Contemporary Educational Technology, 2026, **18**(4), ep704
ISSN: 1309-517X (Online)



OPEN ACCESS

Research Article

The design and impact of a dual-agent artificial intelligence system in personalised mathematics tutoring

Ruonan Wang [1*]
0009-0007-4430-1402

Laura Tuohilampi [2]
0000-0002-8117-4269

[1] School of Mathematics and Statistics, The University of New South Wales, Sydney, NSW, AUSTRALIA
[2] School of Education, The University of New South Wales, Sydney, NSW, AUSTRALIA
[*] Corresponding author: z5445041@zmail.unsw.edu.au

Citation: Wang, R., & Tuohilampi, L. (2026). The design and impact of a dual-agent artificial intelligence system in personalised mathematics tutoring. *Contemporary Educational Technology, 18*(4), Article ep704. https://doi.org/10.30935/cedtech/19950

ARTICLE INFO

Received: 2 May 2026
Accepted: 9 Nov 2026

ABSTRACT

Digital tutoring tools are increasingly used in schools to support mathematics learning, but empirical evidence remains limited on how diagnostic and instructional functions can be coordinated to provide personalised support. This study examined a dual-agent intervention in which a diagnostic agent generated provisional learning-profile classifications (diverging, assimilating, converging, and accommodating) and a personalised tutoring agent delivered profile-responsive, stage-specific guidance within Polya's (1945) problem-solving framework; a general tutoring agent provided non-profile-responsive guidance in the control condition. The quasi-experimental study involved 206 Chinese grade 5 students, including 99 in the experimental group and 107 in the control group. The pre-test and post-test used two parallel mathematics tasks evaluated using the same analytic rubric, and learning gain was calculated from these matched assessment components. The experimental group achieved a significantly larger mean gain than the control group (5.14 vs. 0.94, $t(204) = 4.36$, $p < .001$, Hedges' $g = .61$). A baseline-adjusted sensitivity analysis confirmed the overall group effect ($p < .001$). After Holm correction for four profile-specific comparisons, only the assimilating-profile contrast remained significant (adjusted $p = .0496$). However, the omnibus group-by-profile interaction was not significant ($p = .931$), warranting caution in interpreting profile-specific differences. Overall, the findings support separating diagnostic and tutoring roles and structuring artificial intelligence (AI) guidance around mathematical problem-solving stages.

Keywords: learning profiles, pedagogical AI agents, personalised mathematical tutoring, Polya's problem-solving model, quasi-experimental design

INTRODUCTION

Digital systems are increasingly expected to provide personalised support in mathematics education (Engelbrecht & Borba, 2024). This is in line with personalised experiential learning, aimed to support critical and creative thinking, reflection and independence, all essential mathematics problem-solving skills. This in turn can be further enhanced by integrating collaboration and generative artificial intelligence (GenAI) support (Lin et al., 2025). Indeed, with the development of artificial intelligence (AI), the focus has shifted from whether digital technologies should be used to how they can provide meaningful guidance to individual students. For example, recent studies have examined students' adoption of tools such as ChatGPT (López-Mendoza et al., 2026; Owan et al., 2025), teachers' beliefs about ICT integration in mathematics classrooms (Tong et al., 2026), and possible relationships between AI-supported tools, engagement, self-regulation, and cognitive independence (Ussenova et al., 2025). For even more details, Shoustikova et al. (2026) provided a comprehensive overview of AI literacy in relation to possible applications in education.

Copyright © 2026 by authors; licensee CEDTECH by Bastas. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/).

Within mathematics education more specifically, Mohamed et al. (2022) showed that AI-related research has developed across multiple approaches, and much of this work has focused on the effectiveness of AI-based tools in improving learning outcomes. Schorcht et al. (2024) likewise found that prompt techniques can affect the process-related quality of AI-supported mathematical problem-solving, even when content-related quality remains unchanged. Overall, these studies indicate growing interest in AI-supported learning, but they also suggest that research has more often focused on intervention effectiveness than on how instructional support is structured within mathematical problem-solving or how such support may work differently across learners.

Primary mathematics provides a useful context for examining this issue. Mahmud and Mohd Drus (2023) claimed that the primary stage is a critical period in students' learning of structured reasoning skills. Meanwhile, students' differences in learning pace, confidence, and interest are becoming more apparent (Powell & Kusuma-Powell, 2011). As Muhamad Fadzil et al. (2025) pointed out, primary students who encounter mathematics word problems often find it hard to

(a) construct a correct picture of the problem,

(b) develop a workable strategy,

(c) sustain a consistent strategy throughout a multi-step problem, and

(d) evaluate whether or not the answer makes sense.

These difficulties should be viewed holistically, in relation to when and how students are supported in these early years of developing mathematical problem-solving skills.

Uniform instruction often fails to meet the needs of mathematics students who struggle to visualize problem conditions, sustain a strategy in multi-step problems, and require differential levels of assistance and feedback (Jyrwa & May, 2025). Not surprisingly, Powell and Kusuma-Powell (2011) identified personalised tutoring as an important response to students' differences in problem-solving. However, personalised learning assistance is hard for teachers to sustain due to workload and institutional priorities, as pointed out by Gašević et al. (2023). In this context, pedagogical GenAI agents offer a potentially scalable way to provide personalised guidance. By automating diagnostic and instructional roles, these systems may provide the kind of sustained and responsive support described by Powell and Kusuma-Powell (2011), thereby offering a feasible path toward achieving adaptive and holistic support in classrooms.

While traditional AI systems can effectively automate diagnostic roles by classifying learners into profiles, providing responsive tutoring requires the sophisticated interaction capabilities offered by GenAI. Han et al. (2025) highlighted the potential of GenAI systems to serve as learning partners, capable of providing adaptive feedback and maintaining interactive conversations, a level of interaction that exceeds that of static digital tools. For mathematical problem-solving, this interaction is essential. For example, Nasongkhla and Shieh (2025) showed how employing questions powered by GenAI might encourage more in-depth cognitive processing and reflective engagement (Lai et al., 2024). GenAI can recognize contextual information and generate real-time responses to learners' inputs, replicating the sophisticated instructional framework of human tutors (Apoki et al., 2022).

Importantly, Gibson et al. (2023) pointed out that the instructional effectiveness of such systems depends more on the quality of instructional design than on the technology itself. This highlights the necessity of focusing on the structure and design of pedagogical AI systems. For example, Hong and Guo (2025) proposed the use of prompts as part of the design of AI tutoring systems. Wu et al. (2024), in turn, noted that single-agent systems have limitations: using just one agent is not effective in breaking down complex tutoring tasks into manageable sub-tasks or adapting effectively during decision-making. This is particularly relevant in mathematical problem-solving, where learners may require support that is sensitive both to the stage of the task and to differences in learning profiles. However, there is limited empirical research on how the design of AI-supported guidance influences learners' problem-solving processes and whether its benefits vary across learners.

This study addresses three linked gaps in the existing literature: the limited separation of diagnostic and instructional functions in AI tutoring, the weak alignment of adaptive support with a transparent mathematical problem-solving sequence, and the scarcity of empirical evidence on whether profile-responsive support

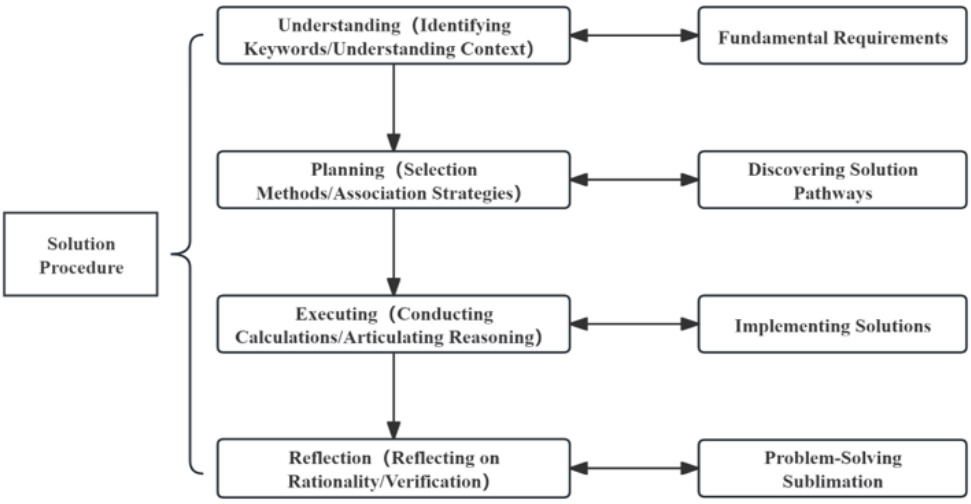


Figure 1. Polya's (1945) problem-solving model (created by the authors based on Polya, 1945)

works similarly across learner groups. To address these gaps, the system employed one diagnostic agent and two tutoring agents. The diagnostic agent generated provisional learning-profile classifications, while the two tutoring agents provided instructional support in the experimental and control conditions, respectively. The personalised tutoring agent used profile information to adapt support within the shared problem-solving framework, whereas the general tutoring agent provided non-profile-responsive guidance. This design enables the comparison of general guidance and personalised support while examining whether the effectiveness of AI tutoring depends on the organization of instructional support rather than merely the presence of AI.

Research Questions

RQ1. Does a dual-agent AI tutoring system providing profile-responsive support within a stage-specific problem-solving framework improve primary students' mathematical word problem-solving performance compared with non-profile-responsive AI guidance?

RQ2. Does the effectiveness of the dual-agent AI tutoring system vary across students with different learning profiles?

Based on the preceding literature, profile-responsive support embedded within a stage-specific problem-solving framework is expected to produce greater overall learning gains than non-profile-responsive AI guidance, while potential differences across learning profiles are examined exploratorily.

LITERATURE REVIEW

Mathematical Problem-Solving

Problem-solving has been described as a developmental process from understanding to planning, execution, and reflection (Polya, 1945). Engagement is also central to this process, as Valquaresma (2024) argued that it supports persistence during problem-solving. Sujatha and Vinayakan (2023) identified active learning, the use of technology, collaborative learning, and differentiated instruction as strategies for supporting student engagement in mathematics learning. In addition, Tuohilampi and Jiang (2024) proposed the ProInterest model, which identifies features of mathematics tasks that can spark and sustain students' situational interest.

Mathematical problem-solving has often relied on Polya's (1945) problem-solving model. The model has four main stages: understanding the problem, planning, carrying out the plan, and looking back (**Figure 1**). In simple terms, students first need to understand what the problem is asking and what information is given. They then decide on a possible method, use that method to solve the problem, and finally check whether the answer makes sense. In this study, Polya's (1945) model is applied to organize personalised AI support so that students receive guidance at different stages of the problem-solving process.

Table 1. GenAI agent tutoring strategies and implementation options across Polya's (1945) stages

Stage	Tutoring strategy	Technical methods
Understanding	Reading and understanding problem statements	Prompt engineering; CoT
	Extract key information (givens, unknowns, constraints)	Prompt engineering; Standardizing math language
Planning	Question-based guidance and follow-up prompts	Prompt engineering; CoT; ReAct; Tool module; RAG
	Clue-based prompting (partial cues)	Prompt engineering; RAG
	Real-life contextualization	Prompt engineering ; RAG
	Verbal encouragement and praise	Prompt engineering (role)
Executing	Immediate verification and checking	Prompt engineering; ReAct; Tool module; RAG
	Encourage productive trial and error	Prompt engineering (role)
Reflection	Error analysis and brief reflection	Prompt engineering (skills, examples)
	Transfer tutoring (similar or scaffolded tasks)	Prompt engineering; RAG; Tool module

Note. The listed technical methods indicate implementation options rather than fixed requirements; Prompt engineering is always used to control roles and skills, while mathematical tools and RAG are triggered only when the tutoring move requires calculation, verification, or retrieval of relevant resources

AI-Supported Scaffolding Across Problem-Solving Stages

Effective instruction of mathematical problem-solving can provide students with the necessary guidance and support by following Polya's (1945) four-stage framework. Lee (2017) pointed out that teachers can effectively support students' learning by providing targeted cognitive prompts and feedback for each stage. To replicate this established pedagogical scaffolding digitally, this study uses a set of technical methods centered on prompt engineering.

Although prompt engineering is the fundamental mechanism for controlling the behavior of the agent, more specialized technologies are selectively activated: chain-of-thought (CoT) helps the agent break down complex logic into a series of steps (Zhang et al., 2025); the reasoning and acting (ReAct) framework enables the agent to combine reasoning with internal verification (Yao et al., 2022), while retrieval-augmented generation (RAG) enables the agent to extract accurate information from the course database (Singh et al., 2025). These implementation options are summarized in **Table 1** and applied across the following four stages.

Stage 1. Understanding the problem

Research indicates that teachers must first assist students in identifying key information and translating verbal descriptions into mathematical representations (Rahmawati & Anwar, 2020). Specifically, the AI tutoring agent focuses on assisting students to comprehend information as well as represent it. The agent helps students to identify key information, label it, accordingly, categorize relevant data, and finally convert spoken information into symbols. The agent asks students questions such as, "How many total heads and legs are given in the problem?" or "Can you tell me how many legs a chicken and a rabbit each have?" The agent also provides students with immediate, affirming feedback while helping them identify mistakes before these develop into more substantial learning difficulties.

Stage 2. Making a plan

In the planning stage, effective tutors scaffold learning by prompting students to use heuristic strategies or draw analogies (Tay & Toh, 2023). This helps students select and refine an appropriate solution strategy. For example, the agent may say: "Let's assume all the animals in the cage are chickens. What would happen to the total number of legs?"

Stage 3. Executing the plan

At this stage, teachers typically encourage self-monitoring and accuracy tracking while students carry out their plans (Oudman et al., 2022). This enables the student's strategy to remain mathematically sound. For example, if a student calculates a result, the agent might intervene: "You calculated 5 rabbits. Do these 5 rabbits and the rest of the chickens equate to the total number of legs in our problem?" This supports students in monitoring the accuracy of their reasoning as they execute the selected strategy.

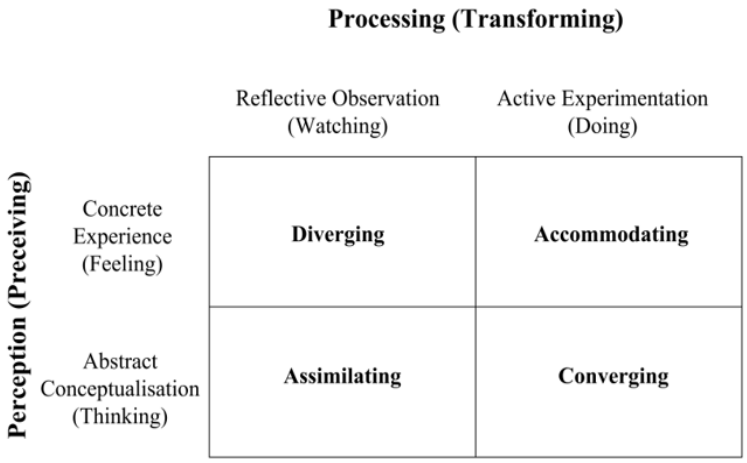


Figure 2. Mapping of Kolb's (2014) learning profiles based on perception and processing dimensions (created by the authors based on Kolb, 2014)

Stage 4. Reviewing and reflecting

The final stage of problem-solving requires teachers to guide students in consolidating their learning and transferring strategies to new contexts, thereby improving their critical thinking skills (Dookurong et al., 2025). With this instructional framework, the agent helps students review how they arrived at a solution through knowledge-based questions. In addition, the agent encourages students to reflect on how their solution strategy might be applied to other situations. For example, "You used the assumption method. Can we use another method such as setting up an equation to solve this problem?" or "How would we be able to apply our strategy to a problem concerning bicycles and tricycles?"

Learning Profiles and Personalised Support

The experiential learning theory (ELT) framework (Kolb, 2014), originating from Dewey's (1916) view of learning by doing, can inform personalised instruction. Since individuals experience learning situations differently, Kolb's (2014) framework suggests that instructional support should be tailored to these personal variations. Students differ in how they perceive a learning situation: they may rely on feeling or thinking. Similarly, they differ in how they process their learning: this occurs through either watching or doing. Based on the combination of these two dimensions, Kolb (2014) identified four distinct learning profiles: diverging, accommodating, converging, and assimilating (see **Figure 2**).

Kolb's (2014) ELT framework has long and widely been applied in educational research, and the theory itself has proven useful. However, the psychometric characteristics of the learning style inventory, which is used to distinguish the four learning styles (**Figure 2**), have been the subject of substantial debate. Criticisms have raised concerns about the reliability of the scale, for example due to the forced item structure (Calderón Carvajal et al., 2021). These concerns are particularly relevant when the scale is used to inform high-stakes educational decisions. However, these critiques should be distinguished from the broader ELT framework itself, which continues to inform research on how learners engage with different forms of experience. In the present study, the learning profiles are therefore treated as provisional design heuristics and an exploratory analytical lens rather than as fixed or stable learner traits.

Within Kolb's (2014) ELT framework, different learning profiles are associated with different preferences for instructional support. Diverging learners often benefit from visual aids and multiple representations. Accommodating learners tend to prefer hands-on experience. In contrast, converging learners focus on logical reasoning and model efficiency, while assimilating learners prioritize conceptual structure and systematic reflection (Kolb, 2014).

Some studies have reported benefits when digital support is adapted to learner characteristics. For example, Akbulut and Cardak (2012) reported positive student experiences in digital environments adapted to Kolb's (2014) profiles. Additionally, Huertas et al. (2017) associated adaptive scaffolding with students' problem-solving capabilities and self-regulation. Specifically in mathematics, Hwang et al. (2012) reported

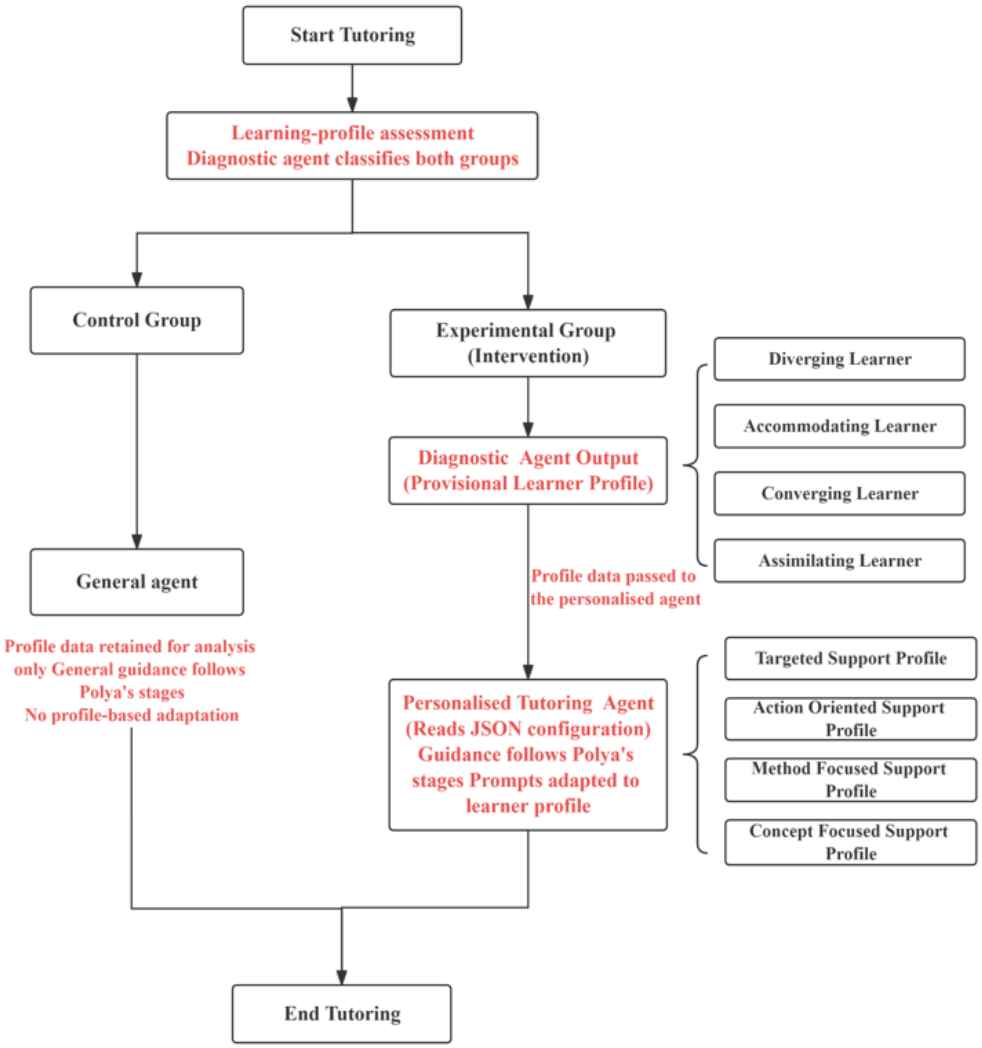


Figure 3. Logical design of the pedagogical intelligent agents, illustrating the profile data transfer from classification to intervention (created by the authors)

higher learning performance and interest among students using a personalised educational game tailored to Kolb (2014) learning styles than among those in non-personalised environments. However, these findings do not establish that matching instruction to a labelled learning style is itself the causal mechanism underlying improved learning outcomes.

METHOD

System Design and Implementation

Agent roles and interaction logic

Across the two study conditions, three agent implementations were used. The experimental condition employed a dual-agent configuration consisting of a diagnostic agent and a personalised tutoring agent, whereas the control condition used a general tutoring agent. The diagnostic agent generated provisional learning-profile classifications from the adapted Kolb (2014) inventory. Students in both conditions completed the profiling procedure so that profile information could also be used in the exploratory subgroup analyses; however, only in the experimental condition was the resulting profile passed to the personalised tutoring agent and used to adapt instructional support.

The two conditions used different tutoring agents built on the same underlying architecture and common stage-specific instructional principles. The general tutoring agent was used in the control group and delivered

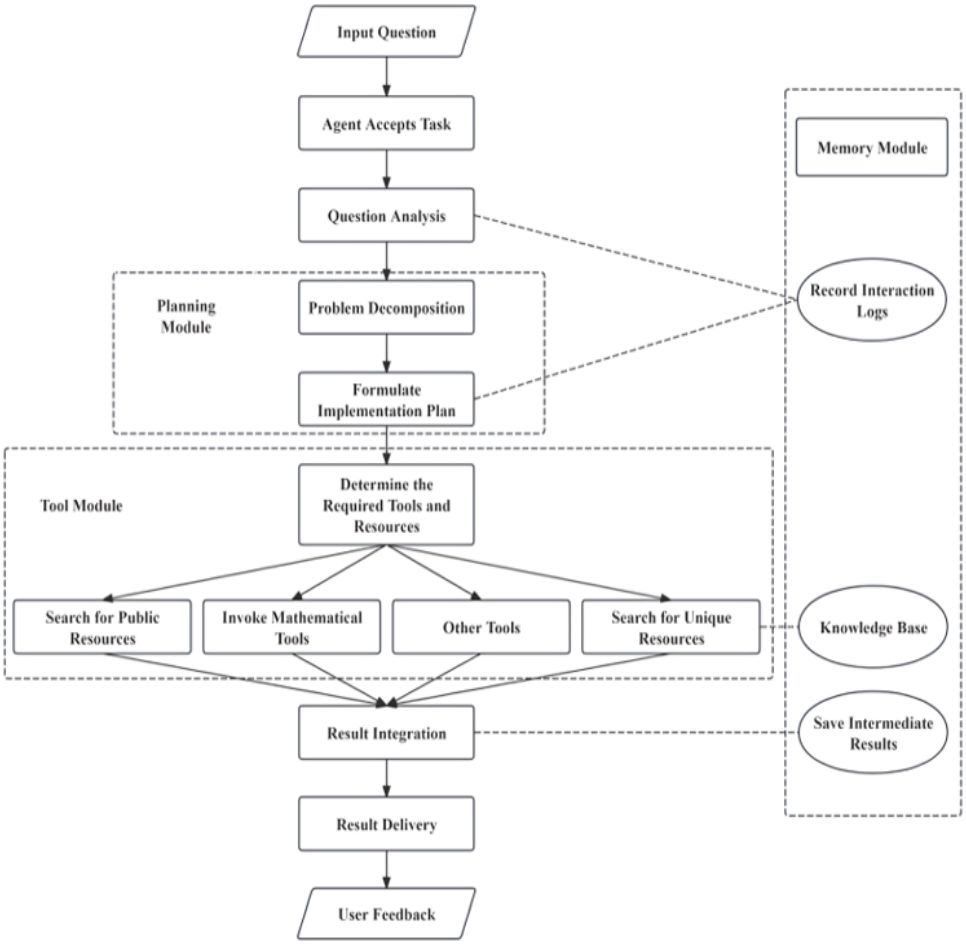


Figure 4. Modular architecture of the agent comprising planning, memory, and tool (created by the authors)

the same non-profile-responsive guidance to all students. The personalised tutoring agent was used in the experimental group and followed the same common problem-solving sequence but additionally adapted the timing and form of support according to the learner's provisional profile. Thus, the key distinction between the two tutoring agents was not the underlying tutoring architecture itself, but whether learner-profile information was incorporated into instructional adaptation (**Figure 3**).

System architecture and modules

Both tutoring agents shared the same underlying agentic architecture, adapted from Wu et al. (2024). The architecture implemented the common stage-specific instructional principles used in both conditions and was organized into three modules: planning, memory, and tool (**Figure 4**).

1. **Planning module:** The planning module managed the sequential structure of the tutoring process and selected appropriate prompt templates according to the current stage of Polya's (1945) problem-solving framework.
2. **Memory module:** The memory module maintained relevant learner-context information across multi-step interactions and retrieved curriculum-aligned materials from the curated knowledge base using RAG.
3. **Tool module:** The tool module incorporated calculators and verification tools that the tutoring agents could call when a problem-solving step required computation or checking.

The memory module supported continuity across Polya's (1945) four stages. Its knowledge base contained the 2022 Chinese mathematics curriculum standards (Ministry of Education of the People's Republic of China, 2022), grade 5 textbooks, and supplementary exercise resources (**Table 2**). These materials were intended to improve curriculum alignment, mathematical relevance, and age appropriateness (**Appendix A**).

Table 2. Summary of sample data for supplementary mathematics materials

Material category	Material name
Curriculum standard	Mathematics curriculum standards for compulsory education (Ministry of Education of the People's Republic of China, 2022)
Authoritative textbook	People's Education Press grade 5 mathematics textbooks for the first and second semesters, used as sources of related application problems (Primary Mathematics Textbook Editorial Committee, Curriculum and Teaching Materials Research Institute, People's Education Press, 2022b, 2022c)
Practice resources	Primary school grade 5 supplementary practice resource library
	Typical examples and solutions (standard methods and step-by-step breakdown examples)
	Typical error problems and analysis (compilation of error-prone points and error correction methods)
	Mathematics exercises by difficulty level (basic questions, advanced questions, challenging questions, etc.)
Supplementary materials	Teaching plan and learning guide for grade 5 primary school students
	Instructional method guidance and interactive teaching cases
	Mathematical culture and enrichment of mathematical fun (mathematical stories, mathematical history, interdisciplinary)

Table 3. Summary of personalised tutoring strategies and corresponding technical methods (experimental group only)

Learner profile	Tutoring strategy	Stage	Technical methods
Diverging	Information trap method	Understanding	Prompt engineering (skills, examples)
	Backward reasoning/goal guidance	Planning	Prompt engineering; CoT
	Regular checks	Executing	Prompt engineering; Tool module; RAG
Accommodating	Development of presentation skills	Planning	Prompt engineering ; CoT; ReAct
	Challenging practical tasks	Reflection	Prompt engineering; RAG; tool module
Converging	Multi-method demonstrations	Planning	Prompt engineering; CoT; RAG
	Independent variation and explanation	Reflection	Prompt engineering; ReAct; RAG; tool module
Assimilating	Demonstration/communication trap method	Planning	Prompt engineering; RAG; tool module
	Answer trap method	Executing	Prompt engineering; CoT; ReAct; RAG
	Differential comparison method	Reflection	Prompt engineering; RAG; tool module

Note. The strategies described above are the personalised intervention strategies that differentiate the experimental group from the control group; the technical means (tool module, RAG) are called upon only when a profile-specific tutoring action involves a calculation or a resource retrieval

The Personalised Tutoring Experimental Design

In addition to the shared architecture and stage-specific instructional principles described above, the personalised tutoring agent used in the experimental condition incorporated profile-responsive adaptations. These adaptations were based on the provisional learner profile generated by the diagnostic agent and were embedded at selected points within the shared Polya-based problem-solving sequence. Profile-sensitive supports were embedded as lightweight interventions at selected points rather than implemented as four entirely separate agents. One or two targeted prompts were used at each stage to preserve continuity and limit cognitive load. The design followed three principles: stage alignment, limited intervention density, and emphasis on reasoning processes rather than answer delivery. Some prompts contained a plausible but incorrect statement or an irrelevant condition and asked the learner to evaluate it. These contrastive verification prompts were intended to elicit checking and independent judgement; they were not presented as authoritative answers (**Table 3**).

For diverging learners (feeling and watching), the personalised tutoring agent prioritized support during understanding and planning. It highlighted key information, used irrelevant-information filtering prompts, compared relevant with irrelevant conditions, guided learners back to the task goal, and inserted brief checks of intermediate reasoning. These supports were intended to stabilize attention and strengthen problem representation before execution.

For accommodating learners (feeling and doing), who benefit from experiential learning, the personalised tutoring agent emphasized guided practice and contextual transfer during the planning and reviewing phases. It offered moderately challenging contextualized problems with gradually reduced scaffolding and prompted

students to explain their reasoning and reflect on misconceptions. Together, these supports formed a cycle of error detection, reflection, correction, and internalization, with an emphasis on transferring strategies across related problem contexts.

For the converging learners (thinking and doing), who prefer a structured approach to reasoning, the personalised tutoring agent emphasized the identification and comparison of multiple solution paths during the planning phase, with a focus on the differences in costs, applicability, and generality. During the reviewing phase, students were encouraged to adapt to the conditions and solve the problems, thus breaking away from a single routine. This reduced reliance on a single routine and encouraged flexible thinking. The integrated chain emphasized path expansion and innovation during both planning and review phases.

For assimilating learners (thinking and watching), the personalised tutoring agent emphasized conceptual coherence and metacognitive monitoring during execution and review. It used contrastive verification prompts, comparison between student work and worked solutions, and questions that asked learners to identify the first point at which their reasoning diverged. These supports were intended to encourage justification, consistency checking, and self-correction.

Participants, Ethics, and Experimental Design

Participants were grade 5 students from six parallel classes at an anonymized primary school in China. Three classes were assigned to the experimental condition (n = 126 before attrition) and three to the control condition (n = 121 before attrition). The school routinely used digital learning tools, allowing the intervention to be implemented within the normal classroom process. After exclusions for incomplete pre-test, post-test, profile, or platform records, the analytic sample comprised 206 students: 99 in the experimental group and 107 in the control group.

Ethical approval and participant information

The study was conducted with permission from the participating school and was reviewed and approved by the Medical Ethics Committee at Changzhi Medical University (approval no. RT2025057). The requirement for signed written informed consent was waived by the ethics committee. Students and their parents or legal guardians were informed about the study procedures, the intended use of the data, privacy and confidentiality protections, and their right to withdraw from participation. All data were anonymized before analysis, and no personal identifiable information was reported in the manuscript.

The problem-solving tasks in the formal experiment were selected based on the opinions of the participating teachers. During the pre-experiment stage, the teachers conducted pilot tests on the initial version and provided feedback that although the classroom process was feasible, the initial task was too abstract and failed to fully stimulate the diversity of students' thinking. Based on this feedback, the research team modified the task by adding richer contextual information, making the connection between the task and classroom application clearer, and allowing for more than one reasonable solution path. The revised version was reviewed by the teachers again and was determined to be suitable for assessing both students' problem-solving processes and the instructional effectiveness of personalised tutoring agents.

The research team adapted the focal task from the traditional chicken and rabbit in the same cage problem in unit 9 of the People's Education Press grade 4 mathematics textbook for the second semester (Primary Mathematics Textbook Editorial Committee, Curriculum and Teaching Materials Research Institute, People's Education Press, 2022a, p. 99). The focal intervention task comprised three related problem-solving items within the same mathematical context. The English rendering of the core problem as used in the intervention is as follows. "A cage contains some chickens and rabbits. Counting from the top, there are 45 heads. Counting from the bottom, there are 100 legs. How many chickens and how many rabbits are there in the cage?" The task requires students to set up and solve relationships between quantities. It also supports observation of how students move through understanding, planning, execution, and looking back while interacting with the pedagogical agent.

The task was designed to align with the grade 5-grade 6 objectives in the Chinese mathematics curriculum standards (Ministry of Education of the People's Republic of China, 2022, section III, stage 3). In this study,

Table 4. Procedural timeline for the quasi-experimental phase

Stage	Time	Participants	Activity	Data
1. Pre-test	20 min	Both groups	Two parallel word problems completed without AI support	Baseline rubric score on the common analytic scale
2. Profiling	5-8 min	Both groups	12-item adapted forced-ranking inventory; profile used for adaptation only in the experimental group	Item rankings, dimension scores, and provisional profile
3. Intervention	45 min	Both groups	One focal task comprising three related problem-solving items; personalised versus general tutoring agent	Interaction logs and solution paths
4. Post-test	20 min	Both groups	Two parallel post-test problems plus one transfer item excluded from the learning-gain calculation completed without AI	Post-test rubric score on the common analytic scale

those objectives were operationalized as identifying and representing quantitative relationships, reasoning through a real-world problem, and using a simple mathematical model.

Experimental Procedure

The formal quasi-experimental phase followed the sequence summarized in **Table 4**.

1. **Pre-test of word-problem-solving performance:** Before the intervention, all students completed a 20-minute pre-test containing two parallel word problems. The tasks assessed the coordination of multiple constraints using an assumption method or a one-variable linear equation. To limit direct practice effects, the pre-test used contexts that differed from the focal intervention problem while preserving comparable mathematical structure and cognitive demand.

   The assessment problems were drawn from standard grade 5 supplementary mathematics materials and reviewed by two experienced mathematics teachers for curriculum alignment. Baseline equivalence between the experimental and control groups was assessed using an independent-samples t-test on pre-test scores.

2. **Digital Survey and Profiling:** Students logged in on individual devices and completed the digital profiling stage, which took approximately 5-8 minutes. Both groups completed the same digitally adapted Kolb (2014) inventory. In the experimental group, the diagnostic result was passed to the personalised tutoring agent; in the control group, the profile data were collected but were not used to adapt instructional support. The teacher team simplified item wording after consultation with course instructors while retaining the four dimensions of concrete experience (CE), reflective observation (RO), abstract conceptualization (AC), and active experimentation (AE).

   The adapted wording used age-appropriate statements about feeling, observing, thinking, and doing. The diagnostic agent used the completed rankings to generate one of four provisional profiles: diverging, assimilating, converging, or accommodating. These classifications were used only to select profile-responsive prompts in the experimental condition.

Learning-profile instrument: Adaptation, classification, and psychometric considerations

The digitally adapted inventory contained 12 forced-ranking items. For each item, students ranked four statements from 1 (most like me) to 4 (least like me). Scores were summed for CE, RO, AC, and AE; a complete response produced a total of 120 across the four dimensions. Profiles were assigned using prespecified difference scores: CE-AC > 0 and RO-AE > 0 indicated diverging; CE-AC > 0 and RO-AE < 0 indicated accommodating; CE-AC < 0 and RO-AE < 0 indicated converging; and CE-AC < 0 and RO-AE > 0 indicated assimilating. The age-adapted wording was reviewed by the teacher team and course instructors for clarity and content relevance. Because the forced-ranking format produces ipsative scores, the classifications were used as provisional pedagogical design heuristics rather than fixed psychological diagnoses, and profile-specific analyses were treated as secondary to the overall group comparison.

3. **Intervention:** The tutoring intervention lasted 45 minutes and centered on the same focal mathematical task in both conditions. Within this task, students completed three related problem-solving items. The experimental group interacted with the personalised tutoring agent, which varied

prompts according to the provisional learning profile and the current Polya (1945) stage. The control group interacted with the general tutoring agent, which delivered the same non-profile-responsive guidance to all students. The mathematical content, available time, devices, and classroom conditions were otherwise held constant.

4. **Post-test and data collection:** Following the intervention, students completed a 20-minute post-test without AI assistance. The post-test contained three word problems. The first two items were parallel to the pre-test items in mathematical structure and cognitive demand and were used to calculate the post-test score for the learning gain analysis. The third item was a supplementary variation task intended to probe strategy transfer and was not included in the learning-gain analyses reported in this study. Interaction logs were archived during the tutoring session for process-oriented follow-up analyses.

Scoring rubric and rating procedure

Prior to scoring, the research team and participating mathematics teachers jointly established the analytic scoring rules and point allocations. The rubric assessed four components aligned with Polya's (1945) framework: accurate representation of givens and unknowns, formulation of a viable strategy or equation, mathematically accurate execution, and checking or explanation of the result. Full-credit anchor responses contained a correct representation, a valid strategy, complete working, a correct conclusion, and an appropriate check. Partial-credit anchors reflected a viable approach with incomplete execution or a local computational error, whereas minimal-credit responses identified limited relevant information without a workable strategy.

Two mathematics teachers independently scored the student responses using the agreed analytic rubric. Where discrepancies occurred, the two raters reviewed the relevant responses and reached consensus through discussion. The same rubric and scoring procedure were applied to both the pre-test and post-test. Since the archived dataset retained the final consensus score for each student at each assessment occasion, these consensus scores were used in all subsequent analyses.

Outcome metric, attrition, and statistical analysis

Learning gain was calculated using only the two parallel pre-test and post-test items. These matched assessment components were comparable in mathematical structure and cognitive demand and were evaluated using the same analytic rubric and common scoring metric. The additional post-test variation item was included as a supplementary transfer task but was not included in the learning-gain analyses reported in this study. Learning gain was therefore defined as the post-test score minus the pre-test score for the two matched assessment components. Attrition by condition was examined using a Pearson Chi-square test, and baseline equivalence in the retained sample was assessed using an independent-samples t-test. The primary analysis compared mean learning gains between conditions. Distributional assumptions were examined using Shapiro-Wilk tests, and homogeneity of variance was assessed using Levene's test. Hedges' g was used to quantify standardized effect sizes, with 95% confidence intervals (CIs) reported for mean differences and effect estimates. The final analytic sample consisted of 206 students. Profile-specific comparisons were prespecified as exploratory and evaluated using standard independent-samples t-tests because the homogeneity-of-variance tests were non-significant within all four profiles. Holm's sequential procedure controlled the familywise error rate across the four comparisons. An omnibus regression including condition, learning profile, and the condition × profile interaction was estimated using HC3 robust standard errors to test whether the intervention effect differed across profiles. As a sensitivity analysis, post-test performance was regressed on experimental condition and centered pre-test score using HC3 robust standard errors.

RESULTS

Overall Learning Gains

Of the 247 initially enrolled students, 27 of 126 experimental-group students (21.4%) and 14 of 121 control-group students (11.6%) were excluded because of incomplete records. Attrition differed by condition, $\chi^2(1) = 4.33$, $p = .037$. In the retained sample, no statistically significant baseline difference was detected: the

Table 5. Descriptive statistics of learning gains in the experimental and control groups

	Group	Number	Mean	SD	Standard error
Learning gain	Experimental group	99	5.14	7.49	0.75
	Control group	107	0.94	6.31	0.61

Note. Learning gain was computed as post-test score minus pre-test score; both assessments used parallel tasks and the same analytic rubric and scoring scale

Table 6. Independent-samples t-test for overall learning gains

Analysis	t	df	p	Mean difference	Standard error	95% CI
Equal variances	4.364	204	< .001	4.200	.962	[2.303, 6.098]
Welch sensitivity	4.336	192.371	< .001	4.200	.969	[2.290, 6.111]

Note. Levene's test was non-significant, so the equal-variance analysis is the primary inferential result; the Welch analysis is reported as a sensitivity check and leads to the same conclusion

Table 7. Descriptive statistics of post-test–pre-test scores by learning profiles and groups

Type	Group	N	Mean	SD	Standard error
Diverging	Experimental group	24	1.54	6.81	1.39
	Control group	27	-1.63	5.55	1.07
Accommodating	Experimental group	24	5.25	6.63	1.35
	Control group	26	1.77	5.12	1.00
Converging	Experimental group	23	9.44	6.46	1.35
	Control group	20	4.85	7.68	1.72
Assimilating	Experimental group	28	4.60	8.05	1.52
	Control group	34	0.05	5.82	1.00

Note. Gain scores were computed from the two matched pre- and post-test items; the additional transfer item was excluded from the learning-gain calculation

Table 8. Holm-adjusted exploratory learning-gain comparisons by provisional profile

Profile	t (df)	Raw p	Holm p	Hedges' g	95% CI	Holm result
Diverging	1.83 (49)	.0733	.1181	.51	[-.05, 1.06]	No
Accommodating	2.09 (48)	.0419	.1181	.58	[.02, 1.14]	No
Converging	2.13 (41)	.0394	.1181	.64	[.03, 1.24]	No
Assimilating	2.58 (60)	.0124	.0496	.65	[.14, 1.16]	Yes

Note. Holm-adjusted p values control the familywise error rate across four exploratory profile comparisons; homogeneity-of-variance tests were non-significant for all four profiles; Hedges' g and approximate confidence intervals were calculated from the group means, SDs, and sample sizes; yes indicates significance after Holm correction; the omnibus group-by-profile interaction was non-significant, F(3, 198) = 0.15; p = .931

experimental group had a pre-test mean of 9.28 (standard deviation [SD] = 3.80) and the control group a mean of 9.16 (SD = 3.24), Welch's t(193.29) = 0.24, p = .811, Hedges' g = .03. Learning-gain distributions did not significantly depart from normality in either the experimental group, W = .988, p = .526, or the control group, W = .979, p = .084. Levene's test was non-significant, F = 2.01, p = .158.

The experimental group achieved a mean gain of 5.14 points (SD = 7.49), compared with 0.94 points (SD = 6.31) in the control group (**Table 5**). The mean difference was 4.20 points, t(204) = 4.36, p < .001, 95% CI [2.30, 6.10], Hedges' g = .61, 95% CI [.33, .88]. The unequal-variance sensitivity result was nearly identical, Welch's t (192.37) = 4.34, p < .001. A baseline-adjusted analysis also supported the intervention effect: the experimental condition predicted a 4.20-point higher post-test score, b = 4.20, SE = .98, z = 4.28, p < .001, 95% CI [2.28, 6.13] (**Table 6**).

Exploratory Profile-Specific Contrasts

Table 7 and **Table 8** report exploratory learning-gain contrasts by provisional profile. The unadjusted p values were .0733 for diverging learners, .0419 for accommodating learners, .0394 for converging learners, and .0124 for assimilating learners. After Holm correction, only the assimilating-profile contrast remained significant at the .05 level (adjusted p = .0496). However, the omnibus group-by-profile interaction was not significant, F(3, 198) = 0.15, p = .931. Thus, the data support an overall intervention effect but do not provide evidence that the size of that effect differed reliably across the four profiles.

Diverging learners: A positive but non-significant contrast

For diverging learners, the experimental group included 24 students and the control group included 27 students. Mean gains were 1.54 (SD = 6.81) and -1.63 (SD = 5.55), respectively. The contrast was $t(49) = 1.83$, $p = .0733$, Holm-adjusted $p = .1181$, mean difference = 3.17, 95% CI [-0.31, 6.65], Hedges' $g = .51$, 95% CI [-.05, 1.06]. The contrast was positive in direction but remained non-significant after Holm adjustment.

Accommodating learners: Exploratory contrast after correction

For accommodating learners, the experimental group included 24 students and the control group included 26 students. Mean gains were 5.25 (SD = 6.63) and 1.77 (SD = 5.12), respectively. The contrast was $t(48) = 2.09$, $p = .0419$, Holm-adjusted $p = .1181$, mean difference = 3.48, 95% CI [0.13, 6.84], Hedges' $g = .58$, 95% CI [.02, 1.14]. Although the unadjusted comparison reached the .05 level, it did not remain significant after Holm correction.

Converging learners: Exploratory contrast after correction

For converging learners, the experimental group included 23 students and the control group included 20 students. Mean gains were 9.44 (SD = 6.46) and 4.85 (SD = 7.68), respectively. The contrast was $t(41) = 2.13$, $p = .0394$, Holm-adjusted $p = .1181$, mean difference = 4.59, 95% CI [0.23, 8.94], Hedges' $g = .64$, 95% CI [.03, 1.24]. The unadjusted contrast was significant, but it did not survive Holm correction.

Assimilating learners: Holm-adjusted contrast

For assimilating learners, the experimental group included 28 students and the control group included 34 students. Mean gains were 4.60 (SD = 8.05) and 0.05 (SD = 5.82), respectively. The contrast was $t(60) = 2.58$, $p = .0124$, Holm-adjusted $p = .0496$, mean difference = 4.55, 95% CI [1.02, 8.08], Hedges' $g = .65$, 95% CI [.14, 1.16].

This was the only profile-specific contrast that remained significant after Holm correction. However, the non-significant omnibus interaction indicates that this finding should not be interpreted as evidence of reliable differences in intervention effects across profiles.

DISCUSSION

The primary finding is that students in the personalised dual-agent condition achieved greater learning gains than those receiving general guidance. The intervention combined profile-responsive support with stage-specific scaffolding, and the overall effect was moderate in standardized terms and remained robust after baseline adjustment. This pattern is consistent with studies reporting benefits from adaptive digital environments (Akbulut & Cardak, 2012; Huertas et al., 2017; Hwang et al., 2012), while extending this line of work by examining a design that separates diagnostic and instructional functions and maps support to Polya's (1945) problem-solving stages. In line with Gibson et al. (2023), the result is best understood as evidence for a pedagogically structured support architecture rather than a generic effect of AI technology.

Although only the assimilating-profile contrast remained statistically significant after Holm correction, the estimated between-condition gain differences were positive across all four profiles and were relatively similar in magnitude (3.17-4.59 points). Thus, the significant overall group effect is not inconsistent with the profile-specific results. The overall analysis used the full sample, whereas the profile-specific analyses involved substantially smaller subgroups and were additionally subject to multiplicity correction, resulting in lower statistical precision.

For diverging, accommodating, and converging learners, the absence of Holm-adjusted significance should therefore not be interpreted as evidence of no benefit. One possible explanation is that a single-session intervention provided insufficient exposure for profile-responsive differences to emerge reliably. Diverging learners may require more time for exploration and representation building, accommodating learners may benefit from repeated cycles of practice, feedback, and revision, and converging learners may require repeated opportunities for method comparison and transfer. These explanations remain hypotheses and should be tested directly in multi-session studies using process-log measures. This interpretation is also

consistent with the non-significant condition-by-profile interaction, which provided no evidence that the intervention effect differed reliably across profiles.

The assimilating contrast provides the strongest profile-specific signal in the present dataset. Its direction is theoretically compatible with prompts that require justification, comparison with a valid solution, and identification of inconsistency. Replication with an independently validated profile measure and longer exposure is needed before translating this result into a stable design rule.

The findings also support the distinction between content delivery and process scaffolding. Schorcht et al. (2024) found that prompt techniques may change process-related quality even when content-related quality is unchanged, while Apoki et al. (2022) described pedagogical agents as adaptive learning partners. In the present system, the educational role of the technology lies in regulating attention, eliciting explanations, checking intermediate steps, and supporting reflection across a shared problem-solving sequence.

Overall, the study shows that the timing and structure of support can be as important as the amount of information provided. This aligns with Gibson et al.'s (2023) argument about the importance of instructional design of technology-based pedagogical systems. The dual-agent architecture makes this pedagogical logic explicit: the diagnostic agent supplies provisional profile information, while the personalised tutoring agent enacts stage-specific scaffolding. The robust overall effect, together with the conservative multiplicity-adjusted analysis, provides support for the architecture as a promising instructional design while appropriately limiting claims about differential profile matching.

IMPLICATIONS, LIMITATIONS, AND CONCLUSION

Theoretical Implications

The study contributes a process-oriented account of dual-agent personalization by specifying how diagnostic information can be separated from instructional action and how tutoring prompts can be organized across understanding, planning, execution, and reflection. Rather than positioning intelligent agents primarily as answer providers, the design treats them as external supports for goal management, strategy selection, monitoring, and verification. The results support further investigation of stage-specific process scaffolding while showing that profile-specific mechanisms require separate empirical confirmation.

Practical Implications For Agent Design and Future Work

Strengthening profile-responsive support across learner groups

Future systems should test profile-responsive supports across multiple sessions rather than assuming that a single exposure is sufficient. For diverging learners, this may involve user-controlled pacing, additional exploration time, clearer representations, and repeated goal-restatement. For accommodating learners, repeated cycles of attempt, explanation, feedback, and revision may provide more opportunities for experiential learning. For converging learners, repeated method comparison, path selection, and transfer across related problems may help clarify the added value of structured strategy support. These features should be evaluated prospectively rather than inferred from the present subgroup contrasts.

Improving transparency and stability through process monitoring

Future work should combine outcome scores with process evidence. Relevant measures include response latency, time in each problem-solving stage, prompt uptake, representation changes, verification attempts, and revisions following feedback.

Linking these log indicators to subsequent performance would clarify whether a prompt changed the learner's reasoning, was ignored, or merely delayed task completion. Such analyses would provide a more direct test of the proposed instructional mechanism than retrospective interpretation of group means.

The system should also record whether requested checks were completed and whether checking led to revision. This would connect instructional intent with observable learner behavior and reduce reliance on post hoc explanations.

Refining adaptive support without increasing cognitive load

A useful design principle is to preserve the shared problem-solving flow while limiting each stage to one or two targeted supports. Future versions should vary prompt density experimentally so that additional support does not inadvertently increase cognitive load.

Limitations

Several limitations define the scope of the conclusions. First, the adapted 12-item forced-ranking instrument was designed for pedagogical profiling and should be further evaluated for classification stability and external validity in primary-school samples. Second, the archived scoring file retained final rubric scores rather than separate independent rater records; future studies should preserve independent ratings alongside consensus scores. Third, attrition was higher in the experimental condition, although no statistically significant baseline difference was detected in the retained sample and the baseline-adjusted analysis yielded a similar overall effect estimate. Fourth, the single-session intervention, one focal problem type, modest profile-specific sample sizes, and recruitment from one school limit generalizability. In addition, since students were recruited from six intact classes, potential classroom-level clustering represents an additional limitation and should be addressed using larger numbers of classes and cluster-aware analytical approaches in future research. Finally, the significant assimilating contrast occurred alongside a non-significant omnibus interaction, so profile-specific design claims should be tested in preregistered, adequately powered replications.

Conclusion

This study provides evidence that a pedagogically structured dual-agent system was associated with greater short-term gains in grade 5 mathematical problem-solving than a general tutoring agent. The overall effect was statistically robust, moderate in magnitude, and confirmed after adjustment for baseline performance. The contribution lies in separating diagnostic and tutoring roles and organising prompts around a transparent problem-solving sequence, thereby providing a structured model for personalised learning support. The assimilating-profile contrast remained significant after Holm correction, but the null omnibus interaction indicates that the primary conclusion concerns the overall advantage associated with the support architecture rather than reliable differences among learner profiles. Future multi-session studies should test the proposed pacing, representation, monitoring, and profile-responsive mechanisms directly.

Author contributions: RW: conceptualization, formal analysis, methodology, writing – original draft; **LT:** supervision, writing – review & editing. Both authors approved the final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Acknowledgments: The authors would like to thank the participating grade 5 students and their mathematics teachers for their valuable time and cooperation in this study. The authors would also like to thank the teachers for their helpful feedback during the pilot testing and task design phases.

Ethics declaration: This study was approved by the Medical Ethics Committee at Changzhi Medical University with approval number RT2025057. The requirement for signed written informed consent was waived by the committee. Students and their parents or guardians were informed about the study procedures, data use, privacy protection, and withdrawal rights. All data were anonymized before analysis, and no personal identifiable information was reported in the manuscript.

AI statement: During manuscript preparation and revision, the authors used ChatGPT (OpenAI) to refine the wording of selected passages and assist with checking the formatting of selected reference entries. The authors reviewed and edited the text and took full responsibility for the content of the published article.

Declaration of interest: The authors declared no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Akbulut, Y., & Cardak, C. S. (2012). Adaptive educational hypermedia accommodating learning styles: A content analysis of publications from 2000 to 2011. *Computers & Education, 58*(2), 835-842. https://doi.org/10.1016/j.compedu.2011.10.008

Apoki, U. C., Hussein, A. M. A., Al-Chalabi, H. K. M., Badica, C., & Mocanu, M. L. (2022). The role of pedagogical agents in personalised adaptive learning: A review. *Sustainability, 14*(11), Article 6442. https://doi.org/10.3390/su14116442

Calderón Carvajal, C., Ximénez Gómez, C., Lay-Lisboa, S., & Briceño, M. (2021). Reviewing the structure of Kolb's learning style inventory from factor analysis and Thurstonian item response theory (IRT) model approaches. *Journal of Psychoeducational Assessment, 39*(5), 593-609. https://doi.org/10.1177/07342829211003739

Dewey, J. (1916). *Democracy and education: An introduction to the philosophy of education*. Macmillan.

Dookurong, I. D., Yarhands, D. A., & Akwerty, E. (2025). A systematic review of the integration of collaborative learning and Polya's problem-solving strategy for enhancing pre-service teachers' critical thinking. *International Journal of Studies in Education and Science, 6*(4), 395-428. https://doi.org/10.46328/ijses.5520

Engelbrecht, J., & Borba, M. C. (2024). Recent developments in using digital technology in mathematics education. *ZDM Mathematics Education, 56*, 281-292. https://doi.org/10.1007/s11858-023-01530-2

Gašević, D., Siemens, G., & Sadiq, S. (2023). Empowering learners for the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 4*, Article 100130. https://doi.org/10.1016/j.caeai.2023.100130

Gibson, D., Kovanovic, V., Ifenthaler, D., Dexter, S., & Feng, S. (2023). Learning theories for artificial intelligence promoting learning processes. *British Journal of Educational Technology, 54*(5), 1125-1146. http://doi.org/10.1111/bjet.13341

Han, Y., Hong, S., Li, Z., & Lim, C. (2025). Defining and classifying the roles of intelligent learning companion systems: A scoping review of the literature. *TechTrends, 69*(3), 567-581. https://doi.org/10.1007/s11528-025-01058-0

Hong, X., & Guo, L. (2025). Effects of AI-enhanced multi-display language teaching systems on learning motivation, cognitive load management, and learner autonomy. *Education and Information Technologies, 30*(12), 17155-17189. https://doi.org/10.1007/s10639-025-13472-1

Huertas, A., López, O., & Sanabria, L. (2017). Influence of a metacognitive scaffolding for information search in b-learning courses on learning achievement and its relationship with cognitive and learning style. *Journal of Educational Computing Research, 55*(2), 147-171. https://doi.org/10.1177/0735633116656634

Hwang, G.-J., Sung, H.-Y., Hung, C.-M., Huang, I., & Tsai, C.-C. (2012). Development of a personalized educational computer game based on students' learning styles. *Educational Technology Research and Development, 60*(4), 623-638. https://doi.org/10.1007/s11423-012-9241-x

Jyrwa, K., & May, S. (2025). Learning difficulties and remedial teaching in mathematics: Insights and interventions for educators. *The Online Journal of Distance Education and e-Learning, 13*(1), 16-18. https://www.tojdel.net/journals/tojdel/volumes/tojdel-volume13-i01.pdf

Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development* (2nd ed.). FT Press.

Lai, L., She, L., & Li, C. (2024). Online teaching model in the context of blended learning environment: Experiential learning and TAM. *Education and Information Technologies, 29*(13), 17235-17259. https://doi.org/10.1007/s10639-024-12465-w

Lee, C. I. (2017). An appropriate prompts system based on the Polya method for mathematical problem-solving. *Eurasia Journal of Mathematics, Science and Technology Education, 13*(3), 893-910. https://doi.org/10.12973/eurasia.2017.00649a

Lin, C.-J., Lee, H.-Y., Wang, W.-S., Huang, Y.-M., & Wu, T.-T. (2025). Enhancing reflective thinking in STEM education through experiential learning: The role of generative AI as a learning aid. *Education and Information Technologies, 30*(5), 6315-6337. https://doi.org/10.1007/s10639-024-13072-5

López-Mendoza, A., Roque-Hernández, R. V., & Salazar-Hernández, R. (2026). Adoption and use of ChatGPT among university IT students: An exploratory study. *Contemporary Educational Technology, 18*(1), Article ep624. https://doi.org/10.30935/cedtech/17864

Mahmud, M. S., & Mohd Drus, N. F. (2023). The use of oral questioning to improve students' reasoning skills in primary school mathematics learning. *Frontiers in Education, 8*, Article 1126816. https://doi.org/10.3389/feduc.2023.1126816

Ministry of Education of the People's Republic of China. (2022). *Yiwu jiaoyu shuxue kecheng biaozhun (2022 nian ban)* [Mathematics curriculum standards for compulsory education (2022 edition)]. Beijing Normal University Press.

Mohamed, M. Z. B., Hidayat, R., Suhaizi, N. N. B., Sabri, N. B. M., Mahmud, M. K. H. B., & Baharuddin, S. N. B. (2022). Artificial intelligence in mathematics education: A systematic literature review. *International Electronic Journal of Mathematics Education, 17*(3), Article em0694. https://doi.org/10.29333/iejme/12132

Muhamad Fadzil, N., Osman, S., Ahmad, J., Jambari, H., & Said Husain, S. K. (2025). Enhancing students' problem-solving skills in algebra word problems: A systematic review of TAPPS and storyboarding strategies. *International Electronic Journal of Mathematics Education, 20*(4), Article em0850. https://doi.org/10.29333/iejme/16835

Nasongkhla, J., & Shieh, C.-J. (2025). AI robots as learning companions in PBL: Effects on cognitive and self-regulatory outcomes. *Contemporary Educational Technology, 17*(4), Article ep603. https://doi.org/10.30935/cedtech/17417

Oudman, S., van de Pol, J., & van Gog, T. (2022). Effects of self-scoring their math problem solutions on primary school students' monitoring and regulation. *Metacognition and Learning, 17*(1), 213-239. https://doi.org/10.1007/s11409-021-09281-9

Owan, V. J., Mohammed, I. A., Bello, A., & Shittu, T. A. (2025). Higher education students' ChatGPT use behavior: Structural equation modelling of contributing factors through a modified UTAUT model. *Contemporary Educational Technology, 17*(4), Article ep592. https://doi.org/10.30935/cedtech/17243

Polya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press. https://doi.org/10.1515/9781400828678

Powell, W., & Kusuma-Powell, O. (2011). *How to teach now: Five keys to personalized learning in the global classroom*. ASCD.

Primary Mathematics Textbook Editorial Committee, Curriculum and Teaching Materials Research Institute, People's Education Press. (2022a). *Yiwu jiaoyu jiaokeshu: Shuxue, si nianji, xia ce* [Compulsory education textbook: Mathematics, grade 4, second semester] (2nd ed.). People's Education Press.

Primary Mathematics Textbook Editorial Committee, Curriculum and Teaching Materials Research Institute, People's Education Press. (2022b). *Yiwu jiaoyu jiaokeshu: Shuxue, wu nianji, shang ce* [Compulsory education textbook: Mathematics, grade 5, first semester] (2nd ed.). People's Education Press.

Primary Mathematics Textbook Editorial Committee, Curriculum and Teaching Materials Research Institute, People's Education Press. (2022c). *Yiwu jiaoyu jiaokeshu: Shuxue, wu nianji, xia ce* [Compulsory education textbook: Mathematics, grade 5, second semester] (2nd ed.). People's Education Press.

Rahmawati, D., & Anwar, R. B. (2020). Translation of mathematical representation: Characteristics of verbal representation unpacking. *Journal of Education and Learning, 14*(2), 162-167. https://doi.org/10.11591/edulearn.v14i2.9538

Schorcht, S., Buchholtz, N., & Baumanns, L. (2024). Prompt the problem – Investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Frontiers in Education, 9*, Article 1386075. https://doi.org/10.3389/feduc.2024.1386075

Shoustikova, T., Zaripova, Z. F., Ibragimov, G., Zaitseva, N. A., Pashanova, O. V., & Lobuteva, A. V. (2026). Exploring the scope of artificial intelligence literacy in education: A bibliometric approach. *Contemporary Educational Technology, 18*(1), Article ep628. https://doi.org/10.30935/cedtech/17892

Singh, A., Ehtesham, A., Kumar, S., & Khoei, T. T. (2025). *Agentic retrieval-augmented generation: A survey on agentic RAG*. arXiv. https://arxiv.org/abs/2501.09136v1

Sujatha, S., & Vinayakan, K. (2023). Engaging students with mathematics: A review of motivation and engagement strategies. *International Journal of Interdisciplinary Research in Arts and Humanities, 8*(2), 55-60. http://doi.org/10.5281/zenodo.16413870

Tay, Y. K., & Toh, T. L. (2023). A model for scaffolding mathematical problem-solving: From theory to practice. *Contemporary Mathematics and Science Education, 4*(2), Article ep23019. https://doi.org/10.30935/conmaths/13308

Tong, D. H., Nguyen, H. T., Nguyen, T.-T., Le, T. T. T., Vo, X. M., Triet, L. V. M., & Dung, T. M. (2026). Factors shaping teachers' beliefs about ICT use in mathematics teaching. *Contemporary Educational Technology, 18*(1), Article ep629. https://doi.org/10.30935/cedtech/17893

Tuohilampi, L., & Jiang, H. (2024). Introducing the ProInterest model: Designing tasks for sustainable positive mathematics experiences. *LUMAT: International Journal on Math, Science and Technology Education, 12*(3), Article 9. https://doi.org/10.31129/LUMAT.12.3.2373

Ussenova, G., Issayeva, G., Urazaliyeva, R., Karimova, G., & Mukhanova, M. (2025). Examining cognitive independence in the context of digital learning: A moderation analysis of student motivation, self-regulation skills, and cognitive engagement. *Contemporary Educational Technology, 17*(4), Article ep618. https://doi.org/10.30935/cedtech/17570

Valquaresma, A. (2024). Creativity, agency and meaning-making: Unfolding developmental possibilities in twenty-first-century learning environments. In A. Valquaresma, L. Dantas de Paula, & T. K. Rodney (Eds.), *Creativity and learning: Navigating transformative perspectives for complex and contemporary environments* (pp. 105-126). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-73393-2_5

Wu, Y., Jiang, Y., Chen, Y., & Zhang, W. (2024). Da yuyan moxing zhichi de duozhinengti: Jishu lujing, jiaoyu yingyong yu weilai zhanwang [Multi-agent systems supported by large language models: Technical pathways, educational applications, and future prospects]. *Open Education Research, 30*(5), 77-89. https://doi.org/10.13966/j.cnki.kfjyyj.2024.05.007

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). *ReAct: Synergizing reasoning and acting in language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2210.03629

Zhang, Z., Yao, Y., Zhang, A., Tang, X., Ma, X., He, Z., Wang, Y., Gerstein, M., Wang, R., Liu, G., & Zhao, H. (2025). Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. *ACM Computing Surveys, 57*(8), Article 206. https://doi.org/10.1145/3719341

APPENDIX A

Pre-Test, Intervention, and Post-Test Tasks

The pre-test included two short, structured word problems designed to assess students' baseline word-problem solving performance. The intervention task served as the focal problem during tutoring. The post-test included three related transfer tasks designed to examine whether students could apply similar reasoning to closely related problem situations.

Pre-test item 1

There are 18 bicycles and cars in the parking lot, with a total of 56 wheels. How many bicycles and how many cars are there?

Pre-test item 2

There are 15 cranes and giraffes in a zoo, with a total of 40 legs. How many cranes and how many giraffes are there?

Intervention task

A cage contains some chickens and rabbits. Counting from the top, there are 45 heads. Counting from the bottom, there are 100 legs. How many chickens and how many rabbits are there in the cage?

Post-test item 1

There are 28 chickens and rabbits in total, with 86 legs altogether. How many chickens and how many rabbits are there?

Post-test item 2

There are 22 turtles and cranes in total, with 68 legs altogether. How many turtles and how many cranes are there?

Post-test item 3

There are 20 bicycles and tricycles in a parking lot, with a total of 48 wheels. How many bicycles and how many tricycles are there?

Detailed Workflow of the Quasi-Experimental Study

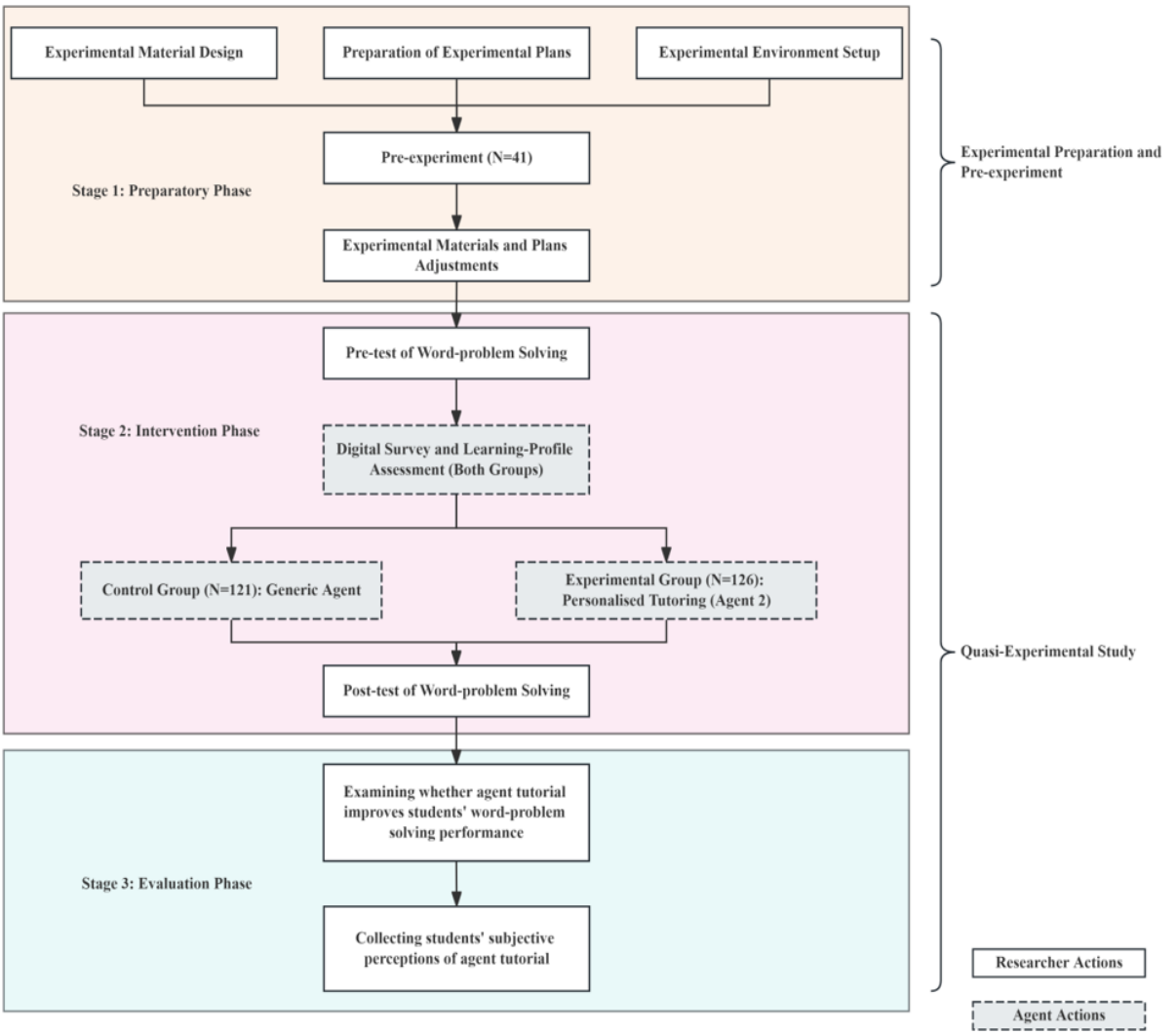


Figure A1. Detailed workflow of the quasi-experimental study (both groups completed the learning-profile assessment; the control group received general tutoring, and the experimental group received personalised tutoring) (created by the authors)