MODESTUM EURASIA Journal of Mathematics, Science and Technology Education, 2026, 22(4), em2816
ISSN:1305-8223 (online)
OPEN ACCESS Research Paper https://doi.org/10.29333/ejmste/18263

Modelling with AI: How complexity and experience shape ChatGPT use by pre-service teachers

César Gallart [1], Irene Ferrando [1], Carlos Segura [1], Lluís Albarracín [2*]

[1] Departmento de Didáctica de las Matemáticas, Universitat de València, València, SPAIN
[2] Department de Didàctica de la Matemàtica i les Ciències Experimentals, Universitat Autònoma de Barcelona, Barcelona, SPAIN

Received 23 October 2025 ▪ Accepted 14 March 2026

Abstract

Despite their potential to foster critical thinking, modelling tasks remain underrepresented in mathematics classrooms. Fermi problems (FPs), as open estimation tasks, are well-suited for introducing modelling in primary education. Given the growing presence of artificial intelligence (AI) in education, it is essential to understand how pre-service teachers (PSTs) engage with tools like ChatGPT in modelling contexts. This study analyses the use of ChatGPT by 133 PSTss solving FPs and examines how this use is shaped by problem complexity and prior experience. Through qualitative and quantitative analysis, three distinct profiles of AI use emerged—expert, assistant, and support—reflecting varying degrees of autonomy and delegation. Results show greater delegation to AI in more complex problems, while prior experience with outdoor problem-solving or ChatGPT fosters more autonomous engagement. These findings provide insights for integrating AI into mathematics education to support reflective, independent, and critical modelling practices.

Keywords: modelling, Fermi problems, artificial intelligence, prompt engineering, problem-solving setting

INTRODUCTION

Research on mathematical modelling has reached a high level of maturity, and its inclusion in international curricula is widely promoted (Schukajlow et al., 2018). The literature has also emphasized that modelling is pursued with various educational objectives and depends largely on the task characteristics, which in turn shape what students can learn from modelling activities (Krawitz et al., 2025). However, modelling practices remain uncommon in classrooms, particularly at the primary education level (English, 2023). As a result, students have limited opportunities to use mathematics as a tool to interpret real-world phenomena or to engage in critical and reflective thinking about complex situations (Borromeo-Ferri, 2021). This persistent gap between curricular intentions and classroom practice highlights the need for instructional approaches and teacher training models that make modelling accessible and meaningful in school settings.

In this context, Fermi problems (FPs) constitute a well-established pedagogical resource within mathematical modelling research (Ärlebäck, 2009). Their open-ended nature, reliance on estimation, and requirement to make assumptions and simplifications place them firmly within the modelling paradigm, while their apparent simplicity lowers entry barriers for novice modelers (Peter-Koop, 2009). Unlike more formal modelling tasks that may require advanced mathematical tools or extensive domain knowledge, FPs foreground key modelling processes—such as structuring a situation, selecting relevant variables, and justifying assumptions—making them particularly suitable for teacher education contexts. Moreover, previous research has shown that sequences of FPs with varying contextual and structural complexity can foster flexibility and adaptative strategy in problem-solving, supporting the progressive development of modelling competence (Segura & Ferrando, 2023). Nevertheless, studies also indicate that pre-service teachers (PSTs) frequently struggle with these tasks (Segura et al., 2025),

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

✉ cesar.gallart@uv.es ✉ irene.ferrando@uv.es ✉ carlos.segura@uv.es ✉ lluis.albarracin@uab.cat **(*Correspondence)**

Contribution to the literature

- This study contributes to the growing body of research on the integration of AI in mathematics education by examining how PSTs engage with ChatGPT when solving modelling problems.
- It advances current understanding by characterizing distinct AI use profiles based on prompting strategies and solver autonomy, and by analyzing how this use is shaped by problem complexity and prior experience.
- The findings enrich existing frameworks on AI-supported problem-solving, offering insights for fostering more reflective and pedagogically grounded uses of AI in mathematics learning contexts.

underscoring the need for targeted support in modelling instruction (Geiger et al., 2022).

To support such training, previous studies have explored different problem-solving settings in which FPs are enacted. These include outdoor settings, where the problem context is directly experienced on site, and indoor settings, where the context is mediated through images or descriptions (Buchholtz, 2021; Jablonski, 2023; Segura et al., 2023). More recently, these settings have been complemented by educational technologies, adding new layers of mediation to the modelling process (Jablonski et al., 2023; Quarder et al., 2025).

Among these technologies, artificial intelligence (AI) has rapidly entered mathematics educational research. Systematic reviews highlight both its opportunities and risks for teaching and learning (Almarashdi et al., 2024; Pepin et al., 2025). From a modelling perspective, AI tools can act as external cognitive resources, offering estimates, assumptions or solution structures. However, these systems rely heavily on user-generated prompts, making the nature of human-AI interaction a central factor in their educational effectiveness (Noster et al., 2024; Schorcht et al., 2024). Emerging evidence suggests that prompting strategies influence the quality of AI-generated responses and the extent to which users rely on them (Spreitzer et al., 2024).

Despite this growing body of research, important gaps remain. Much of the existing work has focused either on the quality of AI-generated solutions or on prompting as a set of local interaction moves, without situating these interactions within the broader modelling process (Fock & Siller, 2025). As a result, we still know relatively little about how PSTs integrate AI across the phases of modelling when tackling open-ended tasks such as FPs, or how this integration may vary depending on task complexity and prior experience in different problem-solving settings. Moreover, although FPs are widely used in modelling research, their specific affordances for examining AI-supported modelling have not yet been explored in depth.

From a teacher education perspective, addressing these gaps is crucial. Recent studies with PSTs also indicate that sustained engagement with AI can shape perceptions, intentions, and classroom-oriented practices (Zhuang & Zhang, 2025). In modelling practices, understanding how and under what conditions PSTs delegate modelling processes to AI is key to fostering critical, autonomous, and reflective engagement with this tool in future classrooms (Celik, 2023; Walter, 2024).

Against this background, the present study adopts a quasi-experimental design to investigate how PSTs interact with ChatGPT when solving a sequence of FPs. Specifically, it examines how different profiles of AI use emerge and how they are shaped by task complexity and prior experience acquired in distinct problem-solving settings. By integrating perspectives from mathematical modelling, problem-solving research, and prompt engineering, this study aims to contribute to a more theoretically coherent understanding of AI-supported modelling in teacher education and to inform the design of instructional practices that promote critical engagement with AI tools.

THEORETICAL FRAMEWORK

Modelling Problems in Mathematics Education

There has been consensus for decades on the central role of problem-solving in the development of mathematical competence, as it provides students with opportunities to engage in mathematical reasoning and to understand the relationships between mathematical concepts and procedures (DiNapoli & Miller, 2022; Stanic & Kilpatrick, 1989). Problems situated in real contexts that require the mathematical analysis of complex real-world phenomena enable students to make sense of mathematical content and connect it with other disciplines (Gravemeijer & Doorman, 1999). Moreover, solving such problems promotes autonomy and critical thinking (Clarke & Roche, 2018; English & Gainsburg, 2015). For this reason, many curricula emphasize the importance of real-world problem-solving (Cevikbas et al., 2022; Schukajlow et al., 2018).

Modelling problems are characterized as authentic—relate to reality—, complex—involving a multi-step solution process— and open—allowing for multiple solutions depending on the assumptions made about the situation set (Maass, 2006). Their solution involves the development of a mathematical model, understood as a system of interrelated concepts and procedures that represent and structure the real-world situation using variables and their relationships (Lesh & Doerr, 2003). In this sense, the complexity of modelling problems does not only stem from mathematical demands, but also from the need to coordinate assumptions, representations, and relationships between variables (Maass, 2010).

The modelling process unfolds through what is known as the modelling cycle (Blum & Leiss, 2007; Czocher, 2016), which involves a double process between the real world and the world of mathematics. Following this cycle, students try to solve the problem by progressing through different stages and may revisit earlier ones to refine their understanding of the situation (Blum & Borromeo-Ferri, 2009). This cyclic perspective provides a useful analytical framework for examining how learners approach modelling tasks and how external tools may intervene at different phases of the process.

Modelling problems are gaining increasing relevance in educational curricula due to their role in fostering mathematical competence and their potential connections with other STEM disciplines (Maass et al., 2019). Among them, FPs stand out as particularly suitable for introducing both students and PSTs to modelling practices, as they combine accessible contexts with varying degrees of complexity (Ärlebäck, 2009), as will be discussed below.

Fermi Problems as Mathematical Modelling Activities for Primary Education

Ärlebäck (2009) defines FPs as "open, non-standard problems requiring the solvers to make assumptions about the problem situation and estimate relevant quantities before engaging in, often, simple calculations" (p. 332). These problems present a real-world situation with little or no specific data, requiring solvers to rely on reasoned estimations (Efthimiou & Llewellyn, 2007). Research has shown that FPs are suitable for introducing modelling practices in primary education (Albarracín & Ärlebäck, 2019; Ärlebäck, 2009), as they foreground key modelling actions such as assumption-making, simplification, and estimation within accessible contexts.

Solving a FP involves making reasonable assumptions, breaking the task into manageable subproblems, and flexibly applying mathematical knowledge (Robinson, 2008). These problems are inherently open-ended, allowing for multiple valid solutions based on the assumptions and initial decisions made by the solver (Segura et al., 2023). The solution process can be described in terms of a modelling cycle, involving the following phases (Albarracín & Ärlebäck, 2019; Borromeo-Ferri, 2006):

- **Understanding the problem:** mentally constructing the situation, simplifying it and identifying the key elements to be considered.
- **Establishing the model:** mathematising the relevant elements by establishing their relationships among them using different strategies (Albarracín & Gorgorió, 2014). Common strategies include using *density* (the number of elements per unit), *unit iteration* (estimating the area occupied by a single element) and *grid distribution* (estimating the number of elements in a row and a column).
- **Working mathematically:** applying the selected model to generate a numerical result.
- **Interpreting and validating:** relating the mathematical result to the real-world context and validating it. If the solution is deemed unsuitable, the model is revised and the process repeats; otherwise, the result is communicated.

For modelling to become a meaningful classroom practice, PSTs need opportunities to experience problem-solving as learners before they can teach it effectively (Thompson, 1985). This calls for equipping them with tools and strategies to approach modelling tasks, particularly those that require managing open-endedness and complexity (Geiger et al., 2022). Within this training context, FPs have been recognized as a powerful entry point to modelling. However, recent studies indicate that many future primary education teachers still struggle to solve FPs effectively (Segura & Ferrando, 2023; Segura et al., 2025), which helps explain why modelling remains difficult to implement in classroom practice.

Complexity and Solving Settings of Modelling Tasks

Complexity in modelling problems can be defined in various ways (Krawitz et al., 2024). It is generally determined by the complexity of the proposed situation and the demands it imposes on the solver (Knabbe et al., 2025; Maass, 2010). According to Maass (2010), greater complexity is associated with an increased number of steps required for problem solution, involving more assumptions, variables, decisions and relationships that must be considered when building or adapting an appropriate model.

Concerning problem-solving settings, some studies have explored how modelling performance differs between outdoor settings, where students experience the problem context directly, and indoor settings, where the context is evoked through images, visual representations and descriptions (e.g., Buchholtz, 2021; Hartmann & Schukajlow, 2021; Jablonski, 2023). In Segura et al. (2023), PSTs developed initial solution plans to FPs in the classroom and later revised their strategies when solving the same task *in situ*, adapting to real-world constraints and refining their assumptions accordingly. These transitions between settings illustrate how contextual engagement can shape the interpretation of the task and the handling of complexity at different stages of the modelling process. These findings suggest that engaging with problems in an outdoor setting may offer valuable scaffolding for developing modelling competence.

Recent studies have examined how modelling processes are shaped not only by where the task is solved (indoor vs. outdoor), but also by the availability of specific digital tools—including simulations, interactive tools and 3D printing—which can expand the range of representations available during problem-solving (e.g., Jablonski et al., 2023; Quarder et al., 2025). From this perspective, the use of AI can be understood as an additional resolution condition that modifies the problem-solving setting, mediating access to information, representations, and possible solution paths. However, despite the growing interest in the educational use of AI, studies investigating its role as a supportive setting for modelling problem-solving remain limited (Noster et al., 2024; Spreitzer et al., 2024).

Use of Generative AI in Mathematics Education and Problem-Solving

In recent years, generative AI tools based on large language models (LLMs), such as ChatGPT, have emerged as interactive systems capable of responding to user prompts through natural language dialogue. Their rapid incorporation into educational contexts has generated both opportunities and challenges (Yu, 2023; Zhang & Aslam, 2021), particularly regarding how they may transform teaching and learning processes, at all educational levels (Lin et al., 2023; Lo, 2023; Walter, 2024).

Almarashdi et al. (2024) and Pepin et al. (2025) provide broad reviews of reported applications of ChatGPT in mathematics education, including lesson planning, task design and assessment. In particular, Almarashdi et al. (2024) note that some studies report increased student engagement and learning motivation when ChatGPT is used in educational settings. In line with this, Guardia-Paniura et al. (2026), in a systematic review on AI-generated feedback in higher education, report that its perceived effectiveness depends on how students interpret and use it, and they emphasize the need for context-sensitive use and human oversight.

Within problem-solving contexts, however, an open question remains as to whether—and under what conditions—such tools can meaningfully support the development of mathematical competencies rather than merely providing solutions (Bani-Hamad & Al-Kalbani, 2024; Schorcht et al., 2024). This aligns with broader work on the *paradox of accessibility*, which suggests that increased access to digital and AI-based tools does not automatically translate into stronger mathematical competence (Lumandas & Taja-on, 2026). Although LLMs have demonstrated the capacity to solve a broad range of mathematical problems, their performance appears to be influenced by both the complexity of the task and the mathematical domain involved (Almarashdi et al., 2024; Fock & Siller, 2025; Schorcht et al., 2024).

In the case of mathematical modelling problems, research suggests that the effectiveness of AI tools such as ChatGPT is influenced more by the contextual complexity of the problem than by its degree of openness. Spreitzer et al. (2024) found that while ChatGPT performs reasonably well on basic modelling tasks, its reliability decreases as the number of variables, assumptions, and contextual relationships increases. Moreover, they observed that generic prompts (e.g., "solve the following task") often lead to responses that lack justification, validation, or explicit modelling assumptions. Similar limitations have been reported in studies on FPs (López-Simó & Rezende, 2024), reinforcing the need to examine how users interact with AI during the modelling process.

Taken together, these findings suggest that while ChatGPT can function as a supportive condition within modelling problem-solving settings, its educational value depends strongly on how it is used. In particular, the quality of prompting strategies and the degree of responsibility assumed by the solver appear to play a key role in shaping the effectiveness of AI support in modelling tasks.

Use of Prompts in Problem-Solving

The conditioning of LLMs through different types of user instructions—collectively referred to as *prompting*—is studied within the field of prompt engineering, which examines how prompts can be designed to increase the likelihood of obtaining accurate responses (Schorcht et al., 2024; Walter, 2024).

Prompting techniques (described in Schorcht et al., 2024, and Walter, 2024) are commonly classified according to the number of instructions provided to the AI. In *zero-shot prompting*, the task is presented directly without additional guidance or details, relying on a single, concise instruction. In contrast, the *few-shot prompting* involves supplying the LLM with prior examples or intermediate prompts that guide the reasoning process and may improve the quality of the response.

Beyond this distinction, AI performance can be further enhanced by combining these approaches with prompts that explicitly structure the reasoning process. For instance, *chain-of-thought prompting* encourages the AI to articulate intermediate steps, making its reasoning more transparent. Another approach, often referred to as *output customization*, consists of constraining the response or assigning the AI a specific role—for example, asking it to act as an expert or to follow predefined rules or conventions. Such strategies are particularly relevant in open-ended modelling tasks, where assumptions, intermediate decisions, and validation play a central role.

In the context of mathematical problem-solving, Noster et al. (2024) analyzed PSTs´ interactions with ChatGPT and found that prompting strategies were strongly task dependent. Most participants relied on *zero-shot prompting*, while more elaborate techniques—such as *few-shot prompting*—emerged only in response to a specific task demands. These findings raise important questions about how task characteristics—such as complexity—and users' prior experience in different problem-solving settings shape the selection and use of prompting strategies when solving modelling tasks, and in particular FPs.

Research Goals

As AI-based tools become increasingly integrated into mathematics education, there is a growing need to understand how PSTs engage with them when solving open-ended modelling tasks. Building on research on mathematical modelling, FPs and problem-solving settings, this study aims to examine how interactions with ChatGPT are embedded within the modelling process, rather than treating AI as a mere source of answers. Specifically, we focus on how PSTs use ChatGPT when solving FPs of varying complexity, and on how this use is shaped by task complexity and by prior experience in different problem-solving settings. In this sense, task complexity may increase the perceived need for AI support, whereas prior experience with modelling in diverse settings may enable PSTs to retain greater responsibility for the problem-solving process. Based on this aim, we pose the following research questions (RQs):

RQ1. What uses do PSTs make of ChatGPT when solving FPs?

RQ2. How does the complexity of the FPs affect the use of ChatGPT in the solution?

RQ3. How does the type of prior experience—solving simple FPs in-class without assistance (RF group), outdoors at the problem context location (out group), or in-class with assistance from ChatGPT (AI group)—influence the use of ChatGPT when tackling more complex FPs?

METHODOLOGY

To address the RQs, we have conducted a quasi-experimental study with a convenient sample of future primary education teachers.

Participants

The study involved a sample of 133 PSTs enrolled in the third year of the primary education degree program at the faculty of teacher training, University of València (Spain). The average age of the participants was 22.2 years, and 67% were women, making this a representative sample of pre-services at this stage of the program. The participants were completing their academic year and had received prior coursework in arithmetic, geometry, measurement, statistics, and probability. Before the study, it was confirmed that none of the participants had any previous experience with FPs. All participants were informed about the aims of the research and took part voluntarily. Before the study, they created accounts on ChatGPT (free-access version 3.5) and were instructed to bring a personal laptop with Internet access on the day of the activity.

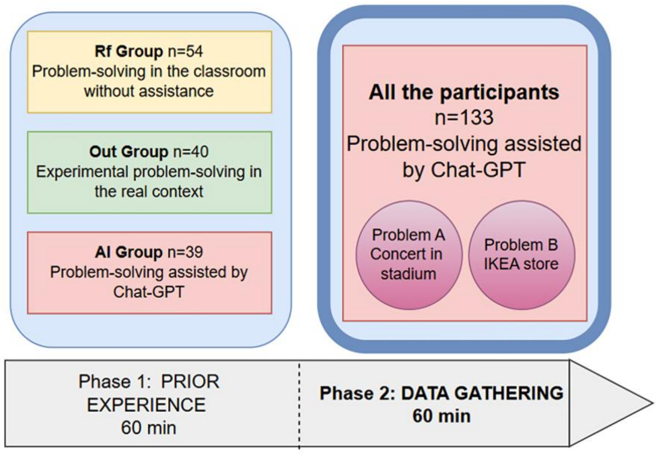

Rf Group n=54
Problem-solving in the classroom without assistance
Out Group n=40
Experimental problem-solving in the real context
AI Group n=39
Problem-solving assisted by Chat-GPT
All the participants
n=133
Problem-solving assisted by Chat-GPT
Problem A
Concert in stadium
Problem B
IKEA store
Phase 1: PRIOR EXPERIENCE
60 min
Phase 2: DATA GATHERING
60 min

Figure 1. Outline of the design of the experience (Source: Authors' own elaboration)

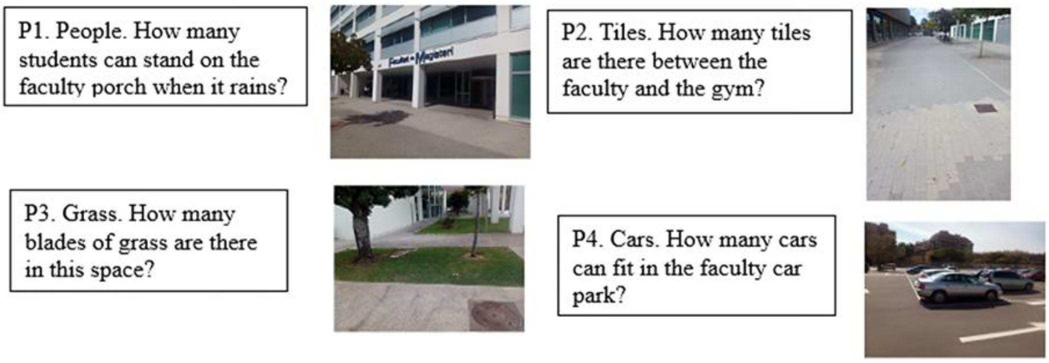
Procedure and Data Gathering

The experience, supervised by two researchers, took place over a two-hour session in two distinct phases. **Figure 1** outlines the design of the experience and the structure used for data collection.

Participants were randomly assigned to three groups—reference (RF), outdoor (out), and AI-assisted (AI)—before the start of phase 1. This first phase served as a preparatory intervention offering different types of prior experience. However, only the data collected in phase 2 were used to address the study's RQs (RQ1-RQ3).

In phase 1, all groups worked on the same sequence of four FPs, but under different conditions. The problems were situated in a familiar context (the area surrounding the Faculty of Education) and all required estimating the number of elements within a bounded space. These FPs sequence had been used in prior studies (Ferrando et al., 2021; Segura & Ferrando, 2023) and can be seen in **Figure 2**.

- In group Rf (n = 54), participants solved the problems independently in the classroom without assistance.
- In group out (n = 40), participants worked directly in the real-world contexts described in the problems. They were free to move around the Faculty of Education, take measures, and engage with the environment.



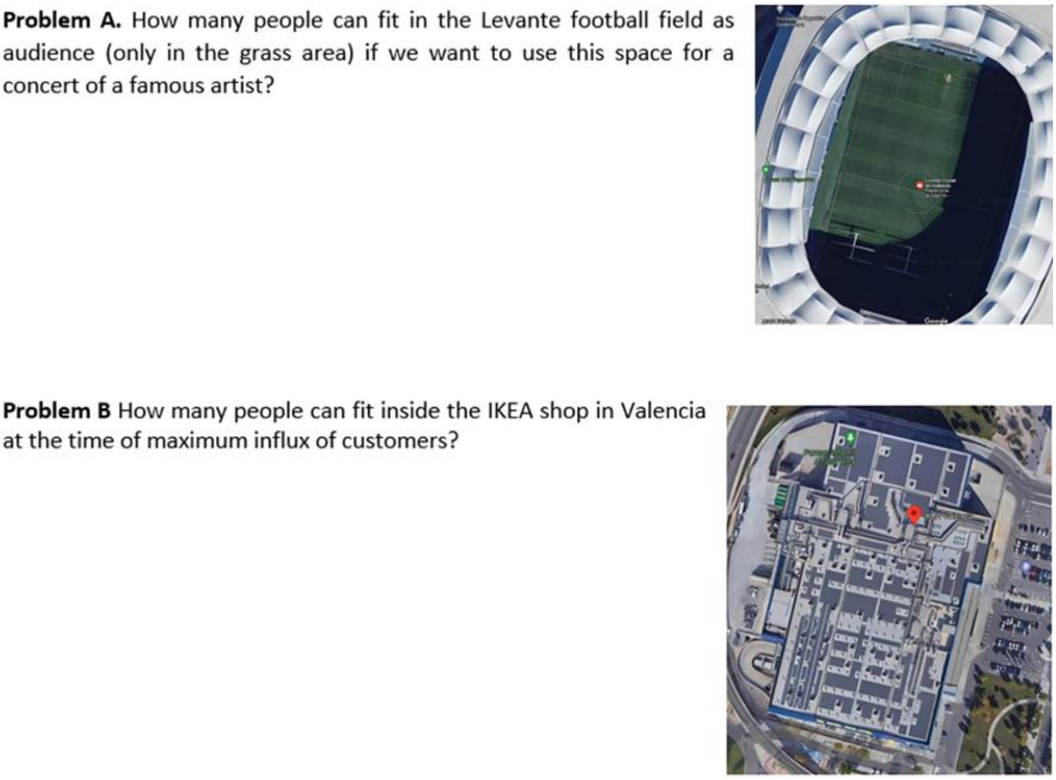
P1. People. How many students can stand on the faculty porch when it rains?

P2. Tiles. How many tiles are there between the faculty and the gym?

P3. Grass. How many blades of grass are there in this space?

P4. Cars. How many cars can fit in the faculty car park?

Figure 2. Problems included in the first phase of the study (prior experience) (Source: Authors' own elaboration)



Problem A. How many people can fit in the Levante football field as audience (only in the grass area) if we want to use this space for a concert of a famous artist?

Problem B How many people can fit inside the IKEA shop in Valencia at the time of maximum influx of customers?

Figure 3. Problems included in the second phase of the study (data gathering) (Source: Authors' own elaboration, including Map data ©2026 Google)

- In group AI (n = 39), participants solved the problems in the classroom using ChatGPT on their personal laptops. They were not permitted to use other search engines or AI tools.

In phase 2, all the participants solved the same two problems in identical conditions: in the classroom, with access to ChatGPT on their personal devices. In this case, participants had to solve two FPs situated in a familiar context, though not directly accessible; the statements of both problems appear in **Figure 3**.

In the first problem of this second phase, problem A (P_A), participants were asked to estimate how many people could fit on the field of the Levante football stadium (located in the same city as their faculty) for a concert. In the second problem, problem B (P_B), they had to estimate how many people could be inside the IKEA store (a well-known shop in a nearby town) during peak hours.

Although both tasks can be classified as Fermi-type modelling problems, P_B was designed to be more complex than P_A. Following Maass (2010), task complexity in modelling is related to the number of relevant variables, the degree of interdependence between them, and the number and nature of assumptions required to construct a model. P_A mainly involves a limited set of predominantly spatial variables (e.g., usable area and density assumptions) within a static context, similar to problem P1 proposed in phase 1, where variables are combined through direct and largely independent relationships.

In contrast, P_B requires coordinating a broader range of spatial, temporal, and behavioral variables, such as accessible area for customers, entry and exit rates,

Table 1. Comparison of modelling task complexity

Aspect	Problem A: Football field concert	Problem B: IKEA at peak time
Context structure	Highly structured, static	Less structured, dynamic
Main goal	Estimate maximum capacity	Estimate occupancy at peak time
Number of relevant variables	Limited	Higher
Nature of variables	Spatial	Spatial, temporal, behavioral
Relations between variables	Simple and independent relationships	Dynamic relationships with mutual dependencies
Required assumptions	Few and explicit	Numerous and implicit
Modelling decisions	Relatively straightforward and structurally familiar	Open and ambiguous in terms of model structure

Table 2. Categorization of prompts based on the phases of FP-solving process

Phase of the FP-solving process	Description	Example
Understanding the problem (UP)	Prompts to gain a more in-depth understanding of the problem.	"How is the maximum capacity of people in a place determined?", in P_A, or "What information needs to be considered to calculate the maximum capacity of a shop?", in P_B.
Establishing the model: dimensions of the space (EM1)	Prompts about the dimensions of the space to be occupied	"What is the area of a football pitch?" or "What are the measurements of the IKEA in Alfafar?"
Establishing the model: density or space occupied by a single person (EM2)	Prompts related to the density or the area occupied by one person.	"How much does a person occupy?" or "How many people fit in one meter squared?"
Establishing inclusion of realistic assumptions (EM3)	Prompts focused on realistic assumptions.	"How much does the stage of a famous artist occupy?" or "How many meters squared are occupied by shelves and furniture in IKEA?"
Working mathematically (WM)	Prompts that require mathematical calculations, using ChatGPT as if it were a calculator.	"Multiply 4 by 7,140" or "How much is 7,140 minus 1,500?"
Interpreting and validating (IV)	Prompts that seek to verify the validity or confirm a result or an assumption.	"Do you think that 420 people can fit in a football pitch if there is a concert?" or "Is five meters squared per customer too much?"

average length of stay, and customer distribution across store areas. This increased situational complexity (Knabbe et al., 2025) is also reflected in the formulation of the question itself: whereas P_A focuses on estimating the maximum possible occupancy, P_B asks for the number of customers present at the moment of maximum influx, which does not necessarily coincide with the maximum capacity and requires modelling customer flow over time. In this case, variables form a dynamic system with multiple and interdependent relationships, so that changes in one variable may affect several others simultaneously (e.g., increasing the rate of customer entry raises occupancy, which may lengthen the average time spent in the store and, in turn, reduce the rate of exit). As a result, P_B relies on a larger number of assumptions—many of them implicit and mutually dependent—related to customer behavior and flow dynamics, whereas P_A can be solved using a small number of explicit and largely independent assumptions. These features justify the higher modelling complexity attributed to P_B. **Table 1** summarizes the main differences between the two tasks in terms of modelling complexity.

In the written questionnaire provided to the participants, we included the following advertisement:

> "Below we propose two problems; to solve them, you can use the ChatGPT tool, but please note that you will have to submit the file with all your interactions with the AI. In your written answer, explain, step by step, how to solve each problem, indicating what data and procedures you have used to reach the solution."

Each questionnaire contained the problem statements and a blank space for the written responses. Researchers collected both the written solutions and the files containing each participant's interactions with ChatGPT.

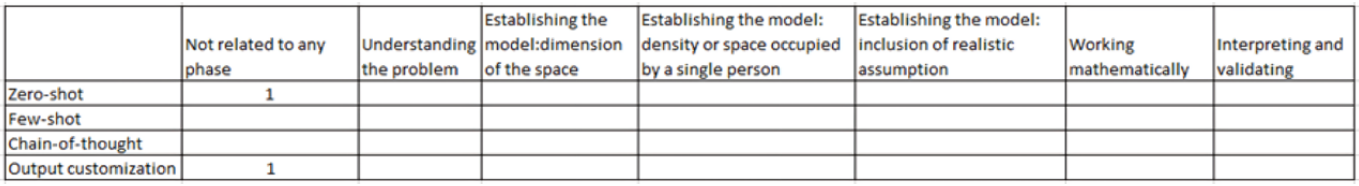
Data Analysis

To address the research objectives, we analyzed the data collected during the second phase of the study. Specifically, since we are interested in the use of AI, we focused on the prompts that participants gave to the ChatGPT tool to solve problems P_A and P_B. This analysis was complemented by examining each participant's written solution. The categorization of prompts follows a double deductive coding approach, grounded in a predefined system of categories established in the specialized literature and outlined in the theoretical framework. This system considers two aspects: the relationship of the prompt with the modelling process and the type of prompt used.

For this aim, first, we analyzed, for each prompt (and in the light of the written solution), whether the prompt can be related to a phase of the FP-solving process, as

Table 3. Description of the categories of prompting techniques

Prompting technique	Description	Example
Zero-shot	The problem is entered directly into the chat by copying its formulation without any prior questions, examples, or additional contextual information.	"How many people can fit in the Levante football stadium as an audience if we want to hold a concert there?" in P_A.
Few-shot	Users pose questions related to specific aspects of the problem before requesting a final solution. These preliminary prompts are aligned with phases of the Fermi problem-solving process and are intended to guide and refine the AI-generated response.	"How many people can fit in 7,140 meters squared?" in P_A, or "If you have 8,000 meters squared, how many people can fit?" in P_B. Also, "… you need to bear in mind that there will be a stage, an area reserved for the technical team and sound system" in P_A or "Calculate the problem now assuming that the size is 30,000 meters squared, but the density of people is 1 per meter squared", in P_B.
Chain-of-thought	After receiving a response to a zero-shot or few-shot prompt, ChatGPT is given a prompt to detail or explain its solution.	"How did you calculate it?"
Output customization	After receiving a response to a zero-shot or few-shot prompt, users request a specific output format from ChatGPT.	"Solve it with mathematical operations", "I need a quantity" or "Calculate it, approximately, with figures".



	Not related to any phase	Understanding the problem	Establishing the model:dimension of the space	Establishing the model: density or space occupied by a single person	Establishing the model: inclusion of realistic assumption	Working mathematically	Interpreting and validating
Zero-shot	1						
Few-shot							
Chain-of-thought							
Output customization	1						

Figure 4. Grid used for the analysis of participants' prompts (filled for participant A017) (Source: Authors' own elaboration)

outlined in the theoretical framework. In **Table 2**, we provide, for each phase of the FP-solving process, a brief description of the characteristics of the prompts and an example.

At this point, we also found some prompts that cannot be directly related to any of the phases of the FP-solving process. For instance, we categorized that prompt as "not related to the modelling process" (NotM). Once these categories were established, productions (266 interaction documents with ChatGPT and 266 written solutions) were divided for categorization into two research pairs, who analyzed and agreed on the categorization. For each prompt, we identified whether it corresponds to one of the 6 phases or not. First, to ensure the reliability of the coding (interrater agreement), each pair of coders analyzed 20% of the documents independently, obtaining for each prompt a Cohen's kappa of $0.70 \leq k \leq 1$ in P_A, and of $0.61 \leq k \leq 1$ in P_B, indicating acceptable to very good agreement. Once a good level of agreement between the two pairs of coders was assured, the remaining documents were divided.

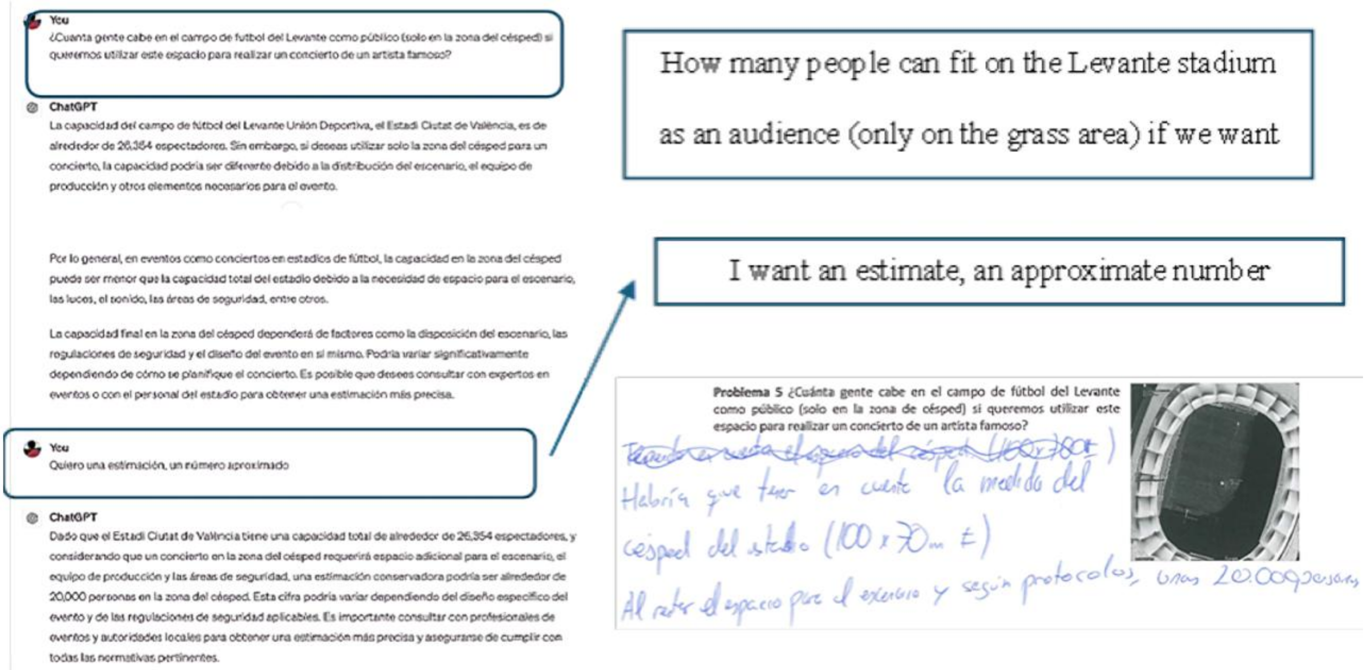
Once we completed this first analysis, we started the categorization of prompts grounded in a predefined system of categories established in specialized literature and outlined in the theoretical framework. **Table 3** presents the complete categorization system used in this study.

Once we agreed on the description of each category, two of the researchers coded the productions of the 133 participants to solve P_A and P_B (266 interaction documents with ChatGPT and 266 written solutions). The presence or absence of each of the four prompts in each participant's interactions was coded. First, to ensure coding reliability (interrater agreement), both researchers analyzed 20% of the documents independently, obtaining for each prompt a Cohen's kappa of $0.65 \leq k \leq 0.84$ in P_A, and of $0.75 \leq k \leq 1$ in P_B, indicating acceptable to very good agreement. Once a good level of agreement between coders was assured, both researchers divided the remaining documents, meeting to agree on the category in case of doubts.

To systematize the analysis, we use, for each participant and each problem, the grid shown in **Figure 4**. In that grid, we mark a "1" in each cell when the participant uses and prompts corresponding to the FP-solving process (see **Table 2**) or to a prompting technique category (see **Table 3**). If no such prompt is used, the cell is left blank.

For example, in **Figure 5**, we show the interaction with ChatGPT and the written solution of participant A017 when solving problem P_A.

Based on the qualitative analysis of participants' interactions with ChatGPT and their written solutions, three distinct profiles of AI use were identified (which responds to **RQ1**). These emergent profiles were then coded as nominal variables to analyze through Chi-square tests, their relationship with problem complexity (**RQ2**) and prior experience (**RQ3**).



You
¿Cuanta gente cabe en el campo de futbol del Levante como público (solo en la zona del césped) si queremos utilizar este espacio para realizar un concierto de un artista famoso?

ChatGPT
La capacidad del campo de fútbol del Levante Unión Deportiva, el Estadi Ciutat de València, es de alrededor de 26,354 espectadores. Sin embargo, si deseas utilizar solo la zona del césped para un concierto, la capacidad podría ser diferente debido a la distribución del escenario, el equipo de producción y otros elementos necesarios para el evento.

Por lo general, en eventos como conciertos en estadios de fútbol, la capacidad en la zona del césped puede ser menor que la capacidad total del estadio debido a la necesidad de espacio para el escenario, las luces, el sonido, las áreas de seguridad, entre otros.

La capacidad final en la zona del césped dependerá de factores como la disposición del escenario, las regulaciones de seguridad y el diseño del evento en sí mismo. Podría variar significativamente dependiendo de cómo se planifique el concierto. Es posible que desees consultar con expertos en eventos o con el personal del estadio para obtener una estimación más precisa.

You
Quiero una estimación, un número aproximado

ChatGPT
Dado que el Estadi Ciutat de València tiene una capacidad total de alrededor de 26,354 espectadores, y considerando que un concierto en la zona del césped requerirá espacio adicional para el escenario, el equipo de producción y las áreas de seguridad, una estimación conservadora podría ser alrededor de 20,000 personas en la zona del césped. Esta cifra podría variar dependiendo del diseño específico del evento y de las regulaciones de seguridad aplicables. Es importante consultar con profesionales de eventos y autoridades locales para obtener una estimación más precisa y asegurarse de cumplir con todas las normativas pertinentes.

How many people can fit on the Levante stadium as an audience (only on the grass area) if we want

I want an estimate, an approximate number

Problema 5 ¿Cuánta gente cabe en el campo de fútbol del Levante como público (solo en la zona de césped) si queremos utilizar este espacio para realizar un concierto de un artista famoso?

~~Teniendo en cuenta el espacio del césped (100x70±~~)
Habría que tener en cuenta la medida del césped del estadio (100 x 70m ±)
Al restar el espacio para el escenario y según protocolos, unas 20.000 personas

Figure 5. Interaction with ChatGPT and the written solution of participant A017 (Source: Authors' own elaboration)

RESULTS

Profiles of ChatGPT Use in Fermi Problem-Solving

Through a triangulated qualitative analysis of participants' prompts and written solutions, three distinct profiles of AI used in modelling problem-solving emerged. First, we observed participants who directly asked ChatGPT to solve the problem without providing any prior prompts, using the *zero-shot prompting* technique. They reproduce in the written solution the response obtained from ChatGPT (as illustrated in **Figure 5**). We categorized this as the *problem-solving expert* (PS-expert) profile.

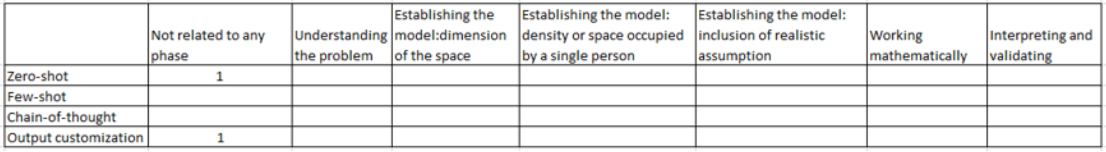
Second, some participants broke down the problem into subproblems—aligning with the phases of the FP-solving process—before asking ChatGPT for a solution, using the *few-shot prompting* technique. These participants rely on AI to generate the solution, which they reproduce in their written response. This behavior was classified as the *problem-solving assistant* (PS-assistant) profile. Finally, other participants followed the phases of the modelling cycle themselves and used ChatGPT strategically at specific stages of the process, employing prompts related to those phases. They independently construct the written solution. This was categorized as the *problem-solving support* (PS-support) profile.

In **Table 4**, we can see examples of prompts used by different participants that allow us to illustrate the difference between the use of ChatGPT as a PS-expert, PS-assistant and PS-support.

Table 4. Complete series of prompts used by three participants in problem PA and the corresponding grid completed during the analysis

ChatGPT as a PS-expert

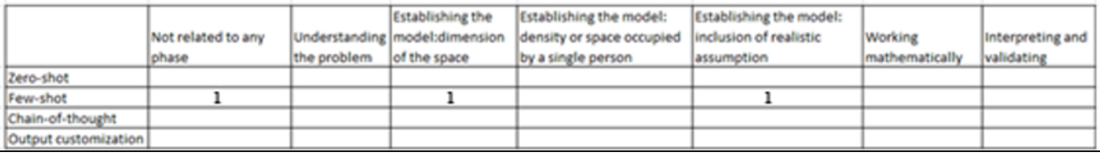
- You are an expert at solving mathematical problems, you have to help me by explaining step-by-step how to solve them and the procedures you would use to obtain the solution.
- How many people can fit in at the Levante football stadium as an audience (only on the grass) if we want to use this space to hold a concert by a famous artist?



	Not related to any phase	Understanding the problem	Establishing the model:dimension of the space	Establishing the model: density or space occupied by a single person	Establishing the model: inclusion of realistic assumption	Working mathematically	Interpreting and validating
Zero-shot	1						
Few-shot							
Chain-of-thought							
Output customization	1						

ChatGPT as a PS-assistant

- What is the surface area of the grass of the Levante football pitch?
- How many people can stand in a space measuring 7,140 meters squared?
- Okay, now we will consider the space occupied by the stage of the singer who will perform in the stadium.

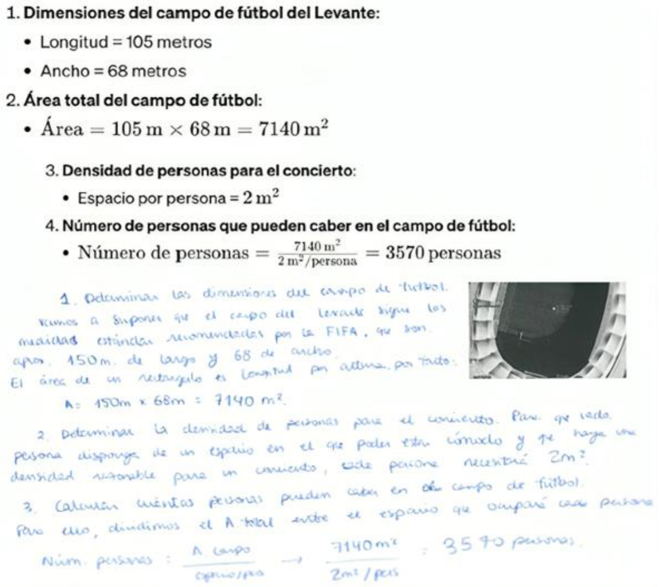


	Not related to any phase	Understanding the problem	Establishing the model:dimension of the space	Establishing the model: density or space occupied by a single person	Establishing the model: inclusion of realistic assumption	Working mathematically	Interpreting and validating
Zero-shot							
Few-shot	1		1		1		
Chain-of-thought							
Output customization							

Table 4 (Continued). Complete series of prompts used by three participants in problem PA and the corresponding grid completed during the analysis

Use of ChatGPT as a PS-support	- How long is the Levante football pitch? - How long is a concert stage usually? - How many people can fit in one square meter?

	Not related to any phase	Understanding the problem	Establishing the model:dimension of the space	Establishing the model: density or space occupied by a single person	Establishing the model: inclusion of realistic assumption	Working mathematically	Interpreting and validating
Zero-shot							
Few-shot			1	1	1		
Chain-of-thought							
Output customization							



1. **Dimensiones del campo de fútbol del Levante:**
   - Longitud = 105 metros
   - Ancho = 68 metros
2. **Área total del campo de fútbol:**
   - Área $= 105\,\text{m} \times 68\,\text{m} = 7140\,\text{m}^2$
3. **Densidad de personas para el concierto:**
   - Espacio por persona = $2\,\text{m}^2$
4. **Número de personas que pueden caber en el campo de fútbol:**
   - Número de personas $= \frac{7140\,\text{m}^2}{2\,\text{m}^2/\text{persona}} = 3570$ personas

1. Determinar las dimensiones del campo de fútbol.
Vamos a suponer que el campo del Levante sigue las medidas estándar recomendadas por la FIFA, que son aprox. 150 m. de largo y 68 de ancho
El área de un rectángulo es longitud por altura, por tanto:
A: 150m × 68m = 7140 m².

2. Determinar la densidad de personas para el concierto. Para que cada persona disponga de un espacio en el que poder estar cómodo y que haya una densidad razonable para un concierto, cada persona necesitará 2m².

3. Calcular cuántas personas pueden caber en el campo de fútbol. Para ello, dividimos el A total entre el espacio que ocupará cada persona

Núm. personas: $\frac{\text{A campo}}{\text{espacio/pers}} \rightarrow \frac{7140\,\text{m}^2}{2\text{m}^2/\text{pers}}$ = 3570 personas.

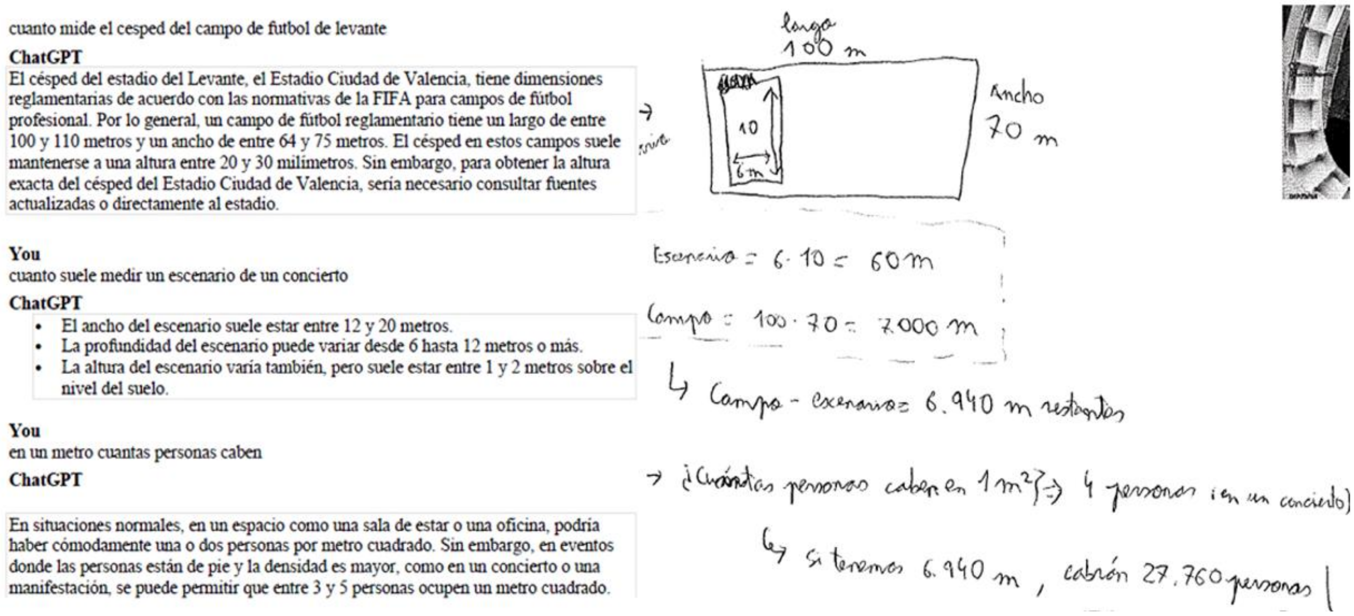
Figure 6. Answer to ChatGPT and written report of a participant solving PA categorized using ChatGPT as PS-expert (Source: Authors' own elaboration)

The first participant (PS-expert) asks ChatGPT to act as an "expert" (*output customization*) and copies the full formulation of the problem into the chat, without any further guidance (*zero-shot*). The AI's response is then transferred almost verbatim into the participant's written report (see **Figure 6**).

The second participant (PS-assistant) begins by asking for the area of the football pitch, corresponding to the FP-solving process (EM1). Then, they ask how many people could fit in that area. At this point, the AI determines both the modelling strategy (e.g., using density or unit iteration) and carries out the mathematical operations, meaning the participant delegates several key phases of the modelling cycle. Finally, the participant introduces a realistic assumption (EM3) in a follow-up prompt, refining the AI's output using the *few-shot* technique.

The third participant (PS-support) uses ChatGPT to obtain numerical estimates of specific variables but never asks for a full solution. In this case, AI plays a purely instrumental role: it is used to quantify elements already selected by the problem-solver (EM1, EM2 and EM3), who retain control over the entire FP-solving process. As shown in the written response (**Figure 7**), it is the participant who performs the necessary calculations to arrive at the final solution independently.

The key distinction between PS-assistant and PS-support lies in if participants delegate critical modelling



cuanto mide el cesped del campo de futbol de levante

ChatGPT

El césped del estadio del Levante, el Estadio Ciudad de Valencia, tiene dimensiones reglamentarias de acuerdo con las normativas de la FIFA para campos de fútbol profesional. Por lo general, un campo de fútbol reglamentario tiene un largo de entre 100 y 110 metros y un ancho de entre 64 y 75 metros. El césped en estos campos suele mantenerse a una altura entre 20 y 30 milímetros. Sin embargo, para obtener la altura exacta del césped del Estadio Ciudad de Valencia, sería necesario consultar fuentes actualizadas o directamente al estadio.

You

cuanto suele medir un escenario de un concierto

ChatGPT

- El ancho del escenario suele estar entre 12 y 20 metros.
- La profundidad del escenario puede variar desde 6 hasta 12 metros o más.
- La altura del escenario varía también, pero suele estar entre 1 y 2 metros sobre el nivel del suelo.

You

en un metro cuantas personas caben

ChatGPT

En situaciones normales, en un espacio como una sala de estar o una oficina, podría haber cómodamente una o dos personas por metro cuadrado. Sin embargo, en eventos donde las personas están de pie y la densidad es mayor, como en un concierto o una manifestación, se puede permitir que entre 3 y 5 personas ocupen un metro cuadrado.

Escenario = 6 · 10 = 60 m

Campo = 100 · 70 = 7.000 m

↳ Campo - escenario = 6.940 m restantes

→ ¿Cuántas personas caben en 1 m²? ⇒ 4 personas (en un concierto)

↳ Si tenemos 6.940 m, cabrán 27.760 personas

Figure 7. Answer to ChatGPT and written report of a participant solving PA categorized using ChatGPT as PS-support (Source: Authors' own elaboration)

Table 5. Frequency of prompts according to the use of ChatGPT (we use *italics* for frequency of use as PS-expert, **bold** for use as PS-assistant, and standard for use as PS-support), we include the percentage of each prompt in each category of use

	NotM	UP	EM1	EM2	EM3	WM	IV
PS-expert							
Zero-shot	*61 (100%)*						
Few-shot							
Chain-of-thought	*10 (16.4%)*						
Output customization	*11 (18.0%)*						
PS-assistant							
Zero-shot							
Few-shot	**106 (100%)**	**9 (8.5%)**	**93 (87.7%)**	**25 (23.6%)**	**17 (16.0%)**	**21 (19.8%)**	**4 (3.8%)**
Chain-of-thought	**7 (6.6%)**						
Output customization	**9 (8.5%)**						
PS-support							
Zero-shot							
Few-shot		3 (3.0%)	92 (92.2%)	50 (50.5%)	33 (33.3%)	17 (17.2%)	2 (2.0%)
Chain-of-thought							
Output customization							

Table 6. Frequency of use of ChatGPT according to the problem posed (P_A and P_B)

	As a PS-expert	As a PS-assistant	As a PS-support
P_A	28 (21.1%)	44 (33.1%)	61 (45.8%)
P_B	33 (24.8%)	62 (46.6%)	38 (25.6%)

Table 7. Frequency of ChatGPT use according to prior experience

	As a PS-expert	As a PS-assistant	As a PS-support
Out group	8 (10.0%)	33 (41.2%)	39 (48.8%)
AI group	12 (15.4%)	27 (34.6%)	39 (50.0%)
RF group	41 (38.0%)	46 (42.6%)	21 (19.4%)

decisions to AI or retain control over the modelling process while using the AI as a technical resource.

When we analyze, globally, the use of ChatGPT in problems P_A and P_B , we observe that most of the participants used AI as a PS-assistant (N = 106, 39.9%), followed by those who used it as a PS-support (N = 99, 37.2%) and those who used it as a PS-expert (N = 61, 22.9%).

Table 5 shows the frequencies observed of the different prompts for each of the three uses of the AI (PS-support, PS-assistant and PS-expert) for problems P_A and P_B. To calculate the frequencies, the first time each type of prompt was identified has been counted and not the number of times that it might have been used during the interaction (for example, due to the reformulation of questions through small semantic or syntactic modifications).

We observe that when using ChatGPT as PS-expert, we always find a *zero-shot* prompt that can be followed by a *chain-of-thought* or *output customization* type prompt to obtain a more accurate answer.

In its use as a PS-assistant, they used *few-shot* prompts with prior questions related to the phases of the FP-solving process (mostly related to the dimensions of the space to be occupied) to condition the IA response.

In its use as a PS-support, there is greater use of the prompt related to density, the area occupied by a person, or to realistic assumptions. It is important to point out that prompt related to understanding the problem contributes to participants being able to construct their model.

Uses of ChatGPT According to the Complexity of the Problem

Table 6 shows the frequency and percentage of use of the AI for each of the problems posed (P_A and P_B).

We observe that the majority of the participants solve P_A using ChatGPT as a PS-support (45.8%), followed by PS-assistant (33.1%) and PS-expert (21.1%), while in P_B, a more complex problem, almost half of the participants use it as a PS-assistant (46.6%), comprising practically the same percentage of use as a PS-support and PS-expert (25.6% and 24.8%, respectively). To determine if there is a significant relationship between the complexity of the problems (increasing from P_A to P_B) and the profile of ChatGPT use in problem-solving, we perform a Chi-square test. The result is χ^2 (2, 266) = 8.81 with $p < .05$ and effect size V = 0.18, which indicates a moderate association. Therefore, the complexity of modelling problems significantly, albeit moderately, affects the profile of ChatGPT use during their solution. In the analysis, it is observed that in P_A (less complex) the use of ChatGPT as a PS-support is 27.5% higher than expected if they were independent variables, while in P_B (more complex), this expected use is 29.2% less.

Use of ChatGPT According to Prior Experience

Table 7 shows the frequency and percentage of ChatGPT use made by participants according to the type of help they have received in prior experience with FPs in close contexts.

Table 8. Use of ChatGPT according to the problem and prior experience

	As a PS-expert		As a PS-assistant		As a PS-support	
	P_A	P_B	P_A	P_B	P_A	P_B
Out group	4 (10.0%)	4 (10.0%)	12 (30.0%)	21 (52.5%)	24 (60.0%)	15 (37.5%)
AI group	3 (7.7%)	9 (23.0%)	12 (30.8%)	15 (38.5%)	24 (61.5%)	15 (38.5%)
RF group	21 (38.9%)	20 (37.0%)	20 (37.0%)	26 (48.1%)	13 (24.1%)	8 (14.8%)

It can be observed that the group that worked on previous problems without any type of assistance (RF group) is the group that has a lower percentage of AI use as a PS-support (19.4%), being by far the group that most delegates problem-solving to ChatGPT as a PS-expert (38% compared to 10% and 15.4% of the other two groups). There are very few differences regarding the use of AI as a PS-support between the other two groups (48.8% for the out group, 50% for the AI group), which means that practically half the participants of these groups assume full responsibility for the solution.

To confirm that the profile of ChatGPT used in the solution of FPs depends, in a statistically significant way, on the type of prior experience solving simpler modelling problems, a Chi-square test is performed, obtaining $\chi^2 = 34.64$ with $p < .001$ and $V = 0.26$, which indicates a medium effect size. Therefore, the type of prior experience has a medium strong influence on the use of ChatGPT during problem-solving. In fact, in the RF group it is observed that the use of ChatGPT as a PS-support is 47.8% lower than that expected if there was no influence from prior experience, while their use of ChatGPT as PS-expert is 65.5% higher than expected. On the contrary, in the out group the use of ChatGPT as a PS-support is 31% higher than expected and as a PS-expert, 56.4% lower. In the AI group a similar pattern is observed: the use of ChatGPT as a PS-support is 34.3% higher than expected and as a PS-expert 32.9% lower.

To complete the previous results, in **Table 8** we can see this use of AI divided according to prior experience (without support, with experimental support, with ChatGPT support) and the complexity of the problem posed (P_A and P_B).

Based on **Table 8**, we can use the McNemar test to compare intra-group proportions to confirm if the changes in ChatGPT use according to the complexity of the problem depend on the type of prior experience. In the case of future teachers with prior experience solving real problems outdoor (out group), it is observed that there is no change in the use of ChatGPT as a PS-expert when the complexity of the problem increases; however, with the McNemar test we obtain that the increase in the use of ChatGPT as a PS-assistant (which goes from 30% to 52.5%) is significant ($p < .01$), as is the decrease in the use of ChatGPT as a PS-support ($p < .01$).

In the case of future teachers with prior experience solving real problems with support from ChatGPT (AI group), we observe that there is a significant increase in the use of ChatGPT as a PS-expert when the complexity of the problem increases ($p < .05$) and a decrease in the use of ChatGPT as a PS-support ($p < .01$). In this group, there is no significant difference in the use of ChatGPT as a PS-assistant when complexity increases ($p = .25$). As regards future teachers with prior experience solving real problems without support (RF group), the use of ChatGPT as a PS-assistant also increases significantly when the complexity of the problem increases ($p < .05$) but there is no significant difference in its use as a PS-expert (it is almost identical) or as a PS-support ($p = .06$).

DISCUSSION

Categorization of the Prompts and Use of ChatGPT

In our analysis of the interactions of PSTs with the AI, we have been able to refine and adapt the prompt types described in prompt engineering (Schorcht et al., 2024; Walter, 2024) to the context of modelling problem-solving. Moreover, this double categorization-based on prompting techniques and phases of the FP-solving process-allows us to categorize the participants according to how they use AI in the context of modelling problem-solving.

We differentiate between those who delegate obtaining the result to the AI (use as PS-expert) and those who use the AI as a support tool for their solution. In this last case, we observe a clear distinction between participants who use it only as support resource, asking very specific questions, mainly aimed at quantifying the variables within their model (use as PS-support), and those who skip certain phases of the modelling process, partially delegating them to the AI as an assistant (use as PS-assistant).

Regarding the use of ChatGPT as PS-expert, participants used a *zero-shot* prompt by directly copying the problem statement into the chat (see **Figure 5**). In a few cases, PSTs asked ChatGPT for more detailed explanations of the procedure, using a *chain-of-thought* prompt.

When we focus on the prompts that can be directly related to the phases of the modelling process (that correspond to use of ChatGPT as PS-assistant or as PS-support), we have found that PSTs identify the dimensions of the space to be occupied as the main variable that they must quantify to be able to solve the problem. Indeed, the most used prompt corresponds to a *few-shot* prompt related to the phase "Establishing the model: dimensions of an element or density" (EM1). However, in some cases, the *few-shot* prompts based on the modelling cycle are also used by those participants (that use ChatGPT as PS-assistant) who want to train or condition the AI before presenting a copy of the problem statement in which nuances or refinements obtained from the previous prompts are introduced. For instance, we observed that the *few-shot* prompts related to the first phase of the FP-solving process (understanding the problem) are mainly used in this way.

The categorization of AI usage proposed here represents a theoretically grounded outcome of our empirical analysis. These profiles serve as the basis for examining how AI use varies with task complexity and prior experience in the subsequent RQs.

Influence of Complexity on the Use of ChatGPT

To investigate how task complexity affects the use of ChatGPT in mathematical modelling, we analyzed participant behavior in two FPs of differing complexity. **Table 6** shows how the distribution of ChatGPT use profiles (PS-support, PS-assistant, and PS-expert) varied between the two problems (P_A and $P_{B)}$, reflecting the impact of task complexity on participants' problem-solving strategies. Although all participants appeared to conflate maximum occupancy (P_A) with maximum influx (P_B), relying mainly on spatial variables in both cases, P_B involved a less structured context than P_A, which was more structurally similar to P1 from phase 1. This increase in complexity is reflected in the greater delegation to ChatGPT observed in P_B. In P_A, most participants adopted a PS-support profile, using ChatGPT strategically while maintaining responsibility for the modelling process. In contrast, P_B led many participants to delegate part or all of the task to AI, shifting towards PS-assistant and, to a lesser extent, PS-expert profiles. This shift reflects how increased contextual and structural complexity influences problem-solving delegation (Knabbe et al., 2025; Maass, 2010). Under higher demands, readily available support tools may be used as a stronger substitute for students' own problem-solving work, especially when confidence is low (Lumandas & Taja-on, 2026).

These findings suggest that when faced with increased complexity, PSTs tend to delegate more of the problem-solving process to AI. This pattern aligns with previous research (Noster et al., 2024) highlighting that task characteristics—particularly complexity—significantly shape the way users interact with AI tools in educational modelling contexts.

Influence of Prior Experience on the Use of ChatGPT

Our analysis reveals that the use of AI by PSTs is influenced not only by problem complexity, but also by their prior experience with FPs. Participants who previously solved problems without any support (RF group) were more likely to delegate responsibility to ChatGPT, frequently adopting the PS-expert profile. Interestingly, this pattern was not observed in participants from the out group, who had also never used AI before but had worked on problems directly in the real-world context.

This suggests that, due to their specific characteristics, FPs benefit from being addressed through an experimental approach that allows solvers to explore the complexity of the situation and thereby enrich their models (Segura et al., 2023). In this way, when participants are faced with solving a problem outside the context in which it is formulated, they are more capable of breaking it down into subproblems and completing the modelling cycle more autonomously. Our findings regarding the value of providing opportunities for experimental work when solving modelling problems align with the results reported by Buchholtz (2021) and Jablonski (2023).

In the case of participants who already had experience using AI in the problem-solving process (AI group), we also observed that they tend to delegate less responsibility to the tool. This is significant, as it suggests that when modelling tasks cannot be addressed experientially, continued interaction with ChatGPT may help participants develop modelling-related skills.

Influence of the Variables of Complexity and Prior Experience on the Use of ChatGPT to Solve Fermi Problems

Lastly, the results presented in **Table 8** allow us to examine how both prior experience and problem complexity jointly influence the use of ChatGPT. When analyzing the variation of the uses across problems P_A and P_B by group, we observe that the shift percentage of PS-support use between these two problems differs depending on prior experience: in the out and RF groups, the change mainly corresponds to an increase in PS-assistant use, whereas in the AI group, it primarily corresponds to PS-expert use. Notably, the percentage for PS-expert use remains very similar across both problems in the out and RF groups, but in the AI group, it increases markedly from 7.7% to 23%.

For the RF group, the increase in complexity leads to slight rise in PS-assistant use with only a mirror decrease in PS-support use, insignificant due to its already low level. In contrast, in the other two groups, increased complexity results in a significant drop in the use of ChatGPT as support. In other words, the more complex the problem, the less autonomy participants show in solving it. However, the behavior of the out group differs while they shift towards using ChatGPT as an assistant, they still assign less responsibility to the tool than the AI group, which tends to use ChatGPT as a PS-expert.

These differences may be explained by the AI group's perception of problem P_B as more complex than P_A. Given their familiarity with the tool, they are more likely to fully delegate the solution process to it. In contrast, prior experience solving FPs in modelling contexts may help the out group formulate more realistic assumptions (Jablonski, 2023). When faced with a more complex task like P_B, they are able to integrate modelling-related prompts into the interaction, refining prompts or complementing the AI's response.

It may seem paradoxical that the AI group, overall, uses ChatGPT mostly as PS-support—approaching the percentage seen in the out group, around 50%, compared to just 19.4% in the RF group, in **Table 7**—considering that having used AI in the earlier part of the study could lead them to delegate unfamiliar tasks to it. As noted earlier, this does occur when complexity increases. However, this familiarity also appears to foster greater autonomy and confidence in taking responsibility for the problem-solving process, as long as the complexity remains comparable to previous tasks.

In both the out and AI groups, we observe a stronger transfer between their initial experience with familiar problems and their approach to problem P_A. This transfer is not evident in the RF group. The higher dependency on AI observed in this group appears to stem from the lack of prior support. In contrast, participants in the other two groups had access to either direct engagement with the problem context (out group) or the opportunity to compare their estimates and strategies with those of the AI (AI group), which may have fostered their trust and autonomy.

Limitations

Several limitations of this study should be acknowledged. First, although the sample size was relatively large, all participants were drawn from a single university. In addition, relevant individual characteristics—such as previous use of ChatGPT, levels of digital or AI literacy, or prior competence in mathematical modelling—were not assessed. These factors may have influenced how PSTs interacted with the AI tool and future studies should consider more diverse samples and learner profiles.

Second, while statistically significant associations were identified, effect sizes were modest. Moreover, the analysis relied primarily on log data and written artefacts, which capture observable interactions but do not provide direct access to participants' intentions, reasoning, or decision-making processes. These methodological choices limit the interpretative scope of the findings, which should therefore be understood as indicative tendencies rather than strong explanatory claims. Complementary qualitative methods, such as interviews or think-aloud protocols, could provide deeper insight into how and why participants delegate modelling processes to AI.

Third, the analysis focused on participants' use of ChatGPT rather than on the mathematical or pedagogical quality of the AI-generated responses themselves. Examining the correctness, coherence, and educational suitability of AI output would offer important complementary perspectives on the affordances and risks of AI-supported modelling. Moreover, the study was conducted using ChatGPT version 3.5. Given the rapid evolution of generative AI systems, newer models may lead to different interaction patterns and outcomes, which limits direct replicability and underscores the need for ongoing research in this area.

Overall, the findings should be interpreted as exploratory, providing an empirically grounded starting point for further research on the integration of generative AI into mathematical modelling problem-solving.

CONCLUSIONS AND IMPLICATIONS

This study explored how PSTs interact with ChatGPT when solving mathematical modelling tasks, specifically FPs. While previous research has demonstrated that the contextual characteristics of these problems influence problem-solving strategies (Ferrando et al., 2021), and that the setting in which they are addressed affects performance (Segura et al., 2023), few studies have examined how the use of AI varies depending on specific contextual variables. Recent work has also highlighted the importance of analyzing prompting techniques (Noster et al., 2024; Schorcht et al., 2024), as well as the role of complexity in shaping the reliability of AI-generated solutions (Spreitzer et al., 2024). Building on these contributions, our study fills a critical gap by examining how prior experience and task complexity shape PSTs' engagement with ChatGPT during FP-solving.

Methodologically, we capitalized on the didactic affordances of FPs to bridge real-world reasoning and mathematical thinking (Ärlebäck, 2009; Peter-Koop, 2009). We analyzed interactions with ChatGPT based on a categorization of the prompts used and the written solutions provided by participants. We applied a double deductive coding strategy, using a predefined framework drawn from the modelling process literature (Borromeo-Ferri, 2006) and prompt engineering studies (Schorcht et al., 2024; Walter, 2024). This framework allowed us to classify each prompt according to its relationship with the modelling process and the type of prompt used. By doing so, we systematically examined how participants engaged with AI.

The application of this framework allowed us to examine how prior experience and task complexity influenced AI use. Participants who had received some type of support—either through previous use of ChatGPT (IA group) or through solving problems in the physical problem context (out group)—tended to use AI in more autonomous and reflective ways, often as PS-support. In contrast, those with no prior support (RF group) more often used AI as an assistant or expert.

Regarding task complexity, the less complex problem (P_A) was generally approached using AI as PS-support, while the most complex problem (P_B) led to more frequent use of AI as an assistant.

Our findings complement previous studies on how the context and setting of FPs influence problem-solving (Ferrando et al., 2021; Segura et al., 2023). We show that prior experience and task complexity also significantly affect the nature of AI use. These results reinforce the importance of instructional design in shaping how PSTs engage with modelling tasks, highlighting the need to place mathematical modelling at the center of learning activities and foster connections between real-world phenomena and mathematical knowledge (Geiger et al., 2022).

The study also offers important implications for practice. While previous research (Bani-Hamad & Al-Kalbani, 2024) has demonstrated that ChatGPT can be integrated into FP-solving tasks, our results highlight the risk of over-dependence on AI when support is not properly scaffolded. At the same time, delegating specific aspects of the modelling process to AI—such as routine calculations or the generation of initial estimates—may be pedagogically justified when it allows learners to focus on higher-level reasoning, interpretation, and validation. From a didactic perspective, this suggests the need for instructional sequences that progressively combine task complexity, problem-solving settings, and forms of AI support (Zhuang & Zhang, 2025). This is especially relevant given evidence that ChatGPT still struggles with complex or domain-specific tasks (e.g., aspects of spatial geometry), underscoring the importance of matching AI support to task demands (Almarashdi et al., 2024; Fock & Siller, 2025). In line with work in teacher education, preparation should therefore address not only how to use ChatGPT but also how to critically appraise its output, acknowledging potential limitations and the need for context-sensitive use and human oversight (Guardia-Paniura et al., 2026; Lumandas & Taja-on, 2026).

This study opens several avenues for future research. Beyond extending the framework to other populations (e.g., school students) and settings (e.g., collaborative modelling), we suggest comparative designs in which ChatGPT support is purposefully restricted to specific phases of the modelling cycle—for instance, supporting estimation or validation only—rather than acting as a general-purpose solver. Future studies could also embed explicit instruction and deliberate practice in prompt design within teacher-education courses and examine whether such preparation changes PSTs´ agency in decision-making and their uptake of AI-mediated suggestions during modelling. In domains where persistent difficulties are expected (e.g., tasks involving spatial reasoning), research could focus on how PSTs review AI output and what kinds of scaffolds foster critical checking. Finally, future work should examine how affective variables—such as perceived value, self-efficacy, and attitudes towards AI—influence prompting behavior and problem-solving outcomes. As AI tools continue to evolve, identifying pedagogical strategies that support critical, responsible use in modelling contexts will remain central to their meaningful integration in mathematics education.

Author contributions: CG, IF, CS, & LA: conceptualization and writing–review & editing; **CG & IF:** writing–original draft; & **CG & CS:** data curation and formal analysis. All authors agreed with the results and conclusions.

Funding: This study was funded by the MCIU/AEI/10.13039/ 501100011033 with the grant no. PID2021-126707NB-I00 and "ERDF a way of making Europe", as well as from the Special Research Actions Grant UV-INV-AE-3668116 and from the Generalitat Valenciana through project CIGE/2024/64.

Ethical statement: The authors stated that the study was approved by the Committee of Ethics and Human Research at the University of Valencia on 8 July 2025 with approval number 2025-MAGPED-3909285. Written informed consents were obtained from the participants.

AI statement: The authors stated that no generative artificial intelligence tools or AI-based software were used in the design, data collection, analysis, or writing of this study.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Albarracín, L., & Gorgorió, N. (2014) Devising a plan to solve Fermi problems involving large numbers, *Educational Studies.in Mathematics*, *86*, 79-96. https://doi.org/10.1007/s10649-013-9528-9

Albarracín, L., & Ärlebäck, J. (2019). Characterising mathematical activities promoted by Fermi problems. *For the Learning of Mathematics*, *39*(3), 10-13. https://ddd.uab.cat/record/296776

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

Ärlebäck, J. B. (2009). On the use of realistic Fermi problems for introducing mathematical modelling in school. *The Mathematics Enthusiast*, *6*(3), 331-364. https://doi.org/10.54870/1551-3440.1157

Bani-Hamad, A. M. H., & Al-Kalbani, M. S. A. (2024). Fermi problem-based learning with artificial intelligence: Is it effective to develop United Arab Emirates cycle three students' twenty-first century skills? In A. Al-Marzouqi, S. A. Salloum, M. Al-Saidat, A. Aburayya, & B. Gupta, (Eds.), *Artificial intelligence in education: The power and dangers of*

ChatGPT in the classroom (pp. 113-125). Springer. https://doi.org/10.1007/978-3-031-52280-2_8

Blum, W., & Borromeo-Ferri, R. (2009). Mathematical modelling: Can it be taught and learnt? *Journal of Mathematical Modelling and Application, 1*(1), 45-58.

Blum, W., & Leiss, D. (2007). How do students and teachers deal with modelling problems. In C. Haines, P. L. Galbraith, W. Blum, & S. Khan (Eds.), *Mathematical modelling (ICTMA 12): Education, engineering and economics* (pp. 222-231). Horwood. https://doi.org/10.1533/9780857099419.5.221

Borromeo-Ferri, R. (2006). Theoretical and empirical differentiations of phases in the modelling process. *ZDM Mathematics Education, 38*, 86-95. https://doi.org/10.1007/BF02655883

Borromeo-Ferri, R. (2021). Mandatory mathematical modelling in school: What do we want the teachers to know? In F. K. S. Leung, G. Stillman, G. Kaiser, & K. L. Wong (Eds.), *Mathematical modelling education in East and West* (pp. 103-117). Springer. https://doi.org/10.1007/978-3-030-66996-6_9

Buchholtz, N. (2021). Processos de modelação dos alunos envolvidos em trilhos matemáticos [Modeling processes of students involved in mathematical tracks]. *Quadrante, 30*(1), 140-157. https://doi.org/10.48489/quadrante.23699

Celik, I (2023). Towards intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education, *Computers in Human Behavior, 138*, 107468. https://doi.org/10.1016/j.chb.2022.107468

Cevikbas, M., Kaiser, G., & Schukajlow, S. (2022). A systematic literature review of the current discussion on mathematical modelling competencies: State-of-the-art developments in conceptualizing, measuring, and fostering. *Educational Studies in Mathematics, 109*(2), 205-236. https://doi.org/10.1007/s10649-021-10104-6

Clarke, D., & Roche, A. (2018). Using contextualized tasks to engage students in meaningful and worthwhile mathematics learning. *The Journal of Mathematical Behavior, 51*, 95-108. https://doi.org/10.1016/j.jmathb.2017.11.006

Czocher, J. A. (2016). Introducing modeling transition diagrams as a tool to connect mathematical modeling to mathematical thinking. *Mathematical Thinking and Learning, 18*(2), 77-106. https://doi.org/10.1080/10986065.2016.1148530

DiNapoli, J., & Miller, E. K. (2022). Recognizing, supporting, and improving student perseverance in mathematical problem-solving: The role of conceptual thinking scaffolds. *The Journal of Mathematical Behavior, 66*, Article 100965. https://doi.org/10.1016/j.jmathb.2022.100965

Efthimiou, C. J., & Llewellyn, R. A. (2007). Cinema, Fermi problems and general education. *Physics Education, 42*(3), Article 253. https://doi.org/10.1088/0031-9120/42/3/003

English, L. D. (2023). Multidisciplinary modelling in a sixth-grade tsunami investigation. *International Journal of Science and Mathematics Education, 21*, 41-65. https://doi.org/10.1007/s10763-022-10303-4

English, L. D., & Gainsburg, J. (2015). Problem solving in a 21st-century mathematics curriculum. In L. D. English, & D. Kirshner (Eds.), *Handbook of international research in mathematics education* (pp. 313-335). Routledge. https://doi.org/10.4324/9780203448946-20

Ferrando, I., Segura, C., & Pla-Castells, M. (2021). Analysis of the Relationship Between Context and Solution Plan in Modelling Tasks Involving Estimations. In Leung, F.K.S., Stillman, G.A., Kaiser, G., Wong, K.L. (Eds.), *Mathematical Modelling Education in East and West. International Perspectives on the Teaching and Learning of Mathematical Modelling*. (pp. 119-128). Springer. https://doi.org/10.1007/978-3-030-66996-6_10

Fock, A., & Siller, H. S. (2025). Generative artificial intelligence in secondary STEM education in the light of human flourishing: A scoping literature review. *International Journal of STEM Education, 12*, Article 67. https://doi.org/10.1186/s40594-025-00589-5

Geiger, V., Galbraith, P., Niss, M., & Delzoppo, C. (2022). Developing a task design and implementation framework for fostering mathematical modeling competencies. *Educational Studies in Mathematics, 109*(2), 313-336. https://doi.org/10.1007/s10649-021-10039-y

Gravemeijer, K., & Doorman, M. (1999). Context problems in realistic mathematics education: A calculus course as an example. *Educational studies in Mathematics, 39*(1), 111-129. https://doi.org/10.1023/A:1003749919816

Guardia-Paniura, C. H., Cueva Luza, T., Cruz Carpio, F. M., Ito Díaz, R. R., Apaza Paco, D. V., Rosas Rojas, N., Mamani Mamani, B., Terrero Pérez, Á., Yaedú, R. Y. F., & Peralta Mamani, M. (2026). Human and AI generated feedback in higher education: A systematic review of effectiveness and student perceptions. *Contemporary Educational Technology, 18*(1), Article ep623. https://doi.org/10.30935/cedtech/17863

Hartmann, L. M., & Schukajlow, S. (2021). Interest and emotions while solving real-world problems inside and outside the classroom. In F. K. S. Leung, G. A. Stillman, G. Kaiser, & K. L. Wong (Eds.), *Mathematical modelling education in East and West* (pp. 153-163). Springer. https://doi.org/10.1007/978-3-030-66996-6_13

Jablonski, S. (2023). Indoors vs. outdoors: Student perception of different modelling settings. *Research in Integrated STEM Education, 1*(3), 403-421. https://doi.org/10.1163/27726673-bja00016

Jablonski, S., Barlovits, S., & Ludwig, M. (2023). How digital tools support the validation of outdoor modelling results. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1145588

Knabbe, A., Leiss, D., & Ehmke, T. (2025). Reality-based tasks with complex-situations: Identifying sociodemographic and cognitive factors for solution. *International Journal of Sciences and Mathematics Education, 23*, 97-120. https://doi.org/10.1007/s10763-024-10463-5

Krawitz, J., Hartmann, L., & Schukajlow, S. (2024). Do task variables of self-generated problems influence interest? Authenticity, openness, complexity, and students' interest in solving self-generated modelling problems. *The Journal of Mathematical Behavior, 73*, Article 101129. https://doi.org/10.1016/j.jmathb.2024.101129

Krawitz, J., Schukajlow, S., Yang, X., & Geiger, V. (2025). A systematic review of international perspectives on mathematical modelling: Modelling goals and task characteristics. *ZDM Mathematics Education, 57*, 193-212. https://doi.org/10.1007/s11858-025-01683-2

Lesh, R., & Doerr, H. M. (Eds.). (2003). *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching*. Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9781410607713

Lin, S. M., Chung, H. H., Chung, F. L., Lan, Y. J. (2023). Concerns about using ChatGPT in education. In Y. M. Huang, & T. Rocha (Eds.), *Innovative technologies and learning* (pp.37-49). Springer. https://doi.org/10.1007/978-3-031-40113-8_4

Lo, C. K. (2023) What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), Article 410. https://doi.org/10.3390/educsci13040410

López-Simó, V., & Rezende, M. F. (2024). Challenging ChatGPT with different types of physics education questions. *The Physics Teacher, 62*, 290-294. https://doi.org/10.1119/5.0160160

Lumandas, E. R., & Taja on, E. P. (2026). The paradox of accessibility: Investigating mathematics struggle among college students in the age of information and artificial intelligence. *Educational Point, 3*(1), Article e147. https://doi.org/10.71176/edup/17801

Maass, K. (2006). What are modelling competencies? *ZDM Mathematics Education, 38*, 113-142. https://doi.org/10.1007/BF02655885

Maass, K. (2010). Classification scheme for modelling tasks. *Journal für Mathematik-Didaktik, 31*(2), 285-311. https://doi.org/10.1007/s13138-010-0010-2

Maass, K., Geiger, V., Ariza, M.R., & Goos, M. (2019). The role of mathematics in interdisciplinary STEM education. *ZDM Mathematics Education, 51*, 869-884. https://doi.org/10.1007/s11858-019-01100-5

Noster, N., Gerber, S., & Siller, H. S. (2024). Pre-service teachers' approaches in solving mathematics tasks with ChatGPT. *Digital Experiences in Mathematics Education, 10*, 543-567. https://doi.org/10.1007/s40751-024-00155-8

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

Peter-Koop, A. (2009). Teaching and understanding mathematical modelling through Fermi-problems. In B. Clarke, B. Grevholm, & R. Millman (Eds.), *Tasks in primary mathematics teacher education* (pp. 131-146). Springer. https://doi.org/10.1007/978-0-387-09669-8_10

Quarder, J., Greefrath, G., Gerber, S., & Siller, H. S. (2025). Pedagogical content knowledge for simulations and mathematical modelling with digital tools: A quasi-experimental study with pre-service mathematics teachers. *ZDM Mathematics Education, 57*, 395-409. https://doi.org/10.1007/s11858-025-01673-4

Robinson, A. W. (2008). Don't just stand there—Teach Fermi problems! *Physics Education, 43*(1), 83-87. https://doi.org/10.1088/0031-9120/43/01/009

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*. https://doi.org/10.3389/feduc.2024.1386075

Schukajlow, S., Kaiser, G., & Stillman, G. (2018). Empirical research on teaching and learning of mathematical modelling: A survey on the current state-of-the-art. *ZDM Mathematics Education, 50*, 5-18. https://doi.org/10.1007/s11858-018-0933-5

Segura, C., & Ferrando, I. (2023). Pre-service teachers' flexibility and performance in solving Fermi problems. *Educational Studies in Mathematics, 113*(2), 207-227. https://doi.org/10.1007/s10649-023-10220-5

Segura, C., Ferrando, I., & Albarracín, L. (2023). Does collaborative and experiential work influence the solution of real-context estimation problems? A study with prospective teachers. *The Journal of Mathematical Behavior, 70*, Article 101040. https://doi.org/10.1016/j.jmathb.2023.101040

Segura, C., Gallart, C., & Ferrando, I. (2025). Influence of pre-service primary school teachers' prior knowledge of measurement and measurement estimation in solving modelling problems. *Journal of Mathematics Teacher Education*, 1-26. https://doi.org/10.1007/s10857-025-09685-3

Spreitzer, C., Straser, O., Zehetmeier, S., & Maass, K. (2024). Mathematical modelling abilities of artificial intelligence tools: The case of ChatGPT. *Education Sciences, 14*, Article 698. https://doi.org/10.3390/educsci14070698

Stanic, G. M., & Kilpatrick, J. (1989). Historical perspectives on problem solving in the mathematics curriculum. In R. I. Charles, & E. A. Silver (Eds.), *The teaching and assessing of mathematical problem solving* (pp. 1-22). NCTM/Lawerance Erlbaum Associates. https://doi.org/10.4324/9781003726807-1

Thompson, P. W. (1985). Experience, problem solving, and learning mathematics: Considerations in developing mathematics curricula. In E. A. Silver (Ed.), *Learning and teaching mathematical problem solving: Multiple research perspectives*. Franklin Institute Press.

Walter, Y. (2024). Embracing the future of Artificial Intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education, 21*, Article 15. https://doi.org/10.1186/s41239-024-00448-3

Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1181712

Zhang, K., & Aslan, A.B. (2021). AI technologies for education: Recent research & future directions. *Computer and Education: Artificial Intelligence, 2*, Article 100025. https://doi.org/10.1016/j.caeai.2021.100025

Zhuang, Y., & Zhang, S. (2025). Pre service mathematics teachers' perceptions of using GenAI for practicing teacher questioning: A semester long study. *Eurasia Journal of Mathematics, Science and Technology Education, 21*(9), Article em2689. https://doi.org/10.29333/ejmste/16764

https://www.ejmste.com