame in which learning takes place.

5. Conclusion

This study examined how large language models respond to geometry classification questions shaped by different didactical traditions. LLMs often use inclusive definitions that diverge from Czech classroom practices, offering coherent but overly general answers that may not align with the local curriculum and sometimes show inconsistencies—especially in Czech.

While LLMs can support math education, their use requires careful guidance. Teachers are essential in interpreting AI outputs and helping students identify which parts are relevant to their learning. Without this support, students risk adopting technically correct but educationally misleading definitions.

5.1 Limitations of the Study

This study used outputs from publicly available large language models as of early 2025. Given potential future updates, changes in prompt interpretation, or system-level settings (e.g., reinforcement learning), results may vary. Since AI outputs are probabilistic, identical inputs can yield slightly different responses across sessions, limiting generalizability and replicability. The study focused on a specific topic—geometry and square–rectangle classification—within the Czech lower-secondary context, which may not reflect other content areas or education systems. Future research should address broader topics and cross-language dynamics in other subjects.

5.2 Future Directions

One promising way to address observed inconsistencies is through localized or semi-localized AI assistants configured with regional curricula and pedagogical norms. Customized GPTs have already demonstrated improved coherence and alignment in mathematics education (Kwon, 2024). Context-aware tools can enhance clarity and reduce didactical ambiguity, especially in language-sensitive areas (Nyaaba, 2024).

Another important direction is investigating language effects. Our data showed that some Czech prompts often produced more detailed responses than English ones, indicating a potential interaction between language, model behavior, and conceptual depth. Whether this applies across other domains and languages remains an open question.

Future research could examine the qualitative nature of linguistic and conceptual issues in AI-generated explanations. This study noted recurring problems—such as terminological mismatches, stylistic ambiguities, factual inconsistencies, and prototype-based framing—but did not analyze them in depth. A more systematic investigation could offer deeper insights into the didactical impact of AI tools and inform the design of more context-aware systems.

Finally, the integration of AI into mathematics education invites a renewed look at classic didactical theory. The concept of the didactic contract, originally formulated by Brousseau, has been expanded by recent authors to encompass digital and technology-mediated settings. For instance, Hortelano and Prudente (2024) show how

such contracts still govern expectations in online and asynchronous learning environments. Teachers play a key role in shaping digital milieus—whether through interface design or classroom framing—and in helping students navigate the implicit norms of algorithmic explanations (see also Pierce, Stacey and Wander, 2010 or Daher, Baya'a, and Jaber, 2022).

In this sense, integrating AI into education is not only a technical challenge, but also a deeply pedagogical task. It requires teachers, curriculum designers, and researchers to engage critically with the affordances of AI—balancing its global power with local meaning.

Ethics Declaration

This study did not involve human participants or personal data and therefore did not require ethical approval.

AI Declaration

ChatGPT (OpenAI) was used to assist in data collection (via parallel prompt submissions), visualization (e.g., clustermap, dumbbell plot, boxplot), and formulation of selected English text passages. All AI-generated outputs were critically reviewed and edited by the authors to ensure accuracy and disciplinary relevance.

References

Awang, L.A., Yusop, F.D. and Danaee, M. (2025) "Current practices and future direction of artificial intelligence in mathematics education: A systematic review", *International Electronic Journal of Mathematics Education,* Vol 20, No. 2, Article em0823. https://doi.org/10.29333/iejme/16006

Brousseau, G. (auth.), Balacheff, N., Cooper, M. Sutherland, R. and Warfield, V. (eds and trans). (2002) *Theory of didactical situations in mathematics*: *Didactigue des mathématigues, 1970–1990,* Springer, Dordrecht. https://doi.org/10.1007/0-306-47211-2

Daher, W., Baya'a, N., and Jaber, O. (2022) "Understanding prospective teachers' task design considerations through the lens of the theory of didactical situations", *Mathematics*, Vol 10, No. 3, 417, pp 1–14. https://doi.org/10.3390/math10030417

Daher, W. and Gierdien, M.F. (2024) "Use of language by generative AI tools in mathematical problem solving: The case of ChatGPT", *African Journal of Research in Mathematics, Science and Technology Education*, Vol 28 No. 2, pp 222–235. https://doi.org/10.1080/18117295.2024.2384676

Dilling, F. and Herrmann, M. (2024) 'Using large language models to support pre-service teachers' mathematical reasoning—An exploratory study on ChatGPT as an instrument for creating mathematical proofs in geometry', *Frontiers in Artificial Intelligence*, Vol 7, 1460337. https://doi.org/10.3389/frai.2024.1460337

E-Bezpečí (2023) *České školy a umělá inteligence* [Czech schools and artificial intelligence], Univerzita Palackého v Olomouci. Available at: https://e-bezpeci.cz/index.php/ke-stazeni/vyzkumne-zpravy/164-ceske-skoly-a-umela-inteligence-prezentace-2023/file

Ergene, O. and Ergene, B.C.. (2025) "AI ChatBots' solutions to mathematical problems in interactive e-textbooks: Affordances and constraints from the eyes of students and teachers", *International Journal of Education and Information Technologies*, Vol *30, pp 509-545*. https://doi.org/10.1007/s10639-024-13121-z

Fischbein, E. (1993) 'The theory of figural concept', *Educational Studies in Mathematics*, Vol 24, pp 139–162. https://doi.org/10.1007/BF01273689

Hendrycks, D., Mazeika, M., & Woodside, T. (2023) *An overview of catastrophic AI risks. arXiv:2306.12001*. https://doi.org/10.48550/arXiv.2306.12001

Hortelano, J.C. and Prudente, M. (2024) " Effects of the theory of didactical situations' application in mathematics education: A metasynthesis", *Journal of Pedagogical Research,* Vol 8, No. 3, pp 246–262. https://doi.org/10.33902/JPR.202426908.

Korkmaz Guler, N., Dertli, Z.G., Boran, E. and Yildiz, B. (2024) "An artificial intelligence application in mathematics education: Evaluating ChatGPT academic achievement in a mathematics exam", *Pedagogical Research,* Vol 9, No. 2, em0188. https://doi.org/10.29333/pr/14145

Kwon, M. (2024) "Development of a customized GPTs-based chatbot for pre-service teacher education and analysis of its educational performance in mathematics", *The Mathematical Education*, Vol 63, No 3, pp 467–484. https://doi.org/10.7468/mathedu.2024.63.3.467

Liang, W. Yuksekgonul, M., Mao, Z., Wu, E. and Zou, J. (2023) "GPT detectors are biased against non-native English writers", *Patterns*, Vol 4, No. 7. https://doi.org/10.1016/j.patter.2023.100779

Mikulčák, J. (2025) *Matematické, fyzikální a chemické tabulky a vzorce*. [Mathematical, Physical and Chemical Tables and Formulas]. 5th ed. Prometheus, Praha.

Nyaaba, M. (2024) "Glocalizing generative AI in education for the Global South: The design case of 21st Century Teacher Educator AI for Ghana", arXiv:2504.07149. https://doi.org/10.48550/arXiv.2504.07149

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

Pierce, R., Stacey, K. and Wander R. (2010) "Examining the didactic contract when handheld technology is permitted in the mathematics classroom", *ZDM Mathematics Education,* Vol 42, pp 683–695, https://doi.org/10.1007/s11858-010-0271-8

Stefanova, T. and Georgiev, S. (2024) "Possibilities for using AI in mathematics education", *Mathematics Education and Methodology*, Vol 53, pp 117–125. https://doi.org/10.55630/mem.2024.53.117-125

Tsamir, P., Tirosh, D. and Levenson, E. (2008) "Intuitive non-examples: The case of triangles", *Educational Studies in Mathematics*, Vol 69, No. 2, pp 81–95. https://doi.org/10.1007/s10649-008-9133-5

van Hiele, P. M. (1986) *Structure and Insight: A Theory of Mathematics Education.* Academic Press.

Vízek, L., Samková, L. and Star, J. R. (2023) "Investigating how lower secondary school students reason about quadrilaterals emerging in dynamic constructions"*, International Journal of Mathematical Education in Science and Technology*, Vol 56, No. 3, pp 495–514. https://doi.org/10.1080/0020739X.2023.2255184

Yan, L., Sha, L., Zhao, L. Li, Y, Mertinez-Maldonado, R., Chen, G., Li, X., Jin, Y. and Gašević, D. (2024) "Practical and ethical challenges of large language models in education: A systematic scoping review", *British Journal of Educational Technology*, Vol 55, No. 1, pp 90–112. https://doi.org/10.1111/bjet.13370

Zheng, L., Yin, L., Xie, Z., Sun, C., Huang, J., Hao Yu, C., Cao, S., Kozyrakis, C., Stoica, I., Gonzalez, J.E., Barrett, C. and Sheng, Y. (2024) *SGLang: Efficient Execution of Structured Language Model Programs*. arXiv:2312.07104. https://doi.org/10.48550/arXiv.2312.07104