Informatics in Education, 2026, Vol. 25, No. 1, 37–57

DOI: https://doi.org/10.15388/infedu.2509.022

Large Language Models for Educational Task Authoring: A Bebras Challenge Case Study

Leonard BUSUTTIL [0000-0003-3779-891X]
University of Malta, Malta
e-mail: leonard.busuttil@um.edu.mt

Abstract. This study explores the application of large language models (LLMs) to create computational thinking tasks for the Bebras International Challenge through a single-case study approach. Using exemplar-based prompting with seven authentic Bebras tasks from the 2024 cycle as contextual input, a task was developed that was subsequently accepted for inclusion in the 2025 international Bebras challenge. Comparison with the exemplar tasks confirmed that the generated content drew from multiple sources rather than replicating any single task, combining grid-based constraint satisfaction, rule-based filtering, and logical deduction into a novel navigation puzzle with engaging narrative context. International expert reviewers evaluated the task using established Bebras quality criteria, confirming successful alignment with core pedagogical requirements including age-appropriateness, clarity, and cultural neutrality. However, two significant gaps emerged in the broader authoring workflow: accessibility compliance in the researcher-authored visual components and technical inaccuracies in the LLM-generated informatics framing. Following collaborative revision by international editors that addressed these concerns while preserving the LLM's creative contributions, the task achieved acceptance for international use. The findings reveal a collaborative pipeline comprising contextual preparation, LLM-guided generation, human technical implementation, expert community review, and collaborative revision. Results from this case suggest that LLMs can efficiently generate educationally sound creative foundations while requiring integrated human expertise to meet specialised standards and ensure inclusive design, with the task's acceptance providing encouraging evidence for the viability of this collaborative approach.

Keywords: Large Language Models, Educational Task Authoring, Computational Thinking, Bebras Challenge, Human-AI Collaboration.

1. Introduction

Recent advances in LLM capabilities have sparked interest in their potential for educational content creation, with studies demonstrating efficiency gains in the generation of assessment items (Hou *et al.*, 2024; Pourcel *et al.*, 2024). However, concerns persist regarding creativity, technical accuracy, and alignment with specialised educational standards. While some research suggests that LLMs can generate diverse and engaging content (Pourcel *et* *al.*, 2024), others highlight a tendency toward pattern reproduction rather than genuine innovation (Shoaib *et al.*, 2025; Kuusemets *et al.*, 2024). These concerns are particularly relevant for contest-style computational thinking tasks, where originality and pedagogical depth are essential.

The Bebras International Challenge on Informatics and Computational Thinking provides an ideal context for examining these challenges. As a global initiative engaging over three million students annually across more than 50 countries, Bebras has established rigorous quality criteria for educational tasks that emphasise clarity, creativity, cultural neutrality, and relevance to informatics (Dagienė and Futschek, 2008; Vaníček, 2014). The challenge's year-long collaborative development process, culminating in expert review by international panels, offers an authentic evaluation framework rarely available in educational technology research.

Despite the growing body of work on LLM-generated assessments, research examining LLMs in the context of Bebras tasks remains limited. While recent work has explored LLM alignment with human experts in evaluating existing Bebras tasks (Asgari *et al.*, 2025), the suitability of LLMs for generating contest-style computational thinking tasks, which require not only linguistic fluency but also conceptual depth, problem-solving originality, and cultural fairness, remains unexplored. This study addresses that gap through a single case study examining how LLMs can be applied to create Bebras-style computational thinking tasks (Yin, 2018). Rather than relying solely on abstract quality criteria, an exemplar-based approach was employed, providing the LLM with seven authentic Bebras tasks from the 2024 international cycle to ground its understanding in established conventions. The resulting task was submitted to the standard Bebras review process, enabling authentic evaluation by international experts in accordance with established community standards.

Through this case study, the research examines whether LLM-generated content can meet the quality standards of an established international competition through authentic expert review. It also illustrates how exemplar-based prompting can contextualise LLM generation in a domain with specific conventions. In doing so, it adds to broader discussions about human-AI collaboration in creative domains by contributing a situated, empirically grounded account of complementary rather than replacement relationships between human expertise and AI-generated content (Laverghetta Jr. and Licato, 2023; Dascalescu *et al.*, 2025).

2. Background

This section provides the necessary context for the study by outlining the origins and development of the Bebras Challenge, its task design process, and the criteria that define high quality tasks. By situating Bebras within broader educational and computational thinking frameworks, we establish the foundation for evaluating how LLMs might contribute to task authoring.

2.1 A Brief History of the Bebras Initiative

The Bebras International Challenge on Informatics and Computational Thinking is a global initiative promoting computer science and computational thinking (CT) among students of all ages (Datzko, 2019). Conceived in Lithuania by Professor Valentina Dagienė in 2003, the first challenge was held in 2004 (Dagienė and Futschek, 2008; Kalelioğlu *et al.*, 2022). The name "Bebras" (Lithuanian for beaver) reflects qualities of perseverance and problem-solving (Dagienė and Futschek, 2008).

From the outset, Bebras was designed as an international initiative. The 2005 Baltic Olympiad in Informatics introduced the concept to other nations, followed by the first international workshop and organizing committee in 2006. By 2007, Bebras had spread to several European countries, and by 2019 it engaged more than 50 nations and nearly three million students annually (Dagienė and Futschek, 2008; Datzko, 2019).

Bebras serves as an accessible entry point into computing. It requires no prior programming background, instead offering age-appropriate, motivating tasks that directly engage learners with CT concepts (Izu *et al.*, 2017; Vaníček, 2014). Many students encounter CT formally for the first time through Bebras.

2.1.1 Pedagogical Aims and Design

Bebras aims to make informatics engaging and accessible, prioritizing logical reasoning and problem-solving over coding (Dagienė and Futschek, 2008; Vaníček, 2014). Tasks are short (about 3 minutes), self-contained, and curriculum-independent, ensuring global applicability. The overarching goals are to broaden participation in computing, highlight its excitement, and introduce fundamental problem-solving strategies.

To reach a wide range of learners, Bebras adopts a "low floor, high ceiling" model: tasks are divided into easy, medium, and hard levels. This allows all participants to succeed at simpler challenges while offering greater complexity for advanced students (Dagienė and Futschek, 2008; Izu *et al.*, 2017). Beyond the annual contest, the task repository serves as a resource for classrooms and teacher training, providing sustainable support for CT education (Datzko, 2019; Kalelioğlu *et al.*, 2022; Dagienė and Stupurienė, 2016).

2.1.2 The Task Authoring and Refinement Process

Bebras tasks are developed through a year-long, collaborative, and iterative process (Datzko, 2019). Ideas originate at the national level, often via workshops where teachers, scientists, and students propose challenges. Promising tasks are reviewed, refined, and submitted internationally.

The central event is the International Bebras Task Workshop, where experts from member countries form working groups to discuss, critique, and polish tasks (Dagienė and Futschek, 2008; Datzko, 2019). Proposals undergo several revisions to improve clarity, engagement, and educational value. The finalized pool becomes the international task set from which each country selects, translates, and adapts items for its national competition.

2.1.3 Criteria for a High-Quality Task

Various studies (Dagienė and Futschek, 2008; Vaníček, 2014; Datzko, 2019) establish the quality criteria that guide task creation. Core requirements are that a task:

- Relates clearly to informatics, computer science, or CT.
- Can be solved in about three minutes.
- Is appropriate for the target age group.
- Requires no curriculum-specific knowledge or software/hardware familiarity.
- Has a clear, concise problem statement presentable on one screen.
- Is solvable on a computer without external aids.
- Avoids cultural, gender, or religious stereotypes.

In addition, desirable features include humor or engagement, functional use of graphics in SVG format to ensure compatibility across the varied platforms and contest systems used internationally, interactive elements, and post-contest educational value (Dagienė and Futschek, 2008; Datzko, 2019). These principles ensure tasks are not only puzzles but also meaningful educational tools. While some debates continue (e.g., around "no pre-knowledge" in digital literacy tasks), the criteria remain the foundation of Bebras quality assurance (Vaníček, 2014).

2.2 LLMs for Educational Task Creation

In this section, we review the emerging literature on the use of large language models in educational assessment. We synthesize findings around their validity, creativity, and ethical implications, and highlight the research gap concerning contest-style, computational thinking tasks such as Bebras.

2.2.1 Validity and Reliability in LLM-Generated Assessments

Studies have highlighted the efficiency of large language models in producing assessment items, particularly for structured knowledge domains (Hou *et al.*, 2024; Bhandari *et al.*, 2023). LLMs can generate a wide range of tasks quickly, with surface-level plausibility and linguistic fluency, and some work reports promising alignment with curricular goals (Pourcel *et al.*, 2024). However, concerns remain regarding the factual accuracy and reliability of generated content. Tasks may include subtle errors, illogical distractors, or shallow reasoning, which undermine their educational soundness (Zeinalipour *et al.*, 2023; Shoaib *et al.*, 2025). Evaluation methods also diverge: while expert reviews highlight issues of clarity or alignment, learner-based trials reveal further inconsistencies, suggesting that validity cannot be assumed across contexts (Laverghetta Jr. and Licato, 2023).

2.2.2 Creativity and Novelty in Task Generation

Another recurring theme is the extent to which LLMs can contribute genuine creativity. Some researchers note that LLMs generate diverse framings and contexts, which can make tasks feel fresh or engaging (Hou *et al.*, 2024; Pourcel *et al.*, 2024). Yet others emphasize that outputs often reproduce familiar patterns or overfit to common task structures, rather than inventing new problem-solving approaches (Kuusemets *et al.*, 2024; Shoaib *et al.*, 2025). This limitation is particularly relevant to puzzle-like assessments. Whereas traditional test items can tolerate structural repetition, competitions such as Bebras depend on originality and a sense of discovery. Although creativity has been discussed broadly in the literature, it has not been systematically evaluated for cognitively rich or contest-style problem formats.

2.2.3 Bias, Ethics, and Explainability

A further body of research addresses the risks of bias and fairness in LLM-generated educational content. Cultural assumptions, stereotyping, and gendered framings have been observed in outputs, raising concerns for equitable assessment design (Kalelioğlu *et al.*, 2022; Zeinalipour *et al.*, 2023). For example, tasks embedding culturally specific references may disadvantage learners unfamiliar with such contexts. Additionally, explainability remains limited: it is often unclear why a model chooses a particular phrasing, distractor, or problem representation (Bhandari *et al.*, 2023). While prompt refinement and human review can mitigate these risks, oversight is indispensable to safeguard neutrality and inclusivity.

2.3 The Case of Bebras-Style Tasks

Taken together, the literature shows that LLMs can speed up item creation and occasionally enhance variety, but issues of validity, creativity, and neutrality persist. Critically, almost no research has examined their suitability for contest-style computational thinking tasks such as Bebras, which require not only fluency but also conceptual depth, problem-solving originality, and cultural fairness (Dascalescu *et al.*, 2025). This motivates our case study, which investigates the potential of LLMs to support the authoring of Bebras-style tasks. We therefore pose the following research questions:

RQ1. How can large language models be applied in the authoring of Bebras-style computational thinking tasks?

RQ2. To what extent does LLM-generated content align with the established criteria for high-quality Bebras tasks (e.g., clarity, creativity, cultural neutrality, and informatical relevance)?

3. Methodology

3.1 Case Study Design

This study adopts a single-case exploratory design, focusing on the use of a large language model (LLM) to support the authoring of a Bebras computational thinking task. The LLM used in this study was ChatGPT (OpenAI, GPT-4o, March 2025 release) via the official ChatGPT web interface. As the interface does not expose parameter controls (e.g., temperature, maximum tokens), all outputs were generated under default system settings. As a large language model trained on extensive web data, ChatGPT would already have been exposed to substantial publicly available material about the Bebras challenge, including task repositories, pedagogical documentation, and competition guidelines freely accessible online. Seven authentic Bebras tasks from the 2024 cycle were therefore provided as exemplars to focus the model on the specific conventions of that cycle and the target age group, rather than to introduce it to Bebras as a concept from scratch. This distinction is consistent with how exemplar-based prompting operates in LLM contexts more broadly: exemplars activate and focus existing knowledge rather than supplying that knowledge from scratch (Brown *et al.*, 2020).

Building on this approach, the case in question concerns the generation and submission of one task, developed through iterative prompting of the model and subsequently evaluated within the international Bebras review process. A case study approach is appropriate here as it enables an in-depth examination of a novel phenomenon within its authentic context, namely the design and review of Bebras tasks. The purpose is not to generalise statistically but to derive situated insights into the opportunities and challenges of using LLMs for contest-style task creation (Yin, 2018).

3.2 Data Sources

Five primary sources of data informed this study, providing comprehensive documentation of the LLM-assisted authoring process and its evaluation within authentic Bebras community standards.

LLM Interaction Logs: The complete set of prompts and outputs generated during the authoring process was retained, providing a detailed record of the iterative interactions with the language model. This corpus included initial task generation prompts, refinement requests, format modifications, and the model’s responsive outputs at each stage. These logs enabled systematic analysis of the collaborative authoring process and the model’s adaptive capabilities.

Exemplar Task Corpus: Seven authentic Bebras tasks from the 2024 international cycle, specifically selected for the 6–8 age group, served as contextual input to orient the LLM toward established Bebras conventions. These exemplar tasks (see Table 1) represented diverse computational thinking concepts including constraint satisfaction, algorithmic thinking, and logical deduction. The corpus provided a foundation for comparative analysis to assess the originality and pattern synthesis capabilities of the generated task.

Table 1. Exemplar Tasks Provided to LLM for Context (2024 Bebras Cycle, Age 6–8)

Title	Core Concept	Key Mechanics
Pizza Party	Optimization / Preference satisfaction	Analyze friend preferences, select 3 toppings
Ana's Drawing	Stack / LIFO operations	Undo last 3 drawing steps
Online Class	Logical deduction	Determine middle student from seating constraints
Birthday Present	Rule-based filtering	Select ball meeting 3 conditions
Drawing Sailboat	Graph traversal	Draw without lifting pen / retracing
Beaver Grid	Constraint satisfaction	Sudoku-like grid with beaver placement rules
Beaver Robot	Rule validation	Identify impossible sentence from word boxes

Expert Reviewer Feedback: Reviewer comments were collected from the International Bebras Task Workshop, where experts from member countries critique and refine proposed tasks before selection. Six reviewers (five assigned, one non-assigned) provided feedback following the standard Bebras evaluation protocol. The task generated with LLM support was formally submitted under the author's name and reviewed as part of this established process, ensuring that the feedback reflected authentic community standards and criteria. Comments addressed task clarity, age-appropriateness, accessibility, informatics relevance, and adherence to Bebras quality guidelines.

Researcher Reflective Notes: Detailed reflective notes were maintained by the researcher throughout the process, capturing immediate impressions of the model's responses, decision-making rationales during prompt engineering, and criteria applied when selecting outputs for further development. These notes documented the human decision-making process within the collaborative authoring framework and provided insight into the researcher's interpretive role in guiding task refinement.

Revised Task Implementation: Following expert review, the task underwent collaborative revision by international Bebras editors who addressed accessibility and technical accuracy concerns. The before-and-after comparison provided concrete evidence of how human expertise resolved LLM limitations while preserving the generated task's educational value and creative elements.

3.3 Analytical Approach

Analysis proceeded by mapping both the LLM-generated task elements and subsequent reviewer feedback against the established Bebras quality criteria (Dagienė and Futschek, 2008). These criteria emphasise clarity, informatical relevance, creativity, cultural neutrality, and age-appropriateness. Adopting a deductive analytic lens ensured that evaluation remained anchored in widely accepted standards for Bebras tasks. A systematic comparison was conducted between the generated task and the seven provided exemplar tasks to assess originality and pattern synthesis. This analysis examined how the LLM adapted, combined, or innovated beyond existing patterns by mapping structural elements, computational thinking concepts, narrative contexts, and problem-solving approaches across tasks. The comparison enabled identification of derivative elements versus novel contributions, addressing questions of creativity and genuine synthesis in LLM-generated content. Expert comments were systematically categorized and mapped against each Bebras quality criterion to identify patterns in assessment outcomes. This involved coding reviewer suggestions by type (accessibility, pedagogical, technical), frequency of concerns raised, and alignment with established quality standards. The analysis distinguished between surface-level presentation issues and fundamental pedagogical or technical problems to understand where LLM-generated content succeeded or required human intervention. Reflexive interpretation was incorporated throughout the analytical process, with explicit documentation of how the researcher's domain expertise influenced interpretation of both LLM outputs and reviewer feedback. This included acknowledging instances where prior knowledge of Bebras conventions shaped the coding of reviewer comments and the assessment of task quality against established criteria. Such reflexive analysis strengthened the transparency and credibility of the evaluation framework.

3.4 Researcher Reflexivity

The researcher occupied a dual role as both task author and investigator. This insider position afforded familiarity with the pedagogical aims and conventions of Bebras but also introduced potential interpretive bias. For example, expertise in computing education informed both the prompting strategies used to elicit outputs and the criteria applied when selecting which outputs to develop further. Reflexivity was therefore maintained throughout the study by explicitly considering how prior knowledge and assumptions influenced each stage of the process. Such acknowledgement strengthens the transparency and credibility of the analysis.

4. Results

4.1 The Generated Task: "Treasure Hunt Map"

Through iterative prompting and refinement, the LLM generated a grid-based navigation puzzle titled *Treasure Hunt Map*, designed for the 6-8 age group with difficulty scaling across multiple age ranges (6-8 years: hard, 8-10 years: medium, 10-12 years: easy). The task presents a constraint satisfaction problem embedded within an adventure narrative, requiring learners to identify a valid path through a 5×5 grid while adhering to three specific movement rules.

The LLM suggested either an interactive or a multiple-choice question format, with the author choosing to proceed with the multiple-choice option. The core challenge involves guiding a character named Alex from a starting position to a treasure location on a grid containing four terrain types: sand (walkable), trees (walkable), mountains (walkable, with constraint), and water (obstacle). Players must select from four pre-determined coordinate paths, each violating different rules to serve as systematic distractors. The three constraint rules are: (1) cannot step on water, (2) must cross at least one mountain, and (3) can only move orthogonally (up, down, left, right).

Since the LLM was unable to produce a usable SVG graphic at the time the exercise was conducted, which as discussed in Section 2.1.3 is a requirement for Bebras tasks, it was prompted instead to generate a text-based grid layout (see Figure 2). This representation was then used by the author to construct the final SVG map. In addition, the LLM provided the standard Bebras pedagogical framing ("This is Informatics" and "This is Computational Thinking"), which highlighted links to constraint satisfaction, logical reasoning, and algorithmic thinking; this framing was retained and incorporated into the task form.

Task Proposal: Treasure Hunt Map

- **Age Group:** 6-8 years (hard), 8-10 years (medium), 10-12 years (easy)
- **Concepts:** Algorithms, Logical Deduction, Navigation
- **Answer Type:** Interactive (Drag & Drop) or Multiple-Choice

---

Story & Challenge:

Alex the explorer has found an old treasure map! The map shows a grid with different symbols representing trees, rivers, mountains, and sand. Alex needs to follow the correct path to find the treasure. However, he must obey three rules:

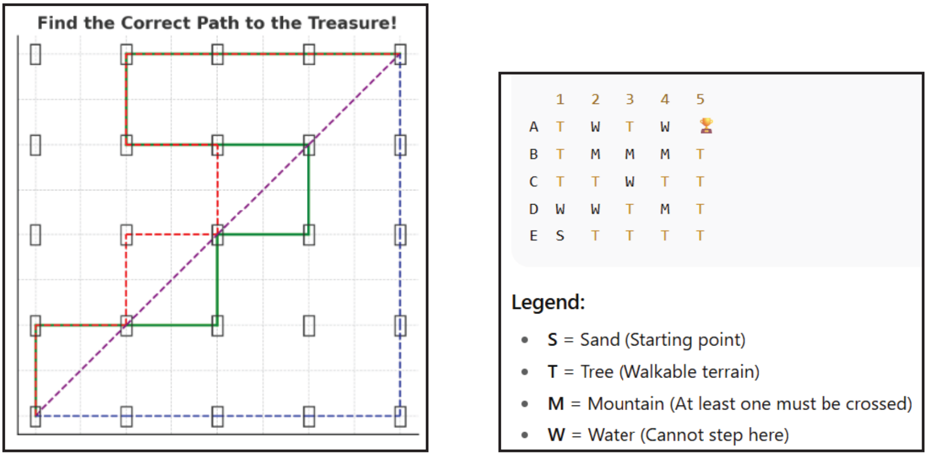
1. He cannot step on water (river).
2. He must cross at least one mountain.
3. He can only move up, down, left, or right—never diagonally.

There are four possible paths on the map. Which path leads Alex safely to the treasure?

Answer Options & Interaction:

- In the interactive version, players drag Alex's route step by step until they reach the treasure. If they break a rule, they must start over.
- In the multiple-choice version, four different pre-drawn paths are given, and players select the correct one.

Fig. 1. Initial proposed task concept: "Treasure Hunt Map."



(a) Initial grid layout Final text based layout

Fig. 2. Development of the "Treasure Hunt Map" layout.

4.2 Pattern Synthesis from Exemplar Tasks

Analysis of the LLM's creative process revealed systematic pattern synthesis across multiple exemplar tasks rather than simple replication of individual concepts. The generated task demonstrated the model's ability to identify, adapt, and recombine elements from the provided corpus while introducing novel variations. Table 2 presents the systematic mapping of influences from exemplar tasks to the generated *Treasure Hunt Map*.

Table 2. Systematic Analysis of Pattern Synthesis from Exemplar Tasks

Exemplar Task	**Core Concept**	**Element Adapted**	**Innovation Applied**
Beaver Grid	Grid-based constraint satisfaction	Structured puzzle format	Applied to pathfinding vs. placement
Online Class	Logical deduction from constraints	Position inference methodology	Route validation through elimination
Drawing Sailboat	Graph traversal rules	Movement constraint framework	Simplified to orthogonal navigation
Birthday Present	Rule-based filtering	Multi-condition selection	Path filtering vs. object selection
Pizza Party	Optimization problem	Selection among alternatives	Applied to route optimization
Ana's Drawing	Sequential operations	Step-by-step validation	Coordinate-based path verification
Beaver Robot	Constraint-based validation	Rule violation identification	Movement rule enforcement

The analysis reveals three key findings regarding the LLM's creative synthesis capabilities. First, no single exemplar provided the complete framework for the generated task; instead, elements were systematically borrowed and adapted from multiple sources. Second, the treasure hunt narrative context was entirely novel, appearing in none of the provided exemplars, suggesting genuine creative contribution beyond pattern recombination. Third, the specific integration of terrain-based navigation with multi-constraint pathfinding represented a new problem formulation that combined familiar structural elements in an innovative configuration.

Most significantly, the LLM demonstrated contextual adaptation rather than direct copying. For example, while the constraint satisfaction concept from Beaver Grid was retained, it was transformed from a placement problem to a navigation challenge. Similarly, the rule-based filtering approach from Birthday Present was adapted from object selection to path validation. This pattern synthesis approach offers one example that counters concerns in the literature about LLMs merely reproducing existing task structures, suggesting that meaningful recombination is at least possible within established pedagogical frameworks.

4.3 Expert Review Process Outcomes

Analysis of reviewer comments revealed consistent patterns across three primary categories: accessibility concerns, pedagogical refinements, and technical accuracy. Table 3 summarizes the frequency and nature of feedback across key evaluation criteria.

Critical Accessibility Concerns

Table 3. Systematic Analysis of Expert Reviewer Feedback

Feedback Category	Frequency	Specific Issues Raised	Assessment
Accessibility	6/6 reviewers	Color-blindness compatibility, visual clarity, pattern differentiation	Critical issue
Age-appropriate language	3/6 reviewers	Terms “diagonally,” “5×5 grid” problematic for youngest learners	Needs refinement
Informatics accuracy	3/6 reviewers	Focus should be constraint satisfaction vs. pathfinding algorithms	Technical correction needed
Task engagement	4/6 reviewers	Positive assessment of narrative and educational value	Strength identified
Problem clarity	5/6 reviewers	Clear problem statement and rule presentation	Strength identified

The most significant finding was universal concern about color-blindness accessibility, with all six reviewers identifying this as a critical issue. Reviewers suggested multiple solutions including pattern-based differentiation (Reviewer 6), numbered labels alongside colors (Reviewer 1), and separate images for each path option (Reviewer 4). It is important to note that the accompanying image was created by the researcher rather than generated by the LLM; consequently, this accessibility oversight cannot be attributed to the LLM's task authoring process. Nevertheless, this finding highlights the importance of accessibility considerations during the broader task development workflow, regardless of whether visual elements are AI-or human-generated.

Technical and Pedagogical Refinements

Three reviewers identified inaccuracies in the informatics framing, with Reviewer 4 noting the task "primarily involves propositional logic" rather than pathfinding algorithms, and Reviewer 1 stating there is "no path finding algorithm in the task, just checking whether certain constraints are satisfied." This suggests the LLM's pedagogical framing, while structurally appropriate, required expert refinement for technical accuracy. Additionally, Reviewer 3 recommended reclassifying the computational thinking focus from pattern recognition to algorithmic thinking and abstraction, indicating the need for more precise pedagogical categorization.

4.4 Alignment with Bebras Quality Criteria

Systematic analysis of reviewer feedback against established Bebras quality criteria revealed both strengths and areas requiring collaborative refinement in the LLM-generated content. Table 4 maps expert assessments to each core criterion and the subsequent resolution outcomes.

Table 4. Alignment with Bebras Quality Criteria: Initial Assessment and Final Resolution

Bebras Criterion	Initial Status	Evidence from Reviews	Final Resolution
Age-appropriate	Met	"age appropriate" (R6), "simple. . . clearly stated" (R2)	Maintained strong alignment
Clear problem statement	Met	"problem statement is clear and concise" (R4)	Maintained strong alignment
Cultural neutrality	Met	No cultural concerns raised	Maintained successfully
Curriculum independence	Met	No prior knowledge requirements noted	Maintained successfully

Bebras Criterion	Initial Status	Evidence from Reviews	Final Resolution
Solvable in 3 minutes	Met	No timing concerns raised	Maintained appropriately
Visual accessibility	Required revision	All 6 reviewers raised color-blindness concerns	Successfully resolved
Informatics relevance	Required refinement	"can use some rewriting" (R1), "should be revised" (R4)	Technically corrected
Creativity/ Engagement	Met	"nice and engaging" (R4), positive narrative reception	Maintained successfully

Successful Initial Alignments: The LLM demonstrated strong performance in core pedagogical areas, meeting fundamental requirements for age-appropriateness, clarity, and cultural neutrality without explicit prompting about these criteria. The engaging narrative and clear rule structure received consistent praise, indicating successful adaptation of Bebras conventions from the exemplar tasks provided.

Areas Requiring Collaborative Resolution: Two significant areas emerged requiring expert refinement. Visual accessibility issues stemmed from the image created by the researcher, which did not explicitly integrate accessibility protocols; this represents a gap in the broader task development workflow rather than a limitation of the LLM-generated content. The informatics framing, while structurally appropriate, contained technical inaccuracies requiring expert correction to properly distinguish between constraint satisfaction and pathfinding concepts.

Successful Resolution Through Human-Expert Collaboration: The revision process systematically addressed all identified concerns while preserving the LLM's creative contributions. Accessibility was resolved through redesigned answer options that eliminated color-dependency, and technical accuracy was achieved through refined informatics explanations with concrete real-world applications. The task's subsequent acceptance for the 2025 international Bebras challenge validates this collaborative refinement approach.

Implications for LLM-Human Collaboration: The analysis reveals that LLMs can successfully generate educationally sound creative foundations that meet high-level pedagogical standards, while specialized technical requirements and accessibility compliance benefit from integrated human expertise. This suggests a sustainable collaborative model where AI efficiency combines with human domain knowledge to produce content meeting international educational standards.

5. Discussion

This study investigated the application of large language models in creating computational thinking tasks for the Bebras International Challenge. The central finding is encouraging: through a structured exemplar-based approach, LLM-assisted authoring produced a task that met core pedagogical requirements, received positive expert assessment, and was accepted for international use. The findings also offer practical guidance about where human expertise adds most value within such a pipeline, which is itself a useful contribution to understanding how LLMs can be deployed effectively in specialised educational contexts.

5.1 Addressing RQ1: LLM Application in Bebras Task Authoring

This case demonstrates that exemplar-based prompting can offer a productive approach to LLM-assisted Bebras task authoring. Providing contextual exemplars oriented the model toward established Bebras conventions while leaving room for novel problem formulations, resulting in a task concept that required relatively little structural intervention before submission. Whether this reflects a reliable capability across different prompting strategies and task types remains an open question for future investigation, but the outcome in this case is a useful proof of concept.

The collaborative authoring process that emerged reveals three distinct phases: (1) exemplar-informed generation, where the LLM synthesised patterns from multiple authentic tasks; (2) iterative human-guided refinement, where the model's initial task concepts were developed through further prompting; and (3) human technical implementation, where the researcher translated LLM outputs into a visually coherent, submission-ready task. This division of labour proved effective, with each phase contributing distinctly to the final outcome.

A structural feature of this workflow is the inability of current LLMs to generate the SVG graphics that Bebras tasks require. This is a practical constraint rather than a fundamental barrier: it simply means that graphical implementation is a distinctly human responsibility within the pipeline, while conceptual and textual content can be developed collaboratively with the model. Knowing this in advance allows the workflow to be planned accordingly.

The pattern synthesis observed in this case offers one illustration of recombination across exemplars rather than replication of any single source. The mappings identified in Table 2 operate at a high level of abstraction, and it would overstate the evidence to claim that the LLM engaged in creativity analogous to human innovation. What the case does show is that the model produced a task that was structurally distinct from each individual exemplar and that was judged by international experts to be engaging and educationally sound. That is a meaningful outcome within an established pedagogical framework, and one that offers a constructive counterpoint to concerns about LLMs merely reproducing existing structures (Kuusemets *et al.*, 2024; Shoaib *et al.*, 2025).

5.2 Addressing RQ2: Alignment with Bebras Quality Criteria

The generated task performed strongly against the core Bebras quality criteria. Age-appropriateness, clarity, cultural neutrality, and task engagement were all met without explicit prompting about these criteria, with multiple reviewers providing positive assessments across each dimension. This suggests that exemplar exposure can effectively convey established domain conventions to the model, supporting findings about implicit pattern learning in LLM contexts (Hou *et al.*, 2024).

Two areas required expert intervention, and understanding their nature is instructive for designing effective pipelines. The first, colour-blindness accessibility in the researcher-authored graphics, was identified by all six reviewers and resolved during the revision process. This oversight is best understood as a workflow gap rather than a limitation of the LLM: since visual components necessarily fall to human contributors, accessibility compliance must be explicitly built into the human implementation phase. The finding is a useful reminder that quality assurance responsibilities are distributed across a human-AI pipeline and need to be planned for at each stage.

The second area, technical accuracy in the informatics framing, is more theoretically interesting. Three reviewers identified that the LLM had characterised the task as involving pathfinding algorithms when it in fact requires propositional logic and constraint checking. This is precisely the kind of issue that domain-expert review is designed to catch, and the review process caught it. The correction was straightforward once identified, and the revised framing was retained in the accepted task. The practical implication is clear: LLM generated informatics framing should be treated as a strong first draft requiring expert verification, rather than as a technically authoritative description. Within a pipeline that includes that verification step, this limitation is manageable.

Taken together, the pattern of findings suggests that LLM alignment with Bebras quality criteria is strongest at the level of pedagogical and narrative conventions, where exemplar exposure appears sufficient, and requires targeted human input at the level of technical accuracy and visual accessibility. This is a practically useful characterisation of where the human contribution is most needed.

5.3 The Complete Collaborative Pipeline: From Generation to Acceptance

The task's acceptance for the 2025 international Bebras challenge is the most concrete finding of this study. It demonstrates that LLM-assisted authoring, when supported by appropriate human expertise and community validation, can produce content meeting rigorous international standards. The revision process is also telling: expert intervention corrected specific technical and accessibility gaps without needing to reconceive the task from scratch, indicating that the LLM's core creative and structural contributions were sound. The pipeline illustrated in Figure 3 worked as intended.

This outcome is encouraging and, while a single case cannot establish general patterns, it provides a concrete basis for further investigation. The pipeline identified here, comprising contextual preparation, LLM-guided generation, human technical implementation, expert review, and collaborative revision, offers a practical model that others working in this space can build on and test.

5.4 Implications for Educational Task Authoring

The study points toward a productive role for LLMs in educational content creation: not as autonomous generators of publication-ready material, but as efficient creative partners that reduce the time and effort required to reach a strong first draft. The efficiency gains are real. Generating a structurally sound task concept with appropriate narrative framing, multiple-choice distractors each designed to violate a specific rule, and pedagogical connections to computational thinking concepts would represent considerable work for a human author starting from scratch. In this case, that foundation was in place after a short prompting session.

Fig. 3. The collaborative pipeline model.

The pattern synthesis observed here also has broader implications. The ability to draw on multiple exemplars and produce something structurally novel suggests that LLMs can contribute meaningfully to domains where established conventions must be respected and originality is still valued, a balance that characterises many specialised educational contexts beyond Bebras. This is a more optimistic picture than concerns about mere pattern reproduction (Kuusemets *et al.*, 2024) might suggest, and it is grounded in authentic expert evaluation rather than researcher assessment alone.

The key design principle that emerges is that human expertise should be concentrated where it adds most value: technical accuracy in domain-specific framing, accessibility compliance in visual implementation, and community validation through established review processes. A pipeline that is explicit about this division of labour is well placed to produce high-quality output efficiently.

5.5 Broader Implications for Human-AI Collaboration

The collaborative model identified here has potential applications beyond Bebras task authoring. The exemplar-informed approach offers a practical method for grounding LLM generation in domain-specific conventions without requiring extensive prompt engineering. The iterative refinement phase shows how human expertise can guide model outputs toward specialised standards through relatively lightweight interaction. And the community validation phase demonstrates that LLM-assisted content can be submitted to established quality assurance processes and succeed within them.

This model addresses concerns about AI displacing human expertise by showing that the relationship is complementary rather than competitive. LLMs handle what they do well: rapid generation of structurally coherent, narratively engaging content drawing on broad pattern knowledge. Humans handle what they do well: domain accuracy, inclusive design, and professional judgment. The result, in this case, was content that met international educational standards and was accepted for use with learners across more than fifty countries, a genuinely positive outcome worth building on (Laverghetta Jr. and Licato, 2023; Dascalescu *et al.*, 2025).

6. Conclusion

This study set out to examine whether LLM-assisted authoring could produce computational thinking tasks meeting the rigorous quality standards of the Bebras International Challenge. The answer, in this case, is yes. The generated Treasure Hunt Map task was positively received by international expert reviewers, successfully revised through the established community process, and accepted for inclusion in the 2025 international challenge. That outcome is the central finding of the study and provides a concrete, encouraging basis for further work in this area.

The study also offers practical guidance about how to achieve that outcome. Exemplar-based prompting effectively oriented the model toward established domain conventions, producing a task concept that was structurally sound and narratively engaging without requiring extensive intervention. Human expertise contributed most distinctly in two areas: correcting technical inaccuracies in the LLM-generated informatics framing, which expert reviewers identified and which were straightforward to resolve once flagged, and ensuring accessibility compliance in the visual components, which falls to human contributors given current LLM limitations in SVG generation. A pipeline that is designed with these responsibilities clearly assigned is well placed to produce content of publishable quality efficiently.

The broader implication is that LLMs are most productively understood as creative partners in specialised educational content creation rather than either autonomous generators or mere autocomplete tools. They contribute genuine structural and narrative value; they require targeted human input at specific points; and the combination, when supported by community validation, can meet international standards. This is a more nuanced and more useful picture than either uncritical enthusiasm or broad scepticism about LLMs in education tends to produce.

This study is limited by its single-case design, and replication across diverse task types, age groups, and models is needed to establish how reliably these patterns hold. That is the natural next step, and the pipeline identified here provides a concrete framework to replicate and extend.

Appendix A: LLM Interaction Log

The following presents the key prompts submitted to ChatGPT (OpenAI, GPT-4o, March 2025 release) during the task authoring session, along with summaries or full reproductions of model responses where methodologically relevant.

Prompt 1: Exemplar Task Summarisation

You are an expert educational content creator specialising in computational thinking tasks for the Bebras International Challenge.

The Bebras Challenge is a global initiative promoting computer science and computational thinking among students. Tasks must:

- *Relate clearly to informatics, computer science, or computational thinking*
- *Be solvable in about three minutes*
- *Be appropriate for the target age group*
- *Require no curriculum-specific knowledge or software familiarity*
- *Have a clear, concise problem statement*
- *Avoid cultural, gender, or religious stereotypes*
- *Include engaging narratives and functional graphics*
- *Have educational value beyond the contest*

Consider these task proposals. Give me a summary of each.

The seven task files (2024-DE-02, 2024-MK-04, 2024-MY-03, 2024-NL-02, 2024-PL03, 2024-SA-02, 2024-VN-01) were uploaded alongside this prompt. The model produced structured summaries of each task, identifying age group, core concepts, task description, learning outcomes, and informatics connections. These summaries are reproduced in Table 5 below, as they represent the model's initial processing of the exemplar corpus.

Table 5. Model-Generated Summaries of Exemplar Tasks

Task	Core Concept	Description
Pizza Party (2024-DE-02)	Optimisation / Preference satisfaction	Select three toppings satisfying the most friends' wishes
Ana's Drawing (2024-MK-04)	Algorithms / Data Structures (LIFO)	Determine image appearance after undoing last three steps
Online Class (2024-MY-03)	Logical deduction	Determine which student sits in the middle from seating constraints
Birthday Present (2024-NL-02)	Rule-based filtering	Select ball meeting three conditions
Drawing Sailboat (2024-PL-03)	Graph traversal	Draw sailboat without lifting pen or retracing lines
Beaver Grid (2024-SA-02)	Constraint satisfaction	Sudoku-like grid with beaver placement rules
Beaver Robot (2024-VN-01)	Rule validation	Identify impossible sentence from word boxes

Prompt 2: Task Generation Request

"Inspire yourself from these 7 tasks to create a new task targeted at this age group. Do not copy."

The model generated an initial task proposal titled Treasure Hunt Map, reproduced in full below as it represents the primary creative output of the session.

Task Proposal: Treasure Hunt Map

Age Group: 6–8 years (hard), 8–10 years (medium), 10–12 years (easy)

Concepts: Algorithms, Logical Deduction, Navigation

Answer Type: Interactive (Drag and Drop) or Multiple-Choice

Story and Challenge:

Alex the explorer has found an old treasure map. The map shows a grid with different symbols representing trees, rivers, mountains, and sand. Alex needs to follow the correct path to find the treasure. However, he must obey three rules:

1. He cannot step on water (river).
2. He must cross at least one mountain.
3. He can only move up, down, left, or right — never diagonally.

There are four possible paths on the map. Which path leads Alex safely to the treasure?

The model also provided standard Bebras pedagogical framing under "This is Informatics" and "This is Computational Thinking," linking the task to pathfinding algorithms, constraint satisfaction, decomposition, algorithmic thinking, and pattern recognition. When asked which exemplar tasks had inspired the generated task, the model identified all seven sources and described specific adaptations, summarised in the pattern synthesis analysis presented in Table 2 of the main paper.

Prompt 3: Format Selection

"Use multiple choice."

The model reformatted the task as a multiple-choice question, maintaining the same narrative context, grid structure, and three constraint rules. It retained the four-path structure with each incorrect option designed to violate a different rule.

Prompt 4: Grid Representation

"Using S for Sand, T for Tree, M for Mountain, W for Water, draw the grid with ABC for rows and 1, 2, 3 for columns. Draw a text grid which the children must navigate."

The model produced the following text-based grid layout, which was subsequently used by the author to construct the final SVG map.

Table 6. Text-Based Grid Layout Generated by the LLM

	1	**2**	**3**	**4**	**5**
A	T	W	T	W	**[Treasure]**
B	T	M	M	M	T
C	T	T	W	T	T
D	W	W	T	M	T
E	S	T	T	T	T

Legend: S = Sand (Starting point); T = Tree (Walkable terrain); M = Mountain (At least one must be crossed); W = Water (Cannot step here).

Acknowledgements

The author would like to acknowledge and thank the members of the international Bebras Challenge community who contributed to the development of the Bebras tasks used in this study. The Bebras tasks and associated images are licensed under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) licence.

References

Asgari, M., Mannila, L., Strömbäck, F. (2025). How Aligned are Humans and Large Language Models in Evaluating Computational Thinking Tasks? In: International Conference on Informatics in Schools: Situation, Evolution, and Perspectives, pp. 111–124. Springer.

Bhandari, S., Liu, Y., Pardos, Z.A. (2023). Evaluating ChatGPT-generated Textbook Questions using IRT. In: NeurIPS'23 Workshop on Generative AI for Education (GAIED).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.*(2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877–1901.

Dagienė, V., Stupurienė, G. (2016). Bebras–A Sustainable Community Building Model for the Concept Based Learning of Informatics and Computational Thinking. Informatics in education, 15(1), 25–44. https:// doi.org/10.15388/infedu.2016.02.

Dagienė, V., Futschek, G. (2008). Bebras International Contest on Informatics and Computer Literacy: Criteria for Good Tasks. In: Mittermeir, R.T., Sysło, M.M. (Eds.), Informatics Education -Supporting Computational Thinking. Springer, Cham, pp. 19–30. https://doi.org/10.1007/978-3-540-69924-8_2.

Dascalescu, S., Dumitran, A.M., Vasiluta, M.A. (2025). Leveraging Generative AI for Enhancing Automated Assessment in Programming Education Contests. Preprint.

Datzko, C. (2019). The Genesis of a Bebras Task. In: Pozdniakov, S.N., Dagienė, V. (Eds.), Informatics in Schools. Closing the Gap Between Tumbling Walls and Rising Towers. Springer, Cham, pp. 240–255. https://doi.org/10.1007/978-3-030-33759-9_19.

Hou, X., Wu, Z., Wang, X., Ericson, B.J. (2024). CodeTailor: LLM-Powered Personalized Parsons Puzzles for Engaging Support While Learning Programming. In: Proceedings of the 11th ACM Conference on Learning @ Scale (L@S '24), pp. 355–369. https://doi.org/10.1145/3657604.3662032.

Izu, C., Mirolo, C., Settle, A., Mannila, L., Stupurienė, G. (2017). Exploring Bebras Tasks Content and Performance: A Multinational Study. Informatics in Education, 16(1), 39–59. https://doi.org/10. 15388/infedu.2017.03.

Kalelioğlu, F., Doğan, D., Gülbahar, Y. (2022). Snapshot of Computational Thinking in Turkey: A Critique of 2019 Bebras Challenge. Informatics in Education, 21(3), 501–522. https://doi.org/10.15388/ infedu.2022.19.

Kuusemets, L., Parve, K., Ain, K., Kraav, T. (2024). Assessing AI-generated (GPT-4) Versus Human Created MCQs In Mathematics Education: A Comparative Inquiry into Vector Topics. International Journal of Education in Mathematics, Science, and Technology (IJEMST), 12(6), 1538–1558. https://doi.org/10. 46328/ijemst.4440.

Laverghetta Jr., A., Licato, J. (2023). Generating Better Items for Cognitive Assessments Using Large Language Models. In: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pp. 414–429.

Pourcel, J., Colas, C., Molinaro, G., Oudeyer, P.-Y., Teodorescu, L. (2024). ACES: Generating diverse programming puzzles with autotelic language models and semantic descriptors. In: NeurIPS 2024 -The 38th Annual Conference on Neural Information Processing Systems. hal-04372580v1.

Shoaib, M., Husnain, G., Sayed, N., Ghadi, Y.Y., Alajmi, M., Qahmash, A. (2025). Automated Generation of Multiple-Choice Questions for Computer Science Education Using Conditional Generative Adversarial Networks. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3530474.

Vaníček, J. (2014). Bebras Informatics Contest: Criteria for Good Tasks Revised. In: Gülbahar, Y., Karataş, E. (Eds.), Informatics in Schools. Curricula, Competences, and Competitions. Springer, Cham, pp. 17–28. https://doi.org/10.1007/978-3-319-09958-3_3.

Yin, R.K. (2018). Case study research and applications (Vol. 6). Sage Thousand Oaks, CA.

Zeinalipour, K., Iaquinta, T., Zanollo, A., Angelini, G., Rigutini, L., Maggini, M., Gori, M. (2023). Italian Crossword Generator: Enhancing Education through Interactive Word Puzzles. In: Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023).