Can LLMs Help Students Learn Accounting? A Performance Analysis Across Models and Prompts

Bàrbara Llacay[a,1], Maria Pilar Curós Vilà[b]

[a] University of Barcelona, Department of Business, Faculty of Economics and Business, Av. Diagonal 690, 08034 Barcelona, Spain. E-mail: bllacay@ub.edu

[b] University of Barcelona, Department of Business, Faculty of Economics and Business, Av. Diagonal 696, 08034 Barcelona, Spain. E-mail: pilar.curos@ub.edu

Abstract

In recent years, the emergence of large language models (LLMs) has transformed the landscape of higher education. Among their most promising applications is the potential to personalise student learning and serve as virtual tutors. This paper investigates whether current versions of the leading LLMs — ChatGPT, Gemini, CoPilot, and Claude — can function as effective learning assistants for accounting students. To this end, we evaluate the accuracy of these models in responding to a set of theoretical and practical multiple-choice questions taken from real university exams. We also examine whether simple prompting strategies, accessible to any student, can improve model performance. Our findings show that while LLMs achieve high accuracy on theoretical questions, their performance declines when solving practical tasks. In particular, accuracy remains insufficient in questions involving the preparation of financial statements. However, when testing more advanced models — ChatGPT-o3 and Claude Opus 4 — we observe a significant leap in performance, suggesting a paradigm shift. These results indicate that, in the near future, accounting students and learners in related fields may be able to reliably use LLMs to support and enhance their independent study.

Keywords: LLM, Generative AI, Accounting, Financial Statements, Journal entry, Higher education

[1] Corresponding author at: Department of Business, Faculty of Economics and Business, University of Barcelona, Av. Diagonal 690, 08034 Barcelona, Spain. E-mail address: bllacay@ub.edu. ORCID ID 0000-0002-5238-1648.

Can LLMs Help Students Learn Accounting? A Performance Analysis Across Models and Prompts

Abstract

In recent years, the emergence of large language models (LLMs) has transformed the landscape of higher education. Among their most promising applications is the potential to personalise student learning and serve as virtual tutors. This paper investigates whether current versions of the leading LLMs — ChatGPT, Gemini, CoPilot, and Claude — can function as effective learning assistants for accounting students. To this end, we evaluate the accuracy of these models in responding to a set of theoretical and practical multiple-choice questions taken from real university exams. We also examine whether simple prompting strategies, accessible to any student, can improve model performance. Our findings show that while LLMs achieve high accuracy on theoretical questions, their performance declines when solving practical tasks. In particular, accuracy remains insufficient in questions involving the preparation of financial statements. However, when testing more advanced models — ChatGPT-o3 and Claude Opus 4 — we observe a significant leap in performance, suggesting a paradigm shift. These results indicate that, in the near future, accounting students and learners in related fields may be able to reliably use LLMs to support and enhance their independent study.

Keywords: LLM, Generative AI, Accounting, Financial Statements, Journal entry, Higher education

1. Introduction

In recent years, the introduction of artificial intelligence (AI) in the education sector has transformed traditional teaching and learning methods (Chen, 2023). Among these developments, the appearance of Large Language Models (LLMs), such as ChatGPT, has shown great potential to revolutionise the educational setting. LLMs function as conversational agents that provide human-like responses to natural-language requests (Yan, et al., 2023). They can support students' learning by explaining concepts, providing feedback, answering questions or personalising course activities to their individual needs ((Dong, Bai, Xu, & Zhou, 2024), (Chugh, et al., 2025)).

The advantages offered by LLMs have attracted growing interest from educators and institutions, who have started discussing and experimenting how to incorporate generative AI tools into their teaching practices to enhance students' learning experiences (Alier, Pereira, García-Peñalvo, Casañ, & Cabré, 2025). However, the introduction of AI tools in education does not come without controversy. The responses generated by LLMs are often overly general, occasionally inaccurate or biased, and can contain errors – especially when applied to reasoning-intensive tasks ((Rahman & Watanobe, 2023), (Krupp, et al., 2024)). Excessive reliance on these tools can hinder

students from developing their critical thinking skills (Sok & Heng, 2023), not to mention integrity concerns such as when students submit AI-generated content as their own work or engage in dishonest practices (Spirgi & Seufert, 2025).

The use of generative AI – especially ChatGPT - as a learning-assisting tool has been evaluated in a variety of academic areas, with most studies focusing on medicine ((Totlis, et al., 2023), (Lower, Seth, Lim, & Seth, 2023)), foreign language learning ((Kohnke, Moorhouse, & Zou, 2023), (Guo & Wang, 2023)) or computer programming (Chugh, et al., 2025). To a lesser extent, LLMs have also been applied in fields such as physics, mathematics, or tourism (Dong, Bai, Xu, & Zhou, 2024). Overall, generative AI tools have shown promising results in promoting active learning, improving learners' engagement and enhancing students' attitude towards course activities.

This paper explores the potential of LLMs to serve as learning assistants in a higher education area that has been far less explored: accounting. This domain presents unique challenges to LLMs, as accounting tasks typically require the interpretation of textual descriptions alongside mathematical computations and reasonings – a task where LLMs still demonstrate limitations. Moreover, accounting involves the use of specialised terminology, rules and domain-specific knowledge – and where LLMs, which have been trained in a general corpus of knowledge can struggle to provide correct and accurate answers ((Garg, Mehndiratta, & Vidushi, 2024), (Biancone & Chmet, 2024)).

Although the use of generative AI tools in accounting education is less widespread than in other disciplines, there is evidence that students are increasingly turning to ChatGPT as a 'virtual tutor' that can answer their questions and explain concepts in a simpler way ((Sundkvist & Kulset, 2024), (Maruszewska, Ziemba, Grabara, & Renik, 2024)). More importantly, learners tend to trust the responses provided by these tools without questioning their validity, which can negatively impact learning – particularly in those tasks where LLMs are known to be error-prone (Sundkvist & Kulset, 2024).

For accounting teachers and students to make the most of the undeniable potential of LLMs as learning assistants, it is crucial to evaluate their accuracy and effectiveness in solving both conceptual questions and practical exercises that students are commonly asked to complete. This paper aims to contribute to this objective by assessing the performance of LLMs on accounting tasks that represent core competences that students are expected to master in introductory accounting courses.

In January 2023, Wood et al. (2023) conducted a large-scale study comparing the performance of ChatGPT-3.5 to that of accounting students in responding to a broad set of accounting assessment questions from 186 institutions. The results showed that students outperformed ChatGPT and that the accuracy of ChatGPT was lower in those tasks that involved mathematical reasoning. The subsequent release of ChatGPT-4 brought noticeable improvements: Abeysekera (2024) demonstrated that ChatGPT-4 improves the scores of ChatGPT-3.5 when answering a small set of multiple-choice questions from introductory and advanced accounting courses; Cheng et al. (2024) showed that ChatGPT-4 performs better than ChatGPT-3.5 when dealing with accounting cases, but it still shows shortcomings in those exercises that require more advanced reasoning, such as preparing financial statements and making journal entries.

The objective of this paper is to update and extend this research. As more powerful LLMs continue to be released, it is crucial to re-evaluate their capabilities in addressing accounting-related assessments. In particular, newer models that have been optimised for reasoning and numerical computation – such as ChatGPT-4o – may offer significant

improvements in areas where earlier models fell short. For this reason, in this paper we focus specifically on accounting questions and exercises involving financial statement preparation and journal entries, which have consistently been identified as challenging for LLMs and are among the most critical learning objectives in introductory accounting courses.

Moreover, we will extend the available literature along three directions:

- All previous studies that evaluate LLMs for solving accounting learning questions have focused solely on ChatGPT. In this paper we compare the performance of ChatGPT to the other most widely used LLMs (DataStudios, 2025), including Gemini (Google), Microsoft CoPilot and Claude (Anthropic). This broader evaluation reflects the diversity of tools available to students today.
- For those tasks where the LLM performance is not good, we assess whether simple prompt engineering strategies – such as few-shot learning, role-based prompting and chain-of-thought – can improve the accuracy of LLM responses, thereby enhancing their effectiveness as learning assistants.
- We evaluate whether the newest models, which are not yet free but are more optimised for logical reasoning—like ChatGPT-o3 and Claude Opus 4—achieve better performance, to get an idea of the evolution we can expect in the near future.

By examining these questions, we aim to provide educators and students with evidence-based insights into how generative AI can be integrated into accounting education in an effective manner.

The remainder of the paper is organised as follows. In section 2 we review the related literature on the use of large language models in accounting education. In section 3 we describe the methodology used to evaluate the different LLMs. Section 4 presents the results of the evaluation and section 5 then concludes our analysis.

2. LLMs in Accounting Education

Since the beginning of the 21st century, the application of AI and machine learning to education has been shown to enhance the quality of pedagogical services, the engagement of students and the learning outcomes ((Okagbue, et al., 2023), (Pratama, Sampelolo, & Lura, 2023)). Among various AI methods and applications, the emergence of generative AI has revolutionised the educational landscape. A key strength of generative AI tools is their capability to facilitate personalised learning - by enabling the customisation of content and resources to address the needs and preferences of learners ((Bozic & Poola, 2023), (Alier, Pereira, García-Peñalvo, Casañ, & Cabré, 2025)) - and active learning – by creating environments where students can explore a topic and find answers to their questions (Rasul, et al., 2023). Moreover, ChatGPT and similar tools have the potential to be used as virtual assistants that support students in their learning needs – answering questions, summarising contents, explaining difficult concepts, generating new ideas, etc. – and enhance their engagement with the learning process ((Faisal, 2024), (Kanont, et al., 2024)).

In recent times, the use of generative AI in higher education has been implemented across such diverse fields as medicine, computer science or language learning. For example, Totlis et al. (2023) explore the capability of ChatGPT to provide anatomical

descriptions in the context of medicine education. Breeding et al. (2024) evaluate the reliability and comprehensiveness of the information produced by ChatGPT on the diagnosis and management of different surgical conditions. Al Hakim, Paiman and Rahman (2024) build a custom AI chatbot to address the questions of engineering students on occupational health and safety issues. Ahmed and Hasnine (2023) assess whether generative AI chatbots can help computer science students to improve their knowledge of computers network and their self-efficacy. Also, various papers have study the possibilities of LLMs to assist students in programming education, either using generic LLM systems or by implementing specific LLM-based assistants ((Chugh, et al., 2025), (Feng, Liu, & Ghosal, 2024), (Savelka, Agarwal, Bogart, Song, & Sakr, 2023)). Wardat et al. (2023) evaluate the potential of ChatGPT to teach and learn mathematics, and Li et al. (2024) develop a ChatGPT-based assistant for STEM education. LLMs have also been shown to be useful assistants for students of foreign languages, as they can engage in realistic dialogues and help learners freely practice their language abilities ((Belda-Medina & Calvo-Ferrer, 2022), (Young & Shishido, 2023), (Slamet, 2024), (Piatykop, Yeva, Pronina, & Balalayeva, 2025)).

The various initiatives exploring the utility of generative AI tools as higher education assistants have shown that these tools provide numerous benefits for students. They can offer instant feedback (Gökçearslan, Tosun, & Erdemir, 2024) and enhance learners' comprehension and engagement in difficult subjects ((Kuhail, Alturki, Alramlawi, & Alhejori, 2023), (Al Hakim, Paiman, & Rahman, 2024)). By adapting to the individual learning pace of each student, the learning experience becomes more pleasant and enjoyable, and the productivity and self-efficacy of learners increase ((Elkhodr, Gide, Wu, & Darwish, 2023), (Elbanna & Armstrong, 2024)).

However, not all aspects are advantageous. The responses provided by ChatGPT and similar tools can be inaccurate, biased or plainly incorrect, and students often are not able to critically evaluate the reliability of chatbot outputs and determine whether these are right or not (Sundkvist & Kulset, 2024). Researchers have expressed concerns about the risk of students becoming overly reliant on generative AI, which could prevent them from developing their creativity and critical thinking skills, and encourage their laziness (Mohamed, 2023). Moreover, students might be tempted to use these tools unethically, presenting work done by AI models as their own and thus threatening academic integrity (Dosumu, Porumb, Stafford, & Zimmer, 2025).

Specifically, in accounting education, generative AI models face particular challenges because accounting tasks require mathematical calculations and logical reasoning, areas where LLMs often struggle to provide correct answers. Additionally, accounting involves specialised terminology and precise rules, and generative AI tools have not been sufficiently trained in this particular area of knowledge. This poses an additional difficulty for LLMs in delivering accurate responses (Garg, Mehndiratta, & Vidushi, 2024).

To evaluate the capability of LLMs to solve accounting problems, some authors have tested the performance of ChatGPT on professional accounting examinations ((de Freitas, Sallaberry, Silva, & da Rosa, 2024), (Eulerich, Sanatizadeh, Vakilzadeh, & Wood, 2024)). The results have shown that ChatGPT-4 performs significantly better than its predecessor ChatGPT-3.5 – which was not able to pass the exams, – and the scores further improve when the LLM is trained with a few examples or given the opportunity to use a calculator or other resources (Eulerich, Sanatizadeh, Vakilzadeh, & Wood, 2024). These results are consistent with the performance observed in proficiency exams of other professions ((Bommarito, Bommarito, Katz, & Katz, 2023), (Katz, Bommarito, Gao, & Arredondo, 2024)).

Closely related to the current study, some papers have evaluated the performance of LLMs on real university accounting assessments. At the beginning of 2023, Wood compared the accuracy of ChatGPT-3.5 and students using a large test set of accounting assessments provided by 327 coauthors from 186 institutions (Wood, 2023). The diversity of coauthors resulted in a large variety of questions and topics covered. The evaluation showed that, on average, ChatGPT performed worse than the students (47% score vs 77% obtained by students), but its score improved when dealing with questions related to AIS and auditing, possibly because these types of questions do not require mathematical calculations or logical reasoning. The study also showed that ChatGPT performed better in true/false and multiple-choice questions than in open-ended questions. Cheng et al. (2024) studied the performance differences between ChatGPT-3.5 and ChatGPT-4 when answering questions related to eight accounting cases. The results showed that, in general, ChatGPT-4 outperformed ChatGPT-3.5, except in questions consisting purely of calculations. However, their accuracy was still modest and lower than that of students, implying that those versions of ChatGPT were not on the point to be reliable assistants for accounting learners. Regarding the topic of questions, both ChatGPT models performed better on cases requiring explanations, application of rules, and ethical evaluation, whereas their performance was worse on cases involving journal entries and financial statement preparation. Abeysekera (2024) compared the performance of ChatGPT-3.5 and ChatGPT-4 using ten multiple-choice questions from an introductory course in accounting and ten from an advanced course. Half of the questions were narrative, while the other half where numerical-based. The author concluded that ChatGPT might be useful for competent learners to validate their answers, but not for novice learners, as it does not scaffold answers adequately, thus preventing them from developing their own competences and possibly hindering their true learning process. Though ChatGPT-4 achieved higher scores than ChatGPT-3.5, neither answered all questions correctly, and they particularly struggled with those questions that were most technical and complex.

The performance of LLMs is continuously improving; thus, we propose here to update previous studies by evaluating the performance of the current version of several LLMs – and not only ChatGPT as in previous studies – using a set of accounting questions and exercises. Building on previous research, which has shown that ChatGPT struggles particularly with those questions that are most technical and require calculations or logical reasoning, we focus our study on those types of accounting questions that have shown to be most challenging for previous versions of ChatGPT: preparation of financial statements and creation of journal entries. Also building on more general work that has shown that better prompts can significantly improve the quality of responses, we aim to contribute to the literature by examining whether simple prompting strategies that can be easily used by any student can help enhance the accuracy of responses.

3. Methodology

3.1. Assessment dataset and procedure

The set of theoretical and practical questions used to evaluate the LLMs' performance was drawn from introductory accounting courses (Accounting I) taught at our university. We focused on these introductory courses because the syllabus focuses precisely on

those tasks that previous studies have shown to be particularly challenging for LLMs (Cheng, et al., 2024): preparation of financial statements and recording of journal entries.

The questions were taken from actual exams prepared by one of the co-authors. The exams of Accounting I contain both a theoretical and a practical part, where all items are formulated as multiple-choice questions. To closely simulate what is expected of students and to assess the extent to which LLMs can be helpful in their self-study, we selected a set of questions that included both theoretical and practical components and covered the entire syllabus that students are expected to master. Each question had four possible answers, with only one correct option.

The evaluation set consists of 75 multiple-choice questions, divided into three parts:

- 25 theoretical questions that cover fundamental accounting principles, conceptual frameworks, and regulatory knowledge
- 25 practical questions involving the record of journal entries
- 25 practical questions related to the preparation of financial statements (balance sheet and income statement)

This dataset provides a comprehensive evaluation framework that ranges from knowledge recall to practical application and numerical computation.

Each set of questions was submitted to the models using their respective chat interfaces. All interactions were conducted through the publicly accessible versions of each LLM, using the default settings available to users as of June 2025. To reduce the impact of response randomness and improve the reliability of the results, each questionnaire was entered three times in independent, memoryless chat sessions to ensure that no contextual information was carried over from one interaction to the next. For each model and question set, we computed the average accuracy across the three runs, as well as the standard deviation, in order to assess both overall performance and response consistency.

All questions were presented in Spanish, preserving their original wording and structure as used in actual student assessments.

3.2. Experiments

Experiment 1: Which LLM is more accurate?

We will compare the performance of the four most widely used LLMs (DataStudios, 2025):

- *ChatGPT*: Launched by OpenAI in 2022, and building on the GPT (Generative Pre-trained Transformer) architecture, it has become the most popular generative AI tool. It has a great versatility in language-related tasks, excelling in answering questions, generating content or summarising texts. However, the latest version (ChatGPT-4o) has also shown a good performance in reasoning and logical tasks.
- *Gemini*: Developed by Google and formerly known as Bard, Gemini distinguishes itself from other LLM in its ability to handle multimodal input, such as text, images, audio or videos. This allows to provide a more comprehensive and versatile experience as it is able to interpret a wide range of information.

- *Microsoft CoPilot*: Developed by Microsoft using the GPT architecture, CoPilot was born as an assistant to enhance the productivity of Microsoft applications users. So, it is deeply integrated within the Microsoft ecosystem, and is intended to help in tasks such as writing content, analysing data or preparing presentations.
- *Claude*: Developed by Anthropic, Claude distinguishes itself by building on "constitutional AI", a set of ethical and safety principles that guide its responses. This reduces the biases of outputs and increases the ability to deal with scenarios that require a higher degree of ethical awareness. At the same time, Claude is known to perform well in reasoning and long-context comprehension tasks.

To replicate the typical conditions under which students are likely to interact with these LLMs, we will: (1) use the most recent free version of each LLM (ChatGPT-4o, Gemini 2.5 Flash, baseline Microsoft Copilot and Claude 4 Sonnet); (2) interact with each LLM directly through its standard user interface (without using the API); and (3) retain the default parameter settings.

Experiment 2: How to formulate the prompt?

It has been repeatedly shown that the quality of the prompt can significantly enhance the accuracy of the response provided by LLMs ((Ekin, 2023), (Knoth, Tolzin, Janson, & Leimeister, 2024)). So, being able to effectively design the input statements used to dialogue with LLMs can be a crucial ability to achieve better results in using these tools for learning. We will test whether simple prompting strategies can increase the accuracy of the different LLMs. Although more sophisticated prompt engineering methodologies are available (Chen, Zhang, Langrené, & Zhu, 2025), our focus here is on basic techniques that could, with almost no effort from students, increase the reliability and power of LLMs as learning assistants.

Specifically, we will test the following prompting strategies:

- *Zero-shot prompting*: We will formulate the question or exercise to be answered without any examples, context, or specific indications on how to answer it. This approach relies on the general knowledge LLMs have been trained on, which allows them to answer to new tasks without additional training (Korzynski, Mazurek, Krzypkowska, & Kurasinski, 2023).
- *Few-shot prompting*: In this case, before formulating the assessment question to be answered, we will provide LLMs with a small set of examples, together with the desired answer. Although few-shot usually outperforms one-shot prompting (Korzynski, Mazurek, Krzypkowska, & Kurasinski, 2023), in some cases zero-shot prompting has turned to be more effective (Chen, Zhang, Langrené, & Zhu, 2025).

  For example, for the questionnaire on journal entries we will use the following prompt:

  > *I want you to solve multiple-choice exercises on journal entries.*
  >
  > *Below, I will give you several example exercises with their correct answers.*
  >
  > *Please learn the format and the logic used to solve them, and then solve the new exercises I will give you next in the same way.*
  >
  > *{EXAMPLE 1}*

…

{EXAMPLE n]

- *Role prompting*: Assigning the LLM the role of an expert in a given field has shown to increase the specificity of its response, probably because this simple instruction allows the model to focus on a particular area within the vast corpus of knowledge it has been trained on (Leung, 2024).

  We will use the following prompt:

  > *Assume the role of a university professor of financial accounting with extensive experience in introductory courses. Your task is to help students understand and correctly solve exercises involving journal entries—following the basic principles of double-entry bookkeeping—as well as exercises involving the preparation of the balance sheet and income statement.*
  >
  > *For journal entry exercises, briefly explain the reasoning behind each entry and identify the accounts involved. For financial statement exercises, correctly group the accounts by type and perform the necessary calculations to prepare the complete statement.*
  >
  > *When responding, use a clear format suitable for first-year students. If there is any ambiguity in the question, explain the possible interpretations and choose the most reasonable one based on the typical context of an introductory course.*

- *Chain-of-thought*: The chain-of-thought technique consists in breaking down complicated tasks into smaller, intermediate steps. This step-by-step reasoning helps increase the accuracy of LLMs, especially for complex tasks (Chen, Zhang, Langrené, & Zhu, 2025). An extremely easy way to implement this strategy is to add the sentence “Think step by step” at the end of the prompt. This simple strategy has shown to increase the accuracy of GPT-3 in mathematical tasks, although it has turned to be less effective on newer models as GPT-4 (Gao, 2023), so we will use a more elaborate prompt, such as the following one for the case of journal entries:

  > *Analyse and solve step by step:*
  >
  > 1. *Identify the accounts involved in the transaction.*
  > 2. *Classify each account as an asset, liability, equity, revenue, or expense.*
  > 3. *Determine whether each account increases or decreases in this transaction.*
  > 4. *Indicate whether each account is debited or credited, justifying your choice.*
  > 5. *Record the journal entry in the general journal, ensuring that debits and credits are balanced.*
  > 6. *Review whether the entry is balanced and whether the account classifications are correct.*

Experiment 3: A glimpse into the immediate future

In the previous experiments, we used the free versions of the various models, as these are the ones students are most likely to access. However, OpenAI and other developers offer more advanced models through their paid plans—models that are better equipped to handle complex reasoning tasks. Since it is only a matter of time before these

advanced models become widely accessible, in this final experiment we set out to test two specific versions of the models that performed best in our earlier evaluations: ChatGPT and Claude.

In this experiment, we re-evaluated the journal entry and financial statement questionnaires using the following models:

- ChatGPT-o3: A proprietary GPT-4-class model offered through OpenAI's premium plans. It is designed to be more efficient and accurate than previous versions, with improved capabilities in mathematical reasoning, multi-step logic, and complex problem solving. It also demonstrates higher consistency and fewer hallucinations across a broad range of tasks.
- Claude Opus 4: The most advanced model currently offered by Anthropic. It features a significantly larger context window (up to 200,000 tokens), allowing it to handle extended inputs and outputs effectively. Claude Opus is optimised for tasks that require careful interpretation of structured data, making it particularly suitable for financial and analytical domains.

With this setup, we aim to provide a glimpse into what the near future may hold in terms of the use of LLMs in accounting education.

4. Results and discussion

4.1. Experiment 1: Which LLM is the most accurate?

In this first experiment, we introduced the questions exactly as they appeared in the sample exams we selected, without any prior introduction or explanation (zero-shot learning). *Table 1* summarises the performance of the four evaluated LLMs—ChatGPT, Gemini, CoPilot, and Claude—across three types of questionnaires: accounting theory questions, practical questions about journal entries, and questions about the preparation of financial statements. The questions (25 for each type) were drawn from actual university exams.

To account for variability in LLM responses, each questionnaire was run independently three times. *Table 1* reports both the mean accuracy and the standard deviation (in parentheses). The standard deviation reflects the stability of each model across iterations, highlighting differences not only in correctness but also in consistency. As described below, results vary significantly depending on the type of questionnaire. Overall, the models perform well on theoretical questions, while performance on the more applied tasks—journal entries and financial statement preparation—is somewhat lower and more variable.

Accuracy	ChatGPT	Gemini	CoPilot	Claude
Theory	96% (0.0%)	95% (1.9%)	96% (0.0%)	89% (1.9%)
Journal entries	85% (1.9%)	91% (1.9%)	85% (1.9%)	95% (1.9%)
Financial statements	69% (3.8%)	49% (3.8%)	59% (1.9%)	71% (7.5%)

Table 1 – Mean and standard deviation of accuracy for each questionnaire section

THEORY QUESTIONNAIRE

All models demonstrate a high level of accuracy when answering theoretical questions on basic accounting topics, suggesting that the fundamental theoretical concepts of accounting are well represented in the training data of these models. Moreover, these results are highly consistent across runs, with the different LLMs achieving nearly identical scores throughout the iterations (and in the case of ChatGPT and CoPilot, the results are exactly the same).

All models correctly answer the vast majority of questions, and the few mistakes they make are concentrated in just a handful of specific items. These are not purely theoretical questions but ones that require more conceptual reasoning or applied knowledge that is specific to the Spanish regulatory framework (*Plan General de Contabilidad*). For example, identifying the correct liability account that represents a bill of exchange in the context of purchasing fixed assets.

These types of questions are more difficult for LLMs because they require not only memorised facts or definitions, but also the ability to apply rules from a specific national framework, which may not be as well represented in their training data or may demand contextual reasoning that exceeds simple pattern recognition.

In this theoretical questionnaire, ChatGPT and CoPilot emerge as the most reliable and consistent models. They not only achieve higher accuracy but also maintain stable performance across runs. Gemini follows closely behind, while Claude makes more errors and also shows the highest variability between iterations. Notably, Claude consistently fails a question about the types of VAT in Spain, even though it involves no conceptual difficulty. While it is true that answering such a question requires up-to-date knowledge of the Spanish tax system, the model is in fact able to retrieve the correct information about the three current VAT rates but still fails to identify the correct answer.

Beyond the number of correct answers, it is also important to assess the explanatory structure of the responses provided by each model. This aspect is relevant because we are not only evaluating LLMs as test-solving agents but also as potential learning assistants. If these tools are to be used in educational settings, their ability to communicate reasoning clearly and effectively is essential. A model that provides the correct answer but fails to explain why it is correct—or why the alternatives are incorrect—may be of limited use to students who are trying to learn.

From this perspective, ChatGPT stands out for its highly structured responses. Even without being prompted to do so, it typically begins by stating the correct answer, then explains why it is correct, and finally offers brief justifications for why the remaining options are incorrect. This format is especially well suited for learners, as it promotes a deeper understanding of the underlying concepts and helps reinforce the correct reasoning process.

Gemini's responses, in contrast, are less structured. It tends to provide a more general explanation of the topic addressed in the question, from which it then infers the correct answer. Often, the answer is embedded within the explanation and not clearly stated at the outset, which may require additional effort from the learner to identify the correct response.

CoPilot follows a structure very similar to that of ChatGPT. It begins with the correct answer, followed by an explanation of why that answer is correct, and then offers shorter comments on why the other options are incorrect.

Claude also adopts a similar structure, but it tends to offer more detailed explanations for why the selected answer is correct. While this level of elaboration can be helpful, it may also introduce unnecessary complexity if not well targeted to the learner's level.

In sum, the clarity and structure of explanations—especially the ability to contrast correct and incorrect options—can significantly enhance the educational value of these models. From this standpoint, ChatGPT and CoPilot currently offer the most pedagogically effective responses.

JOURNAL ENTRY QUESTIONNAIRE

In this more practical section, the average accuracy is somehwat lower, ranging from 85% (achieved by ChatGPT and CoPilot) to 95% (achieved by Claude). This confirms that the practical application of theoretical knowledge presents a greater challenge for LLMs. Interestingly, and in contrast to the theory section, the performance rankings shift significantly: Claude—previously the lowest-performing model in theoretical questions—now obtains the highest accuracy, with an average of 95%. Meanwhile, ChatGPT and CoPilot, which had led in the theory section, drop to an average of 85%. Gemini remains in an intermediate position with an average accuracy of 91%.

These results may suggest that ChatGPT and CoPilot were trained with greater exposure to theoretical content, whereas Claude may have been exposed to more practical accounting examples, making it better suited to the applied aspects of accounting tasks. This supports the idea that there is no straightforward correlation between theoretical proficiency and the ability to apply that knowledge in practice.

The standard deviations across runs indicate that model outputs are fairly consistent, which is desirable given that these tasks typically have a deterministic correct answer based on formal accounting rules. As in the theoretical section, the errors are concentrated in a small subset of questions. These tend to involve more complex types of journal entries, such as those combining multiple operations (e.g., sales or purchases involving promissory notes), intricate VAT handling (models occasionally confuse input and output VAT), or more advanced procedures like year-end closing or adjustments. These entries require not only rule application but also contextual understanding and higher-order reasoning.

As in the theory section, response structure and pedagogical value varies significantly across models. ChatGPT provides well-organised answers: it identifies the type of journal entry, explains the necessary steps to complete it, determines the correct response, and then justifies why the other options are incorrect. Gemini typically begins by stating the correct answer, then describes how this type of journal entry works, performs the relevant calculations, and concludes by analysing why the alternative options are incorrect. Its explanations are generally more detailed than ChatGPT’s.

CoPilot is far more concise: it provides the correct answer and offers a brief explanation, typically mentioning the relevant accounts or necessary calculations. However, it does not usually explain why the other options are incorrect. Claude takes a more balanced approach, walking through the required calculations, identifying the accounts involved, and offering good explanations of why the incorrect options are wrong.

Beyond simply identifying the correct entry, it is crucial that LLMs also explain how the answer is derived. This aspect is especially important in an educational context, where students benefit from learning the reasoning process behind each entry. In this regard, ChatGPT and Claude appear to be the most pedagogically effective: they strike a balance between clarity and detail, offering sufficient explanation for learners to understand both how to solve similar problems and why incorrect alternatives should be discarded.

FINANCIAL STATEMENT QUESTIONNAIRE

In this third section of the questionnaire, the performance of all models drops significantly. Claude and ChatGPT achieve the highest scores, but their accuracy levels—71% and 69%, respectively—are still far lower than in the previous questionnaires. Gemini performs worst in this section, with an average accuracy of just 49%. Although Claude outperforms the other models, it also shows the highest variability across runs, indicating inconsistent output.

Once again, questions involving more complex calculations prove to be the most challenging. In particular, items that require multiple computations or involve integrating several line items from the financial statements are especially problematic. In such cases, any intermediate error leads to an incorrect final result. For example, calculating working capital is especially difficult, as it requires accurately identifying and summing both current assets and current liabilities. Similarly, computing equity is often problematic because models frequently omit the profit or loss for the period derived from the income statement.

Moreover, when a model's initial answer does not match any of the options provided, some LLMs attempt to revise their calculations by selectively including or excluding certain line items in a kind of trial-and-error approach aimed at reverse-engineering a plausible result. While this strategy may occasionally produce a correct answer, it is pedagogically unhelpful when the goal is to use generative AI as a tool to help students understand how to solve such problems correctly.

Regarding response structure, ChatGPT provides well-organised answers, explaining the calculations step by step. Gemini tends to engage in overly lengthy reasoning when it struggles to find the correct answer, often trying different combinations to arrive at a solution. CoPilot offers the most concise and minimal explanations, whereas Claude typically explains the steps clearly and methodically, guiding the user through the process to reach the correct result.

OVERALL ASSESSMENT

As question complexity increases, the accuracy of the models declines. This is particularly evident in the financial statement section, which demands not only computational ability but also a sound understanding of the structure of balance sheets and income statements, as well as the ability to manipulate multiple figures across these documents. Among the four models evaluated, Claude appears best suited for handling complex reasoning tasks, although it also exhibits more inconsistency across runs. ChatGPT follows closely in practical tasks and shows outstanding performance in theoretical questions. Gemini performs strongly in the theoretical section but its reliability drops sharply when applied to tasks requiring more advanced reasoning. CoPilot also

shows a decline in accuracy when solving applied problems, although it remains quite reliable on theory questions and basic journal entries.

The overall low level of accuracy in this type of task raises questions about the current applicability of generative AI models as learning assistants for accounting students. In the next experiment, we investigate whether model performance can be improved using simple prompting techniques accessible to any typical LLM user.

4.2. Experiment 2: How to formulate the prompt?

For this experiment, we focused on the journal entry and financial statement questionnaires, as these are the areas where all LLMs encountered the most difficulties and where there is the greatest potential for improvement. These tasks require not only factual knowledge but also the ability to perform structured reasoning and handle multi-step calculations—capabilities that current models still struggle with to varying degrees.

PROMPT STRATEGY 1: FEW-SHOT LEARNING

To explore whether model performance could be improved with guided examples, we implemented a *few-shot learning* approach. In contrast to the *zero-shot setting*—where models were given the test questions without any prior context—few-shot learning involves providing the model with a small number of examples that illustrate the type of task being asked. These examples serve as contextual cues, helping the model infer the correct structure, logic, and approach needed to solve subsequent questions.

To re-evaluate the journal entry questionnaire, we preceded the test questions with six worked examples, each paired with its correct solution. These examples included cases that had proven particularly challenging in the zero-shot setting, such as year-end adjustments and sales with discounts. Similarly, for the financial statement tasks, we provided six example exercises similar in format and structure to those that followed in the test. While the examples targeted different line items, they reflected the same types of reasoning and calculations required in the evaluation tasks.

The results of this second experiment are mixed (see *Table 2*). For the *journal entry questionnaire*, some models responded positively and improved their performance based on the examples provided. Claude achieved the best results, reaching 100% accuracy (up from 95% in the previous experiment) and delivering perfectly consistent answers across all iterations. ChatGPT also showed a modest improvement, increasing its average accuracy from 85% to 88% with the inclusion of prior examples.

In contrast, CoPilot and Gemini did not appear to benefit from the few-shot approach. CoPilot's accuracy remained unchanged, and Gemini's performance declined slightly. These findings suggest that Claude and ChatGPT are able to generalise from a small number of structured examples, while CoPilot and Gemini do not appear to leverage this contextual information effectively. This divergence likely reflects differences in the internal architecture and training of the models, particularly in how they utilise contextual input. Some models may have a stronger foundational understanding and are therefore better positioned to use the few-shot context to improve their output. Others, possibly trained on less relevant data or with different learning strategies, may not gain much from a limited set of examples.

Accuracy	ChatGPT	Gemini	CoPilot	Claude
Journal entries				
Zero-shot	85% (1.9%)	91% (1.9%)	85% (1.9%)	95% (1.9%)
Few-shot	88% (3.3%)	89% (1.9%)	85% (1.9%)	100% (0.0%)
Financial statements				
Zero-shot	69% (3.8%)	49% (3.8%)	59% (1.9%)	71% (7.5%)
Few-shot	56% (3.3%)	53% (1.9%)	67% (13.2%)	64% (5.7%)

Table 2 – Mean and standard deviation of accuracy after using a few-shot learning prompt strategy

However, when applying the same prompting strategy to the *financial statement questionnaire*—focused on the balance sheet and income statement—a completely different picture emerges (see *Table 2*). Claude and ChatGPT, which had previously shown the greatest ability to learn from examples, both experienced a drop in accuracy. In this case, having access to prior examples did not help; in fact, it negatively affected their performance.

Conversely, CoPilot and Gemini—previously the weaker performers—were now able to improve their accuracy. Notably, CoPilot showed the largest improvement in average accuracy. However, this came at the cost of high inconsistency: it exhibited a standard deviation of 13%, indicating that its responses varied significantly across iterations. This erratic behaviour undermines its reliability and renders it unsuitable as a learning assistant for students.

These contrasting results suggest that few-shot learning is not equally effective across all task types. In more repetitive and structured tasks, such as journal entries, few-shot prompting can provide useful templates and allow models to generalise from context. However, tasks that involve producing financial statements require integrating multiple line items and performing arithmetic operations, making it much harder for models to learn from just a few isolated examples. In fact, the examples may introduce noise rather than support generalisation, ultimately reducing the quality of the model's output.

PROMPT STRATEGY 2: ROLE-BASED PROMPTING

In this case, the LLMs were instructed to assume the role of a university professor of financial accounting with extensive experience in introductory courses. This role was selected based on the assumption that it would help the models better showcase their potential as learning assistants. However, this role-based prompting strategy proved entirely ineffective—and in some cases, even counterproductive (see *Table 3*).

Accuracy	ChatGPT	Gemini	CoPilot	Claude

Journal entries				
Zero-shot	85% (1.9%)	91% (1.9%)	85% (1.9%)	95% (1.9%)
Role-based	85% (1.9%)	92% (5.7%)	84% (3.3%)	95% (3.8%)
Financial statements				
Zero-shot	69% (3.8%)	49% (3.8%)	59% (1.9%)	71% (7.5%)
Role-based	57% (1.9%)	48% (3.3%)	57% (5.0%)	68% (0.0%)

Table 3 – Mean and standard deviation of accuracy after using a role-based prompt strategy

In the journal entry questionnaire, the average accuracy across models remained virtually unchanged, while variability across iterations slightly increased, especially for Gemini. This indicates that role-based prompting offered no advantage over the baseline direct-prompting strategy.

In the financial statement questionnaire, some models even performed worse. Notably, ChatGPT's accuracy dropped from 69% to 57%, and Claude's from 71% to 68%. While the consistency of their responses improved, this gain did not compensate for the overall decline in accuracy.

Thus, the role-based prompting strategy did not yield any improvements. This may suggest that the models are already well-calibrated for these types of tasks and do not benefit from assuming an expert teaching persona. Moreover, in the context of financial statement preparation, adopting a "professor" role appeared to interfere with the models' arithmetic reasoning, leading to less accurate computations.

These findings suggest that, despite being widely reported as a strategy to enhance model performance, role-based prompting is not universally effective. For certain technical tasks, this added contextual framing may actually hinder rather than help, reducing model performance and reliability.

PROMPT STRATEGY 3: CHAIN-OF-THOUGHT REASONING

In this experiment, we tested a chain-of-thought prompting strategy, which instructs models to reason step by step. In principle, this approach should enhance model accuracy, especially in accounting tasks that involve multiple calculations—such as the preparation of financial statements. These tasks appear well-suited to a prompt that explicitly requires intermediate steps, including drafting the balance sheet and income statement, reviewing transactions, and performing final checks. However, the results were both disappointing and paradoxical (see *Table 4*).

In the journal entry questionnaire, model performance remained virtually unchanged. The chain-of-thought prompt did not improve accuracy, but neither did it degrade it—possibly because these tasks do not involve many intermediate reasoning steps. In contrast, the performance on the financial statement questionnaire deteriorated significantly. Step-by-step reasoning not only failed to improve accuracy (with the modest exception of Claude, which improved from 71% to 75%) but in some cases led to markedly worse outcomes. ChatGPT's accuracy dropped from 69% to 57%, while Gemini's fell from 49% to 37%, with an accompanying increase in response variability.

Accuracy	ChatGPT	Gemini	CoPilot	Claude
Journal entries				
Zero-shot	85% (1.9%)	91% (1.9%)	85% (1.9%)	95% (1.9%)
Chain-of-thought	87% (3.8%)	92% (3.8%)	85% (1.9%)	95% (1.9%)
Financial statements				
Zero-shot	69% (3.8%)	49% (3.8%)	59% (1.9%)	71% (7.5%)
Chain-of-thought	57% (1.9%)	37% (12.4%)	57% (10.0%)	75% (6.8%)

Table 4 – Mean and standard deviation of accuracy after using a chain-of-thought prompt strategy

These results highlight that prompting models to reason step by step does not universally enhance performance. Depending on the model and task type, it can interfere with the model's internal reasoning processes and result in degraded performance. Tasks involving the preparation of financial statements do not necessarily follow a linear logic; instead, they require a holistic understanding of financial reports, domain-specific knowledge, and the ability to perform calculations based on provided data. Moreover, making all intermediate steps explicit can lead to error propagation—mistakes in early calculations may compound across multiple questions. In contrast, zero-shot prompts may allow models to apply heuristics or shortcut strategies that occasionally produce correct answers, even without showing their reasoning.

Additionally, some LLMs may have been optimised for concise responses, and the verbosity required to express intermediate steps could push them outside their optimal response range. Prior research (Lee, 2025) has shown that longer outputs can negatively impact model accuracy. This may partly explain the poor performance observed on financial statement tasks, which typically require a large number of tokens. By contrast, Claude—the best-performing model in this experiment—has both a significantly larger context window and has been explicitly trained to reason transparently through complex tasks (Anthopic, 2025). As a result, prompting it to articulate extended reasoning steps may align more naturally with its design, whereas it could disrupt the performance of other models.

In summary, the effectiveness of chain-of-thought prompting is not universal. It depends on both the task domain and the model architecture. While Claude appears well-suited to this strategy, it may be counterproductive for other models in certain knowledge domains.

4.3. Experiment 3: A glimpse into the immediate future

Given that the free models we evaluated show low accuracy on certain types of accounting questions—and are therefore not yet reliable or useful enough to serve as standalone learning assistants—in this final experiment we sought to assess whether more advanced models available to paying users of OpenAI and Anthropic (specifically, ChatGPT-o3 and Claude Opus 4) demonstrate superior performance when answering

accounting-related questions. This allows us to gain insight into what we can expect from LLMs in the near future, and whether significant improvements in accounting-related tasks are already on the horizon or will require more time to materialise.

Table 5 compares the mean and standard deviation of accuracy obtained by ChatGPT-o3 and Claude Opus 4, relative to the baseline scenario (zero-shot prompting using ChatGPT-o3 and Claude Sonnet 4). The results show that ChatGPT improves in performance with its latest version, particularly on the journal entry and financial statement tasks, whereas the performance of the latest version of Claude in accounting tasks does not improve significantly. The improvement is more pronounced in the financial statement questionnaire, where the baseline performance was lower and thus left more room for improvement. ChatGPT-o3 clearly appears to be the stronger model: it delivers slightly lower accuracy than Claude Opus 4 on the journal entry questions but significantly outperforms it on the balance sheet and income statement tasks, with greater consistency across responses. This suggests that OpenAI's latest model represents a substantial architectural upgrade that enhances its ability to perform advanced reasoning and overcomes limitations present in earlier versions. By contrast, Anthropic's model does not exhibit the same level of paradigm shift as ChatGPT.

Accuracy	ChatGPT-4o	ChatGPT-o3	Claude Sonnet 4	Claude Opus 4
Journal entries	85% (1.9%)	89% (1.9%)	95% (1.9%)	92% (5.7%)
Financial statements	69% (3.8%)	91% (1.9%)	71% (1.9%)	72% (8.6%)

Table 5 – Mean and standard deviation of accuracy of most advanced models

These findings suggest that model architecture may play a more critical role than prompt design in achieving accurate and reliable results, as ChatGPT-o3 was able to deliver superior answers without any additional prompting instructions. This experiment gives reason for optimism: although current LLMs may still fall short as dependable study companions for accounting students, the progress demonstrated by ChatGPT-o3 indicates that, in the near future, general-purpose LLMs may indeed be viable virtual learning assistants.

5. Conclusions

This study aims to evaluate the capabilities of widely used large language models (LLMs)—ChatGPT, Gemini, CoPilot, and Claude—in solving accounting questions and exercises, with the goal of assessing their potential as learning assistants for university students enrolled in introductory accounting courses. We focus on three types of tasks: theoretical questions, journal entries, and the preparation of balance sheets and income statements. These tasks reflect the core content of early accounting curricula and have been identified in prior research as particularly challenging for LLMs. To carry out this evaluation, we compiled a dataset of multiple-choice questions drawn from real examination materials. Our objective was to determine whether current LLMs can serve

as reliable learning aids, and whether prompt engineering strategies or the use of more advanced models can enhance performance.

The models demonstrated strong performance on theoretical questions (with most achieving average accuracies above 95%), somewhat lower performance on journal entry tasks (average accuracy between 80% and 90%), and markedly weaker results on financial statement preparation (average accuracy ranging from 50% to 70%). These findings suggest that the current free versions of these LLMs are not yet capable of consistently assisting accounting students in a reliable manner.

Prompting strategies had a moderate and uneven effect. In the case of journal entries, few-shot prompting and chain-of-thought reasoning improved performance for some models. However, for financial statement tasks, these techniques were occasionally counterproductive. Role-based prompting had minimal impact across both task types.

Overall, our findings indicate that current LLMs are not yet sufficiently reliable to serve as autonomous learning assistants in financial accounting. While they can occasionally produce correct answers, their inconsistency and limited conceptual understanding present risks for independent student use. If these tools are to be integrated into accounting education, domain-specific assistants should be developed and validated by instructors to ensure both accuracy and pedagogical soundness.

Nevertheless, models such as Claude Opus 4 and, in particular, ChatGPT-o3, were able to complete many of these tasks with higher accuracy without requiring specific prompt engineering. This suggests that we may be approaching a turning point. In the near future, students may be able to rely on general-purpose LLMs as dependable virtual tutors, without requiring technical expertise or prompt design skills. This development could unlock new opportunities for accessible, autonomous learning in accounting and other disciplines.

A key limitation of this study lies in the relatively small number of iterations (three per model) and the limited dataset (75 questions in total). For more generalisable and robust results, future work should include a larger and more diverse set of exam questions from multiple academic institutions, along with an increased number of trials per model. This would allow for more precise estimates of average performance and response variability.

Conflict of interest

The authors declare there is no conflict of interest.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Funding statement

No funding was received for this study.

References

Abeysekera, I. (2024). ChatGPT and academia on accounting assessments. *Journal of Open Innovation: Technology, Market, and Complexity, 10*.

Ahmed, M., & Hasnine, M. (2023). Improving essential knowledge and self-efficacy in computers network course: The potential of chatbots. *Procedia Computer Science, 225*, 3929-3937. doi:https://doi.org/10.1016/j.procs.2023.10.388

Al Hakim, V., Paiman, N., & Rahman, M. (2024). Genie-on-demand: A custom AI chatbot for enhancing learning performance, self-efficacy, and technology acceptance in occupational health and safety for engineering education. *Computer Applications in Engineering Education, 32*. doi:https://doi.org/10.1002/cae.22800

Alier, M., Pereira, J., García-Peñalvo, J., Casañ, M., & Cabré, J. (2025). LAMB: An open-source software framework to create artificial intelligence assistants deployed and integrated into learning management systems. *Computer Standards & Interfaces, 92*. doi:https://doi.org/10.1016/j.csi.2024.103940

Anthopic. (24 / February / 2025). *Claude's extended thinking*. Recollit de https://www.anthropic.com/news/visible-extended-thinking

Belda-Medina, J., & Calvo-Ferrer, J. (2022). Using Chatbots as AI Conversational Partners in Language Learning. *Applied Sciences, 12*(17). doi:https://doi.org/10.3390/app12178427

Biancone, P., & Chmet, F. (2024). Role of ChatGPT in the Accounting Field. A A. Bem Machado, M. Sousa, F. Dal Mas, S. Secinaro, & D. Calandra, *Digital Transformation in Higher Education Institutions* (p. 139-153). Cham: Springer.

Bommarito, J., Bommarito, M., Katz, D., & Katz, J. (2023). GPT as Knowledge Worker: A Zero-Shot Evaluation of (AI)CPA Capabilities. *arxiv*. Recollit de https://arxiv.org/abs/2301.04408

Bozic, V., & Poola, I. (2023). *Chat GPT and education.* Preprint. Recollit de https://www.researchgate.net/publication/369926506_Chat_GPT_and_education

Breeding, T., Martinez, B., & Patel, H. (2024). The Utilization of ChatGPT in Reshaping Future Medical Education and Learning Perspectives: A Curse or a Blessing? *The American SurgeonTM*, 1-7.

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patters*. doi:https://doi.org/10.1016/j.patter.2025.101260

Chen, S. (2023). Generative AI, Learning And New Literacies. *Journal of Educational Technology Development and Exchange, 16*(2), 1-19. doi:https://orcid.org/0000-0003-1066-6650

Cheng, X., Dunn, R., Holt, T., Inger, K., Jenkins, J., Jones, J., . . . Wood, D. (2024). Artificial intelligence's capabilities, limitations, and impact on accounting education: Investigating ChatGPT's performance on educational accounting cases. *Issues in Accounting Education, 39*(2), 23-47.

Chugh, R., Turnbull, D., Morshed, A., Sabrina, F., Azad, S., Mamunur, R., . . . Subramani, S. (2025). The Promise and Pitfalls: A Literature Review of Generative Artificial Intelligence as a Learning Assistant in ICT Education. *Computer Applications in Engineering Education, 33*. doi:https://doi.org/10.1002/cae.70002

DataStudios. (May / 2025). *The Most Used AI Chatbots in 2025: Global Usage, Trends, and Platform Comparisons of ChatGPT, Gemini, Copilot, and Claude*. Consultat el 8 / June / 2025, a DataStudios.org: https://www.datastudios.org/post/the-most-used-ai-chatbots-in-2025-global-usage-trends-and-platform-comparisons-of-chatgpt-gemini?utm_source=chatgpt.com

de Freitas, M., Sallaberry, J., Silva, T., & da Rosa, F. (2024). Application of Chatgpt 4.0 for Solving Accounting Problems. *Journal of Globalization Competitiveness and Governability, 18*(2), 49-64. doi:10.58416/GCG.2024.V18.N2.03

Dong, B., Bai, J., Xu, T., & Zhou, Y. (2024). Large Language Models in Education: A Systematic Review. *6th International Conference on Computer Science and Technologies in Education (CSTE)*, (p. 131-134). doi:10.1109/CSTE62025.2024.00031

Dong, M., Stratopoulos, T., & Wang, V. (2024). A scoping review of ChatGPT research in accounting and finance. *International Journal of Accounting Information Systems, 55*. doi:https://doi.org/10.1016/j.accinf.2024.100715

Dosumu, O., Porumb, V., Stafford, A., & Zimmer, A. (2025). In the wake of ChatGPT: early reflections on marking open-book online accounting assessments. *Accounting Education*. doi:10.1080/09639284.2025.2487487

Ekin, S. (2023). *Prompt Engineering For ChatGPT: A Quick Guide To Techniques,Tips, And Best Practices.*

Elbanna, S., & Armstrong, L. (2024). Exploring the integration of ChatGPT in education: adapting for the future. *Management & Sustainability: An Arab Review, 3*(1), 16-29. doi:https://doi.org/10.1108/MSAR-03-2023-0016

Elkhodr, M., Gide, E., Wu, R., & Darwish, O. (2023). ICT students' perceptions towards ChatGPT: An experimental reflective lab analysis. *STEM Education, 3*(2), 70-88. doi:10.3934/steme.2023006

Eulerich, M., Sanatizadeh, A., Vakilzadeh, H., & Wood, D. (2024). Is it all hype? ChatGPT's performance and disruptive potential in the accounting and auditing industries. *Review of Accounting Studies, 29*, 2318–2349. doi:https://doi.org/10.1007/s11142-024-09833-9

Faisal, E. (2024). Unlock the potential for Saudi Arabian higher education: a systematic review of the benefits of ChatGPT. *Frontiers in Education, 9*. doi:https://doi.org/10.3389/feduc.2024.1325601

Feng, T., Liu, S., & Ghosal, D. (2024). CourseAssist: Pedagogically Appropriate AI Tutor for Computer Science Education. *Proceedings of the 2024 on ACM Virtual Global Computing Education Conference V. 2*, (p. 310-311). doi:https://doi.org/10.1145/3649409.3691094

Gao, A. (2023). *Prompt Engineering for Large Language Models.* SSRN. Recollit de https://ssrn.com/abstract=4504303

Garg, M., Mehndiratta, V., & Vidushi. (2024). Accountancy Capabilities of Large Language Models. *4th International Conference on Advancement in Electronics & Communication Engineering (AECE).* doi:10.1109/AECE62803.2024.10911393

Gökçearslan, S., Tosun, C., & Erdemir, Z. (2024). Benefits, challenges, and methods of artificial intelligence (AI) chatbots in education: A systematic literature review. *International Journal of Technology in Education, 7*(1), 19-39.

Guo, K., & Wang, D. (2023). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*, 8435–8463. doi:https://doi.org/10.1007/s10639-023-12146-0

Kanont, K., Pingmuang, P., Simasathien, T., Wisnuwong, S., Wiwatsiripong, B., Poonpirome, K., . . . Khlaisang, J. (2024). Generative-AI, a Learning Assistant? Factors Influencing Higher-Ed Students' Technology Acceptance. *Electronic Journal of e-Learning, 22*(6), 18-33. doi:https://doi.org/10.34190/ejel.22.6.3196

Katz, D., Bommarito, M., Gao, S., & Arredondo, P. (2024). Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A, 382*. doi:10.1098/rsta.2023.0254

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence, 6*. doi:https://doi.org/10.1016/j.caeai.2024.100225

Kohnke, L., Moorhouse, B., & Zou, D. (2023). ChatGPT for Language Teaching and Learning. *RELC Journal, 54*(2). doi:10.1177/00336882231162868

Korzynski, P., Mazurek, G., Krzypkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review, 11*(3). doi:10.15678/EBER.2023.110302

Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K., Lukowicz, P., Kuhn, J., . . . Karolus, J. (2024). Unreflected Acceptance - Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education. A F. e. Lorig, *HHAI 2024: Hybrid Human AI Systems for the Social Good* (p. 199-212). IOS Press. doi:10.3233/FAIA240195

Kuhail, M., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies, 28*(1), 973-1018. doi:https://doi.org/10.1007/s10639-022-11177-3

Lee, A. (2025). *How Well do LLMs Compress Their Own Chain-of-Thought? A Token Complexity Approach.* arXiv. Recollit de https://arxiv.org/html/2503.01141v1

Leung, C. (2024). Promoting optimal learning with ChatGPT: A comprehensive exploration of prompt engineering in education. *Asian Journal of Contemporary Education, 8*(2), 104-114.

Li, P., Kinshuk, & Huang, Y. (2024). Enhancing ChatGPT in POE Inquiry Learning for STEM Education to Improve Critical Thinking Skills. A Y. P. Cheng (Ed.), *Innovative Technologies and Learning. ICITL 2024. Lecture Notes in Computer Science, vol 14785.* Springer. doi:https://doi.org/10.1007/978-3-031-65881-5_4

Lower, K., Seth, I., Lim, B., & Seth, N. (2023). ChatGPT-4: Transforming Medical Education and Addressing Clinical Exposure Challenges in the Post-pandemic Era. *Indian Journal of Orthopaedics, 57*, 1527–1544. doi:https://doi.org/10.1007/s43465-023-00967-7

Maruszewska, E., Ziemba, E., Grabara, D., & Renik, K. (2024). The determinants of ChatGPT usage among accounting students: the role of habit, social influence, and facilitating conditions. *Zeszyty Teoretyczne Rachunkowości, 48*(3), 215-232. doi:https://doi.org/10.5604/01.3001.0054.7264

Mohamed, A. (2023). Exploring the potential of an AI-based Chatbot (ChatGPT) in enhancing English as a Foreign Language (EFL) teaching: perceptions of EFL Faculty Members. *Education and Information Technologies, 29*, 3195–3217. doi:https://doi.org/10.1007/s10639-023-11917-z

Okagbue, E., Ezeachikulo, U., Akintunde, T., Tsakuwa, M., Ilokanulo, S., Obiasoanya, K., & et al. (2023). A comprehensive overview of artificial intelligence and machine learning in education pedagogy: 21 Years (2000–2021) of research indexed in the scopus database. *Social Sciences & Humanities Open, 8*(1). doi:https://doi.org/10.1016/j.ssaho.2023.100655

Piatykop, O., Yeva, A., Pronina, O., & Balalayeva, E. (2025). Researching artificial intelligence language models for developing the virtual language learning assistant. *CEUR Workshop Proceedings*, (p. 132-142).

Pratama, M., Sampelolo, R., & Lura, H. (2023). Revolutionizing education: Harnessing the power of artificial intelligence for personalized learning. *Klasikal: Journal of Education, Language Teaching and Science, 5*(2), 350-357.

Rahman, M., & Watanobe, Y. (2023). ChatGPT for Education and Research: Opportunities, Threats, and Strategies. *Applied Sciences, 13*(9). doi:https://doi.org/10.3390/app13095783

Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira, F., & et al. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching*, 41-56. doi:https://doi.org/10.37074/jalt.2023.6.1.29

Savelka, J., Agarwal, A., Bogart, C., Song, Y., & Sakr, M. (2023). Can generative pre-trained transformers (gpt) pass assessments in higher education programming courses? *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V.1* (p. 117-123). New York: Association for Computing Machinery. doi:https://doi.org/10.1145/3587102.3588792

Slamet, J. (2024). Potential of ChatGPT as a digital language learning assistant: EFL teachers' and students' perceptions. *Discover Artificial Intelligence, 4*(4). doi:https://doi.org/10.1007/s44163-024-00143-2

Sok, S., & Heng, K. (2023). ChatGPT for Education and Research: A Review of Benefits and Risks. *SSRN*. doi:http://dx.doi.org/10.2139/ssrn.4378735

Spirgi, L., & Seufert, S. (2025). GenAI as a Learning Assistant, an Empirical Study in Higher Education. *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025)*, *2*, p. 27-34. doi:10.5220/0013199300003932

Sundkvist, C., & Kulset, E. (2024). Teaching accounting in the era of ChatGPT – The student perspective. *Journal of Accounting Education, 69*. doi:https://doi.org/10.1016/j.jaccedu.2024.100932

Totlis, T., Natsis, K., Filos, D., Ediaroglou, V., Mantzou, N., Duparc, F., & Piagkou, M. (2023). The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. *Surgical and Radiologic Anatomy, 45*, 1321–1329. doi:https://doi.org/10.1007/s00276-023-03229-1

Wardat, Y., Tashtoush, M., AlAli, R., & Jarrah, A. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(7). doi:https://doi.org/10.29333/ejmste/13272

Wood, A. e. (2023). The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions? *Issues in Accounting Education, 38*(4), 1-28.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., . . . Gasevic, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*. doi:https://doi.org/10.1111/bjet.13370

Young, J., & Shishido, M. (2023). Investigating OpenAI’s ChatGPT potentials in generating Chatbot's dialogue for English as a foreign language learning. *International journal of advanced computer science and applications, 14*(6). doi:10.14569/IJACSA.2023.0140607