Italian Economic Journal
https://doi.org/10.1007/s40797-025-00354-1

RESEARCH PAPER

Artificial Intelligence in Education: Computer-Assisted Learning and AI-guided Tutors

Almudena Sevilla[1] · Pilar Cuevas-Ruiz[1,2] · Luz Rello[3] · Ismael Sanz[1,4]

Received: 14 July 2025 / Accepted: 29 October 2025

Abstract
Artificial Intelligence (AI) and Computer-Assisted Learning (CAL) offer powerful tools to improve foundational skills and close educational gaps, with evidence showing meaningful gains in student performance, especially in mathematics. Recent advancements in these technologies have generated optimism about their transformative potential in classrooms worldwide. These technologies are increasingly being piloted at scale, reshaping the way teachers deliver content and students engage with material. However, their impact depends less on access to devices and more on how they are integrated into teaching—through curriculum alignment, teacher training, and interactive design that promotes active learning. Without careful implementation, these tools risk widening existing inequalities. Using new evidence from Italy, we show that digital divides in AI adoption persist across schools and regions, reflecting broader social and economic disparities. Our findings suggest that realising the potential of AI in education requires inclusive policies and targeted investment to ensure no student is left behind, and that the benefits of digital innovation are shared equitably.

Keywords Artificial intelligence in education · Computer-Assisted learning · AI-guided tutors · Digital divide · PISA · Italian schools

JEL Codes I21 · O33 · C88 · D83

✉ Ismael Sanz
Ismael.Sanz@urjc.es

[1] London School of Economics, London, England, UK

[2] Universidad de Sevilla, Sevilla, Spain

[3] IE University, Segovia, Spain

[4] Universidad Rey Juan Carlos, Madrid, Spain

Published online: 10 November 2025

1 Introduction

Persistent educational inequalities remain one of the most pressing challenges in advanced economies. Gaps in foundational skills—particularly literacy and numeracy—undermine students' long-term prospects and constrain social mobility and economic growth (Heckman et al. 2006; OECD 2023a). Recent PISA 2022 data show that nearly one in three students across OECD countries fails to reach basic proficiency in mathematics, and more than a quarter fall short in reading. These learning deficits, exacerbated by the pandemic, are concentrated among disadvantaged groups and show few signs of narrowing. This paper explores whether and how AI-guided tutoring and computer-assisted learning (CAL) technologies can help reduce these gaps—and under what conditions they may instead reinforce them.

A growing body of causal evidence shows that CAL and AI-powered educational tools can produce substantial learning gains, especially in mathematics. Randomised evaluations of scalable interventions in developing and advanced economies alike find effects of 0.2 to 0.3 standard deviations from well-designed CAL programs (Muralidharan et al. 2019; Büchel et al. 2022; Bhatt et al. 2024). These tools allow for personalised feedback, adaptive pacing, and flexible delivery models, making them a potentially cost-effective alternative or complement to high-dosage human tutoring—the current gold standard. Importantly, evidence suggests that the most effective AI-based systems are those that guide students through hints and scaffolding rather than simply providing answers, promoting deeper engagement and cognitive autonomy (Bastani et al. 2024).

Yet the promise of these technologies is not guaranteed. Their success depends critically on thoughtful integration into teaching practice—alignment with curricula, sustained teacher training, and institutional support (Oreopoulos et al. 2024). Without these conditions, CAL and AI tools risk becoming ineffective or even counterproductive. Concerns include cognitive offloading, algorithmic bias, and especially the reinforcement of existing educational inequalities through unequal access to devices, connectivity, and teacher preparedness (Oakley et al. 2025). These risks are particularly salient in countries with uneven digital infrastructure or high regional disparities.

To investigate these dynamics, we present new empirical evidence from Italy—a country with both a strong policy push for digital education and persistent internal inequality. Using weekly region-level data from Google Trends on ChatGPT usage across all Italian regions, we track the adoption of generative AI over time and analyse how it varies with regional income. We interpret data from Google Trends on ChatGPT usage in the category of education and employment as a proxy for the adoption of AI tools in education, reflecting broader patterns of engagement with generative AI. Our empirical strategy uses regional fixed effects and controls for economic conditions to isolate the structural drivers of adoption.

Our findings show that initial engagement with generative AI was concentrated in higher-income regions but gradually diffused as digital infrastructure improved, suggesting convergence. However, continued disparities in usage highlight the challenges of equitable integration. These results confirm that AI adoption in education reflects broader structural divides—and that, without targeted policy, technological advances may amplify rather than mitigate existing gaps.

This paper contributes to a growing but still fragmented literature examining whether and how AI-guided tutoring and Computer-Assisted Learning (CAL) can reduce educational inequalities—or instead risk reinforcing them. While studies such as Escueta et al. (2020) and Bastani et al. (2024) provide rigorous evidence that these technologies can generate meaningful learning gains, they largely focus on average treatment effects in controlled or pilot settings (Bhatt et al. 2024). Crucially, they offer limited insight into how such tools are adopted and implemented across the diverse and unequal contexts where education systems operate. This is a problem, because whether AI and CAL reduce or widen inequalities depends not only on their effectiveness, but on who accesses them, where, and under what conditions. In particular, the literature has not systematically examined how digital divides—in infrastructure, school capacity, and regional inequality—shape the real-world diffusion of these technologies. We address this gap by using novel region-by-week Google Trends data on ChatGPT searches across Italy to study the diffusion of generative AI in education. This high-frequency, spatially disaggregated data allows us to move beyond controlled evaluations and observe how adoption unfolds in practice—across regions with varying digital infrastructure and socioeconomic conditions. The findings underscore that realising the potential of AI in education requires addressing structural barriers to adoption and ensuring that implementation reaches the students who stand to benefit the most.

The rest of the paper proceeds as follows. Section 2 reviews the causal literature on CAL and AI tutors and their impact on student learning. Section 3 discusses design features and implementation challenges, drawing on recent empirical studies. Section 4 analyses the risks and limitations of technology use in education. Sections 5 and 6 focuses on the Italian context and presents new evidence on regional disparities in AI adoption. Section 7 concludes with implications for policy and the broader goal of reducing educational inequality.

2 The State of the Art Regarding the Use of Educational Technology

This section reviews the causal literature not only to assess average effects of educational technologies, but to understand how these interventions perform across different contexts and student backgrounds—an important consideration for evaluating their role in addressing educational inequalities. In recent years, there has been a rapid expansion in the use of educational technology, accompanied by significant investment in technological tools, including computers, tablets, mobile phones, and the Internet, aimed at enhancing academic quality. The literature review by Escueta et al. (2020) in the *Journal of Economic Literature* analyses rigorous articles that provide precise estimates of the causal effects of technological interventions, such as those obtained through Randomised Controlled Trials (RCTs) and Regression Discontinuity Designs (RDDs). The authors focus on the impact of technology in education, focusing on four possible interventions: (a) access to technology, (b) computer-assisted learning (CAL), (c) online courses, and (d) technology-enabled behavioural interventions.

Regarding the first technological intervention, the authors show that providing technological devices, such as computers or tablets, does not guarantee significant improvements in academic performance. For these devices to be effective, they need to be accompanied by specific educational programs and pedagogical support. About technology-assisted learning, CAL programs have proven particularly effective in mathematics, where personalised teaching and immediate feedback can improve student performance. In comparison, the impact of CAL programs in areas such as reading and writing is less clear and requires more research to determine their effectiveness. The third intervention, online courses and MOOCs (Massive Open Online Courses), are valuable tools for expanding access to quality education. Still, they face significant challenges, such as high dropout rates and low average student engagement. Escueta et al. (2020) emphasise that retention and engagement are relevant for the success of online courses, highlighting the need for effective strategies to increase participation and completion rates in this type of training. The fourth intervention, technology-enabled behavioural interventions, such as strategies for sending reminders and messages to increase motivation, have shown the potential to improve attendance and academic performance. However, their effectiveness varies depending on the design and frequency of the interventions.

Escueta et al. (2020) review highlights both the promises and limitations of technology's role in education. The key to maximising educational technology lies in its careful and contextualised implementation, considering the specific needs of students and the capabilities of teachers. The success of technological interventions depends on personalisation and adequate support for teachers and students. Integrating technological tools effectively into the educational curriculum is essential to maximise their benefits.

In this section, we will examine the conclusions from rigorous causal literature on the use of CAL programs in education. CAL programs facilitate personalised instruction tailored to each student's learning pace. They offer exercises and activities that can be repeated as needed, providing immediate feedback to students, teachers, and schools regarding correct responses and errors. These educational software tools can complement skill development by addressing challenges faced by educators, such as managing diverse learning levels within a single classroom. Additionally, some CAL programs are adaptive, leveraging increasingly sophisticated artificial intelligence to adjust content based on users' cognitive abilities and progress. They can deliver individualised feedback and swiftly collect data on student performance, tasks that might be challenging for educators due to time constraints.

Notably, CAL programs have demonstrated a positive and significant impact on mathematics education, though there is less evidence regarding their effectiveness in other subjects, such as language. In their literature review, Escueta et al. (2020) examined 31 RCTs to provide causal evidence on the impact of computer-assisted programs on student learning. Many of these studies focused specifically on algebra and elementary education. Of the 31 studies reviewed, 21 reported statistically significant positive effects, with many estimates being precise and of substantial magnitude. The majority (16 out of 21) of studies that found a positive impact concentrated on mathematics programs.

Table 1 combines the most relevant articles from the economics of education literature reviewed by Escueta et al. (2020) with recent RCTs published after their review, which provide new evidence on the effectiveness and implementation mechanisms of CAL in real-world school settings. The studies are listed in chronological order of publication, to highlight the evolution of evidence over time.

One question raised in previous research is whether the use of software improves outcomes because students are spending more time learning, or due to the digital tool itself. In other words, it is possible that if students had more hours of traditional classes instead of increased use of digital tools, their academic performance might also improve.

Büchel et al. (2022) examined the relative effectiveness of a freely available CAL program. To distinguish between the effects of additional teaching and software use, the RCT included three treatments that did not interfere with regular lessons. The first treatment, consisting of 40 classes, included additional traditional mathematics lessons (without software use and outside of school hours) taught by a teacher. In the second and third treatments, which also took place outside school hours, a computer-assisted mathematics learning program was used. The second group was monitored by support staff (39 classes), while the third was supervised by teachers (another 39 classes). Each of the three treatments consisted of two 90-minute mathematics lessons per week over six months, nearly doubling the number of math classes students received during the program.

Additionally, there were two control groups: (a) schools that did not receive the treatment, constituting the "pure" control group (29 schools), and (b) students from the 28 treated schools who did not participate because their classes were randomly excluded from the program. Within the 28 treatment schools, 118 classes received the intervention, and 40 did not participate. The latter constituted the second control group aimed at measuring "externalities," i.e., whether students who did not benefit from the program improved their results because other classmates in other classes of the same school were treated.

Using Intention-to-Treat (ITT) estimates, Büchel et al. (2022) demonstrated that being assigned to additional lessons with the CAL program, monitored by support staff (treatment 2), resulted in a 0.21 standard deviation (SD) increase in math scores, and a 0.24 SD increase when supervised by teachers (treatment 3). In both cases, the magnitude of improvement is equivalent to students advancing more than half a school year in mathematics. Additional traditional classes (treatment 1) also increased academic performance in mathematics, but by 0.15 SD, a significant difference compared to treatment 3, but not to treatment 2. Furthermore, Büchel et al. (2022) found that the use of CAL programs enhances learning, even in large classes with heterogeneous student levels —a benefit not observed in traditional courses. When using treatment assignment as an instrumental variable (IV) for attendance, the estimated effects of treatments 2 and 3 increase to 0.38 and 0.40 SD, respectively.

Table 1 Summary of experimental evidence on Computer-Assisted learning (CAL) Programs, listed in chronological order

Author(s)	Publication & Year	Country	Output Measure	Direction of Effect	Effect Size	Intervention	Comments	Methodology	Objective	Sample Size
Dynarski et al.	US Dept of Education (IES) Final Report, 2007	USA	Standardized math & reading test scores	No significant effect	n.s.	16 types of math/reading software implemented in K-12 schools	Large sample, low implementation fidelity, short exposure; software often not aligned with curriculum	Cluster RCT (randomized at teacher/class level)	Assess effectiveness of widely-used math/reading software in real-world settings	439 teachers, 132 schools, 33 districts
Barrow, Markman & Rouse	American Economic Journal: Economic Policy, 2009	USA (Chicago)	Algebra test scores	Positive; strongest for largest, most heterogeneous classes; higher absenteeism	+0.17–0.18 SD (full sample); up to +0.23 SD (largest/most heterogeneous classes)	"I Can Learn" interactive CAL algebra program, integrated into regular lessons	Effects larger in bigger, more diverse classes; strong design, uses administrative data	RCT at class level	Test impact of CAL algebra on urban students' performance	1,605 students, 142 classes, 10 schools
Rockoff	NBER Working Paper, 2015	USA (NYC)	Math test scores	No significant effect	n.s.	"School of One" personalized math program in middle schools	Large-scale, rotational implementation; challenging to scale; "dose" varied across schools	RCT at school level	Assess effect of personalized math rotation model	5,070 students, 8 schools

Table 1 (continued)

Author(s)	Publication & Year	Country	Output Measure	Direction of Effect	Effect Size	Intervention	Comments	Methodology	Objective	Sample Size
Van Klaveren et al.	Economics of Education Review 2017	Netherlands	Test scores (economics, biology, history, language)	No significant impact	n.s.	Adaptive CAI vs. static digital content, various subjects, secondary school	No significant difference for adaptivity; covers multiple subjects, randomized within schools	RCT at student level within classes	Compare adaptive vs. non-adaptive CAI across multiple domains	1,021 students, 4 schools
Büchel, Monteiro, Justman & Wolf	Journal of Labor Economics, 2022	El Salvador	Math test scores	Positive, significant; larger for CAL than for extra traditional lessons	+0.15 SD (extra traditional), +0.21 SD (CAL+staff), +0.24 SD (CAL+teacher); up to +0.38–0.40 SD (IV)	Three arms: extra math classes, CAL with support staff, CAL with teachers	CAL effects robust to class size, strongest with teacher support; persistent in heterogenous/large classes	Cluster RCT (schools/classes), plus IV analysis	Test effects of CAI with different supervision vs. traditional instruction	~6,400 students, 57 schools
Hirata, Guilherme	Journal of Human Capital, 2022	Brazil	Arithmetic test scores	Positive and significant (short and medium term)	+0.56 SD (short term); +0.17 SD (one year)	In-class, game-based CAI for arithmetic; primary (grades 1–3); ~20 min/day	Impact greater for basic skills and younger students; medium-term effects persist for accuracy, not for speed	Cluster RCT (schools/classes), random assignment	Evaluate CAI for foundational math skills in early primary	~870 students, 12 schools, 36 classes

Table 1 (continued)

Author(s)	Publication & Year	Country	Output Measure	Direction of Effect	Effect Size	Intervention	Comments	Methodology	Objective	Sample Size
Oreopoulos, Gibbs, Jensen & Price	Ed-Working Paper, 2024	USA (Texas, Tennessee)	Math achievement (STAAR)	Positive, significant; larger with more practice time	+0.12–0.22 SD (depending on usage)	Khan Academy mastery-based CAL + weekly "khoach" teacher support	Larger effects with > 35 min/week practice; implementation fidelity key; two district-wide RCTs	Two cluster RCTs (teachers/classes)	Assess impact of mastery-based CAL + coaching on math achievement	224 teachers, 10,979 students
Bhatt, Guryan, Khan, LaForest-Tucker & Mishra	NBER Working Paper, 2024	USA (Chicago, NYC)	Standardized math test scores, Math GPA, failure rates	Positive, significant; effects sustained after 1 year	TOT: +0.23 SD (math); +0.24 GPA; −22% math failures	Hybrid 4-to-1 high-dosage tutoring model: pairs alternate daily between tutor and CAL (ALEKS) for 50 min/day	Cost per student reduced by 30%; tutors needed halved; effects comparable to traditional 2-to-1 model; persistent effects at 1 year	Student-level RCT, 2 cohorts (2018–19, 2019–20)	Test scalability/effectiveness of high-dosage tutoring with integrated CAL	3,906 students, 7 schools, 72 tutors

The results of this study show that students in control classes within treated schools (“control for measuring externalities”) in the intervention with more traditional math classes (treatment 1) achieve better results than those in schools where no class participated in the treatment (“pure” control), particularly among students with a low prior level. The treatment groups that used CAL, whether supervised by support staff or teachers (treatments 2 and 3), exhibited significantly higher math performance than the pure control group across the distribution. However, the gap seems to close for students with higher pre-experiment performance. In summary, Büchel et al. (2022) provide evidence that advances in CAL can, at least in part, be attributed to the software rather than solely to the increase in the number of math lessons. Lessons delivered through computer-assisted programs lead to more significant learning and are less sensitive to class size and student ability heterogeneity.

The second article, published after the literature review by Escueta et al. (2020) and also included in Table 1, is by Hirata (2022), who analysed a computer-assisted program used during class time. Like Büchel et al. (2022), this randomised experiment isolates the impact of software use from the effect of having more instructional time, which, in this case, does not occur because the total number of instructional hours remains constant. Hirata (2022) examined the impact of using a software tool to learn and practice mathematics (arithmetic) through games designed for primary school students in Brazil. Students in the treatment group used the software for up to 20 min during the 4-hour school day over two months. First-, second and third-grade primary students who used the software increased their math scores by 0.56 of the SD in the short term (immediately after the intervention) and by 0.17% in the medium term (one year after the intervention ended). The impact of many educational measures fades in subsequent years, although in this case, one-third of the initial impact remains. Hirata (2022) suggests that the more significant effect found in this randomised experiment compared to previous research may be due to CAL being more effective in improving student outcomes in lower grades, given that the skills taught and learned are more basic.

The third article published after the Escueta et al. (2020) review highlighted is by Oreopoulos et al. (2024), which focuses on the implementation of mastery-based learning through technology and continuous teacher support. This study evaluates a program designed to foster greater mastery learning in mathematics at both primary and secondary education levels. The intervention includes the use of CAL combined with weekly teacher support through a “coach.” These coaches offer teachers proactive guidance on how to effectively utilise CAL tools to personalise learning and monitor student progress.

Oreopoulos et al. (2024) conducted two randomised experiments in Nashville and Arlington (both in the US) to evaluate the impact of this intervention. The results show significant improvements in mathematics performance, ranging from 0.12 to 0.22 SD, depending on the amount of weekly practice time with the CAL program. Students who participated in classrooms that achieved an average of at least 35 min of weekly practice with CAL showed more notable improvements. Key factors contributing to the program’s success included high initial teacher engagement, a clear implementation strategy for practice, and teachers' willingness to closely monitor progress and follow up with students who were struggling.

The importance of fidelity in implementation and teacher commitment is a fundamental finding in Oreopoulos et al. (2024). This study makes four key contributions: it demonstrates the effectiveness of a program that primarily uses existing resources to facilitate more personalised learning; it provides evidence of the effectiveness of Khan Academy in a developed country setting; it highlights how the effects of the intervention critically depend on fidelity in implementation and training; and it offers insights into why some teachers can implement more CAL practice time than others. Institutional support, exclusive dedication to the program, belief in its effectiveness, and active participation are all factors that influence the amount of practice time teachers implement. The results suggest that the efficacy of CAL depends more on the quality of its implementation than on the platform itself, underscoring the need for continuous and structured support for teachers in utilising these technological tools.

A further relevant contribution to the debate on the use of technology in scalable tutoring models is provided by Bhatt et al. (2024), who evaluate the impact of integrating computer-assisted learning (CAL) into high-dosage tutoring programs in U.S. public high schools. The study is based on a randomised experiment conducted across three public high schools in Chicago and four in New York City during the 2018–2019 and 2019–2020 academic years, with 9th-grade students as participants.

The intervention was structured as a "4-to-1" tutoring model: four students shared a table. They alternated daily between working in pairs with a human tutor and engaging in mathematics activities on a CAL platform during a 50-minute daily session. This model, called "Saga Technology," was designed to reduce both the costs and staffing requirements associated with the traditional daily 2-to-1 tutoring model. The cost per student was reduced by 30%, and the number of required tutors by 50%, while maintaining implementation during regular school hours.

The experiment's results show a significant impact on academic outcomes: the treatment-on-the-treated (TOT) effect was 0.23 standard deviations in mathematics, a magnitude comparable to the daily 2-to-1 tutoring model evaluated by Guryan et al. (2023). Improvements in mathematics (+0.24 points) and a 22% reduction in failure rates for this subject were also observed. Moreover, the effects were partially replicated in the study's second year (2019–2020), and positive, persistent effects on mathematics achievement were found one year after the intervention.

In contrast to other studies focused on AI conversational virtual tutors, this research explores a hybrid approach in which technology does not replace the tutor but rather frees up part of their time, enabling greater scalability without sacrificing effectiveness. The authors highlight that even without personalised interaction with an intelligent system, the strategic use of CAL in combination with human tutoring can yield substantial and sustained improvements in real-world school contexts.

3 Applications, Promises, and Challenges of AI: AI-Guided Tutors in the Classroom

The growing body of evidence on CAL provides valuable insights into how technology can support student learning and help address educational inequalities. However, the field is now witnessing a new technological paradigm: the arrival of generative AI models. These advances go beyond traditional CAL software, offering new ways to adapt instruction, deliver feedback, and simulate aspects of human tutoring at scale. It is worth noting that the most recent research of Oreopoulos et al. (2024) illustrates the future potential of integrating CAL approaches and AI-guided tutoring, to personalize and scale mastery learning further, especially as advances in large language models may allow virtual tutors to support both students and teachers with real-time feedback, progress monitoring, and assignment design.

The emergence of generative AI—such as GPT-4 models capable of generating text, maintaining conversations, and solving complex problems—represents a paradigm shift. These tools can personalise student interaction, adapting content, pace, and type of support in response to each student's answers. Automated conversational tutoring of this kind can facilitate teaching practice, support metacognitive skills, and activate prior knowledge, provided that interface design and pedagogical principles are coherent.

What sets generative AI apart is not only its technical capabilities, but its ability to replicate key aspects of human tutoring at scale—a potential the OECD (2023a) has highlighted as especially relevant for today's classrooms, where teachers often face the challenge of supporting students at varying levels within the same group. The real opportunity lies in using generative AI to enrich and diversify learning, for example by creating multiple types of explanations or analogies, and by guiding students through self-reflection and planning their learning. The adaptive nature of these tools makes them particularly valuable for students with learning difficulties or those in under-resourced settings.

Beyond the classroom, AI can promote autonomous learning by helping students summarise information, improve their writing, or explore topics of interest independently. However, the successful integration of AI into schools depends not only on the technology but also on teacher training, ethical standards, and safeguards for data protection. As the OECD (2023b) emphasises, the long-term benefits will ultimately depend on thoughtful implementation within robust pedagogical and institutional frameworks, rather than on technological adoption alone.

AI now can deliver individualised tutoring at scale—a goal that once seemed out of reach due to high costs. Today's technology enables the envisioning of scenarios where every student can interact with an AI assistant capable of explaining concepts, resolving doubts, or providing support tailored to each learner's pace. The potential impact is particularly significant for students who struggle most, as these are the learners who tend to fall behind in traditional models. AI-powered solutions can address the true diversity of learning levels and styles found within a single classroom. However, the benefits of AI in education will not materialise on their own. Real improvements in learning and reductions in educational inequality will require sustained institutional effort. Among the priorities identified are rigorous evaluation of pilot programs, the development of tools grounded in sound pedagogical principles, comprehensive teacher training, and ongoing support for effective classroom integration.

A particularly relevant contribution to the literature on the educational effects of generative AI-based tutors is the experimental study by Bastani et al. (2024) in Turkey. Unlike many studies focused on university or simulated settings, this intervention was designed and implemented in collaboration with the Turkish Ministry of Education in real classroom conditions, with 3,200 secondary students in a low-resource context. Its goal was to analyse not only whether AI tutors improve academic achievement, but also how the design of the user interface—that is, the way students interact with the model—shapes the tool's impact on learning.

The intervention compared three groups: (1) a control group with no AI access; (2) a group with access to a standard GPT-4-based AI tutor (GPT-Base); and (3) a group using a modified GPT-4 version integrating pedagogical principles (GPT-Tutor). Both models helped students solve math problems. Whereas GPT-Base provided direct answers—including complete solutions when requested—GPT-Tutor was configured to avoid giving complete answers, instead offering partial hints, counterexamples and guiding questions, following a scaffolding approach inspired by Vygotsky's pedagogy and human tutoring practices. Educational scaffolding is a strategy in which the teacher provides temporary support to help students accomplish tasks that are not yet entirely within their grasp. As students gain autonomy, this support is gradually withdrawn to foster active learning and the development of complex skills.

During the practice phase, both AI groups significantly outperformed the control group, with gains of + 0.137 points for GPT-Base and + 0.361 points for GPT-Tutor. However, in the subsequent unaided test, only the GPT-Tutor group-maintained performance comparable to that of the control group, while the GPT-Base group performed worse (–0.054 points, representing a 17% drop). In other words, students who previously solved problems with GPT-Base learned less than those who received no assistance.

This finding suggests that AI tutoring does not guarantee improved learning on its own and may even be counterproductive without proper guidance. By providing complete answers, GPT-Base encouraged a passive approach, with students outsourcing cognitive effort to the machine—a phenomenon described as cognitive offloading.

The adverse effect of GPT-Base was particularly pronounced among students with lower initial performance, whereas GPT-Tutor had the most significant positive impact on this same group. Thus, pedagogical design not only improved overall results but also reduced inequalities, serving as a compensatory mechanism. This is especially relevant for equity-focused education policies, as it demonstrates that careful technological design can help close gaps, while an uncritical approach may exacerbate them.

Beyond quantitative outcomes, the study also analysed student interactions with the AI. Users of GPT-Base tended to ask superficial, answer-seeking questions (e.g., “What’s the solution?“), while GPT-Tutor users engaged in richer interactions, asking for clarifications, interpreting hints, and reconsidering their strategies. Qualitatively, a deeper learning environment emerged, with students taking on a more active and reflective role rather than simply receiving information.

Another notable aspect of the study is its implementation: the intervention required no proprietary software or costly infrastructure and used computers already available in schools. Yet the educational impact depended entirely on the system’s pedagogical design. The value of AI in education lies not in its technical sophistication per se, but in the educational intent that guides its use.

In short, Bastani et al. (2024) provide strong evidence that the instructional design of AI tutors is a key determinant of their effectiveness. Tools like GPT-4 can have positive or negative effects depending on how student interaction is structured. When limited to answer-giving, they may inhibit autonomous learning; when configured to support reasoning and self-regulation, they can enhance learning and promote equity.

A recent intervention with individualised AI-guided tutors in low-income contexts, evaluated by Henkel et al. (2024), is that of the Rori conversational tutor. The study was conducted in eleven schools in Ghana, part of the Rising Academies network. The study involved nearly 500 primary students (grades 3–8), with schools randomly assigned to either a treatment group (236 students) or a control group (241 students).

Rori is an AI math tutor accessible via WhatsApp, designed to function on basic mobile phones with limited network capacity. Its content is organised into over 500 micro-lessons aligned to the Global Proficiency Framework for Mathematics, each offering a short explanation, practice exercises, and scaffolding. If a student makes a mistake, the system first provides a hint, then an answer. Natural language interaction simulates a personalised tutoring experience.

Treatment group students used Rori for two 30-minute weekly study hall sessions over a period of 32 weeks, supervised by teachers but requiring no additional staff, training, or curricular changes. Results show a statistically significant improvement in math performance for the treatment group, with an effect of 0.36 SD—a substantial impact in the economics of education literature, roughly equivalent to an extra year of learning. The intervention cost was about $5 per student, supporting its viability as a cost-effective and scalable solution in resource-constrained systems. Although limited to the first year of implementation, these initial results underscore the potential of conversational AI tutors like Rori to improve learning outcomes in low- and middle-income countries.

A complementary approach is found in the study by Thomas et al. (2024), which examines a hybrid tutoring model that combines algorithmic personalisation through AI software with direct human tutor interaction. In contrast to Henkel’s solution, designed for minimal infrastructure, Thomas et al.’s intervention targets vulnerable secondary students in urban U.S. schools. Their model leverages AI to detect learning patterns in real time, enabling human tutors to provide targeted emotional, motivational, and pedagogical support.

The program was evaluated using three quasi-experimental cases, comparing students who received hybrid tutoring with those who only used adaptive math software. Results show significant gains in both achievement and engagement, particularly among lower-performing students. Human tutors, informed by real-time system data, intervened more effectively and personally than they would have without such insights. This effective combination of AI and human intelligence is central to the model's success.

The study further notes that the hybrid approach is scalable, with an annual per-student cost of approximately $700, which is significantly lower than that of fully human tutoring, making it viable for resource-limited districts with basic infrastructure. Thomas et al. (2024) emphasise the importance of tutor dashboards and maintaining a low tutor-to-student ratio to ensure genuinely personalised support.

Overall, Thomas et al. (2024) demonstrate that strategic AI-human combinations can improve learning cost-effectively and sustainably. Their conclusions echo those of Henkel et al. (2024): the key is not only the power of the algorithm, but how it is integrated into a robust, student-centred pedagogical framework.

AI can also help alleviate administrative burdens for teachers. Tools that automate lesson preparation, grading, or material search could significantly reduce time spent on routine tasks, allowing teachers to focus on the essentials: guiding, motivating, observing, and providing close support to each student.

A further significant contribution, centred on teachers and tutors, comes from Wang et al. (2024), who evaluate Tutor CoPilot. This system provides real-time expert support to human tutors during math sessions. Unlike student-facing programs, this approach combines generative AI with active tutor mediation, aiming to amplify pedagogical capabilities. In the first randomized trial of a Human-AI system in live tutoring, Wang et al. (2024) partnered with FEV Tutor and a large Southern U.S. school district, involving 900 tutors and 1,800 K-12 students from historically underserved communities. Tutors were randomly assigned to either receive access to Tutor CoPilot or not, and the intervention ran for two months.

Tutor CoPilot is designed to assist tutors during live sessions by offering expert-like, context-specific guidance through a dedicated interface. Tutors can request suggestions based on the ongoing chat, lesson topic, and selected pedagogical strategies, such as prompting students to explain their reasoning or providing targeted hints. Notably, the system enables tutors to customize or choose from multiple suggested strategies, maintaining autonomy while elevating instructional quality. To protect privacy, the system de-identifies names and limits data sent to external AI services.

The impact was notable: students whose tutors had access to Tutor CoPilot were 4% points more likely to master math content. The effect was even greater for students with less experienced or lower-rated tutors, who saw improvements of up to 9% points compared to control. Tutor CoPilot also proved highly cost-efficient, with an estimated annual cost of about $20 per tutor based on usage patterns during the study.

A distinctive feature of the study is its scale and richness of data. Over the two months, the analysis encompassed 4,136 tutoring sessions, resulting in more than 550,000 chat messages exchanged between tutors and students. These messages were systematically analysed using natural language processing classifiers to identify the pedagogical strategies employed. The results revealed that tutors with access to Tutor CoPilot were significantly more likely to use evidence-based teaching strategies—such as asking guiding questions or prompting students to explain their reasoning—and less likely to give away answers, aligning with high-quality instructional practices.

Qualitative interviews with approximately 20 treatment tutors complemented the quantitative findings. Tutors reported that the real-time support provided by Tutor CoPilot helped them respond more effectively to student needs, especially in explaining complex concepts and breaking down difficult topics. However, they also noted that the relevance and grade-appropriateness of AI-generated suggestions could still be improved, highlighting the ongoing need to fine-tune such systems.

Overall, the study reinforces the idea that generative AI can serve as a "pedagogical co-pilot," helping to scale instructional quality without replacing the human role. By combining AI-driven expertise with human judgment and adaptability, Tutor CoPilot demonstrates the potential to bridge gaps in instructional skill and deliver high-quality education at scale.

Among the most recent empirical studies on generative AI in real-world educational settings is De Simone et al. (2025), who conducted a randomised trial with a generative language model in a low-income educational context in Benin City, Nigeria. The intervention consisted of a six-week after-school tutoring program for first-year secondary students in nine public schools.

Over six weeks, students participated in twelve 90-minute lab-based sessions guided by teachers, utilising Microsoft Copilot to reinforce their English, digital, and AI skills. Fifty-two per cent of eligible students opted in, and treatment assignment was random, allowing for robust causal inference.

Results show significant gains: the treatment group improved by 0.31 SD in combined outcomes, with a 0.23 SD gain in English, the primary program focus. This corresponds to 1 year of conventional learning. Positive effects were observed across the achievement spectrum, with greater benefits for high-performing students and girls (helping close a gender gap). Each additional day of attendance led to a 0.031 SD gain; projecting to a full academic year, the total effect could reach 1.55 SD (or even 2 SD with full attendance). Teacher supervision was central: while AI provided main support, teachers were trained to guide sessions and prevent over-reliance on the tool, reinforcing, as in earlier adaptive learning research (Muralidharan et al. 2019), the importance of integrating technology with pedagogical supervision.

Cost-effectiveness was notable: at $48 per student, the program delivered 3.2 years of schooling for every $100 invested, and utilised free software without pre-set question banks, supporting scalability in resource-limited contexts. The study was conducted under adverse conditions (internet outages and power cuts), which reinforced the robustness of its findings and the potential of this model for similar settings.

Taken together, these studies provide robust evidence that well-designed, properly implemented AI tutoring can significantly improve student learning, especially in disadvantaged contexts. Table 2 summarises the empirical studies reviewed in this section on individualised AI-guided tutoring.

4 Research Gaps, Risks, and Open Questions for CAL Programs and AI-guided Tutors in Education

The previous sections have illustrated how CAL and AI-guided tutoring systems—when well designed and effectively implemented—hold considerable promise for improving student learning and reducing educational disparities. Yet as these technologies become increasingly integrated into classrooms and educational systems, it is equally important to acknowledge and scrutinise the risks and unresolved questions they introduce.

4.1 Unresolved Issues and Implementation Challenges of CAL Programs

CAL programs have demonstrated the potential to complement traditional education, particularly by addressing challenges faced by educators, such as managing heterogeneous learning levels within a classroom. Additionally, some of these programs are adaptive, using artificial intelligence to tailor content according to users' cognitive abilities and progress. A significant challenge, as noted by Bulman and Fairlie (2016), is determining whether CAL not only improves student performance but also provides better results than traditional instruction. Understanding this is essential for effectively guiding educational policies and technological investments in the education sector. Without this knowledge, resources could be invested in technologies that are no more effective than traditional teaching practices, thereby missing opportunities to enhance education genuinely. Until recently, the lack of data and the difficulty of conducting controlled experiments that capture all the factors involved have made it challenging to answer this question. Variability in implementation and dependence on local contexts have also hindered comparative analysis. Recent studies, such as those by Büchel et al. (2022) and Hirata (2022), pointed out in Sect. 2 and Table 1, have employed rigorous experimental designs that enable the analysis of the trade-off between the use of software and traditional classes.

Another key research challenge is understanding how long the effects of CAL last. One major issue is the difficulty of tracking the same students over time. Additionally, changes in educational context and variations in implementation complicate long-term comparisons. Hirata (2022) addressed this through a study in three Brazilian municipalities, assessing students before, right after, and one year following the intervention. The study showed gains of + 0.56 SD in math in the short term and + 0.17 SD in the medium term, revealing that initial effects fade but meaningful gains persist.

Table 2 Comparative evaluations of AI-Guided tutoring: Designs, Contexts, and educational impacts

Author(s)	Publication	Country	Output Measure	Direction of Effect	Effect Size	Intervention	Comments	Method	Objective	Sample Size
Bastani, Bastani, Sungu, Ge, Kabakcı & Mariman	Wharton School Research Paper (2024)	Turkey	Normalized score on exercises and test	GPT-Tutor significantly improves practice and avoids adverse effects on post-test; GPT-Base improves practice but harms subsequent performance.	GPT-Tutor: +0.361 (practice), −0.004 (test); GPT-Base: +0.137 (practice), −0.054 (test)	Virtual tutor using GPT-4 (standard or pedagogical version) in math sessions with laptops in class	GPT-Tutor provides hints, common errors, and scaffolding; GPT-Base gives complete answers. Negative effect of GPT-Base is more pronounced among low-performing students	RCT with classroom-level random assignment	Assess how the design of interaction with AI tutors (GPT-Base vs. GPT-Tutor) affects performance and autonomous learning	839 students
Henkel, Horne-Robinson, Kozhakhmetova & Lee	Effective and Scalable Math Support (2024). Springer chapter book.	Ghana	Math achievement (standardised test)	Positive and significant. Higher improvement in the group with Rori	0.36 SD increase in scores ($p < 0.001$)	Conversational tutor via WhatsApp. Two 30-min sessions per week with Rori during "study hall"	Works on basic phones. Natural language interaction. Cost: $5 per student. No curricular changes or extra training needed	RCT with school-level assignment	Assess impact of an accessible, scalable AI conversational tutor on primary math learning	477 students with complete data (236 treatment, 241 control)
Thomas, Marsh, Allbright, Johnson & Jennings	Improving Student Learning with Hybrid Human-AI Tutoring. Communcation Conference	USA (PA, CA)	Math achievement and engagement	Positive; higher impact for initially low-performing students	Effect size not reported in SD; statistically significant improvement vs. control	Hybrid tutoring: adaptive software + human tutors informed by system data	Dashboards and a low tutor-student ratio (1:4) allow effective personalisation. Costs ≈ $700/year/student	Three quasi-experimental studies with a control group	Assess whether combining AI and human support improves math achievement and engagement	≈ 500 students in three urban schools

Table 2 (continued)

Author(s)	Publication	Country	Output Measure	Direction of Effect	Effect Size	Intervention	Comments	Method	Objective	Sample Size
De Simone, Ogundeyi, Adesanya, Bello, Adeniran, Bello, Daramola & Jhingran	World Bank Policy Research Working Paper No. 10,747 (Feb 2025)	Nigeria (Benin City)	Standardised test in English, digital, and AI skills	Positive, especially in English and among girls and higher performers	+0.31 SD overall; +0.23 SD in English; +0.031 SD per extra session	Copilot (ChatGPT-4) tutoring after school: 12 × 90-min sessions over 6 weeks, with teacher supervision	No proprietary software or question banks. Cost: $48/student. High educational ROI and viability in constrained settings	RCT with individual random assignment in 9 public schools	Assess the impact of generative AI tutors in real-world, low-income educational contexts	857 eligible students; 447 participants; analysis with 414 complete data
Wang, Ribeiro, Robinson, Loeb & Demszky	Stanford University Working Paper (2024)	USA	Mastery of math content (exit tickets), tutoring, and pedagogical quality	Positive, especially for less experienced tutors	4% points increase in overall mastery ($p < 0.01$); up to 9% points for students with less effective tutors	Tutor CoPilot, real-time pedagogical support via generative AI (GPT-4) for human tutors in virtual tutoring	Effective pedagogical strategies, very low cost ($20/year per tutor). Greater impact for tutors with less experience	Pre-registered RCT with qualitative analysis and NLP on 550,000 messages	Assess whether AI can amplify real-time pedagogical capacity of human tutors, improving outcomes for disadvantaged students	900 tutors and 1,800 students (K-12)

More time on CAL doesn't necessarily mean more learning: beyond a certain threshold, additional usage can lead to diminishing—or even negative—returns. Bettinger et al. (2023) demonstrate that increasing exposure beyond a basic level doesn't always yield better results. Clarifying this issue is vital for guiding educational policy, as it helps define optimal usage levels that minimise waste and maximise learning. However, the effects of intensity may vary depending on the academic setting, software features, and student profiles.

Another challenge about CAL is whether these Programmes complement traditional teaching methods. Scalability may be compromised if intensive teacher supervision is required, which can increase costs and complicate implementation. The question of complementarity remains underexplored due to the lack of precise data and the complexity of designing experiments that measure interactions between CAL and conventional methods. Rodríguez-Segura (2022) and Abbey et al. (2024) emphasise that a lack of adequate teacher training and support can hinder the integration and scalability of CAL tools. Some recent studies have begun to fill this gap. Büchel et al. (2022) compared traditional teacher-led classes, CAL with support staff supervision, and CAL with teacher involvement. Their results suggest that teacher-supervised CAL performs best. Gray-Lobe et al. (2022) analysed a program in Kenya using standardised curricula, detailed teacher guides, and tablets with centralised feedback. Results showed learning gains equivalent to a whole school year. Standardisation, continuous monitoring, and consistent implementation helped teachers use digital tools effectively, highlighting the value of structured integration.

Scalability remains a significant hurdle. Many CAL programs are run by charities rather than governments, especially in developing countries. While NGOs often provide key resources and expertise, these programs tend not to survive once NGOs leave due to limited local capacity (Beg et al. 2023). Without sustainable models, CAL can't become embedded in education systems, leading to only short-term impacts. This is evident when the substantial initial investment by NGOs fades after handover to local authorities. The challenge is to design interventions that work within the existing public education infrastructure. Beg et al. (2023) show that government-led CAL programs using current school personnel can succeed. Their RCT in Ghana found that school principals, when acting as instructional leaders, improved both teaching practices and student learning outcomes using existing resources.

Another key issue is the effectiveness of CAL programs in areas such as reading and writing, where results are not as clear. The math curriculum is particularly well-suited for personalised learning software due to the objective nature of its problems and cognitive processes. However, studies like Escueta et al. (2020) indicate that the impact on other subjects, such as language arts, is minor. The average effect size of randomised experiments with CAL programs in mathematics, as summarised by these authors, is 0.23 SD, equivalent to what a student learns in just over six months of classes. In the case of language arts, reading comprehension, or spelling, the average impact of the reviewed articles is calculated at 0.15 SD, equivalent to just over four months of classes.

In summary, although CAL programs have the potential to revolutionise education, it's necessary to address these challenges to fully understand their mechanisms and maximise their effectiveness in diverse educational contexts.

4.2 Risks of AI-guided Tutors: Misinformation, Algorithmic Bias, and Cognitive Offloading

While CAL programs have already prompted debates around implementation, sustainability, and the need for rigorous impact evaluation, the rapid emergence of AI-guided tutors—including those based on large language models—amplifies existing concerns and presents new systemic challenges. Educational institutions face several risks, including algorithmic biases and a lack of transparency, erosion of socio-emotional skills, growing technological dependency, loss of control over personal data, and the spread of misinformation. Even when outputs from CAL or AI-guided tutoring systems appear well-structured and articulate, they may be inaccurate, incomplete, or reflect underlying biases, posing distinct challenges for student learning and critical engagement.

One possible disadvantage of AI-guided tutors is "cognitive offloading": students excessively delegate comprehension, memory, or reasoning, reducing their active involvement and critical capacity. This phenomenon, already observed with previous technologies like GPS, could worsen if a culture of reflective and metacognitive use of AI tools is not established. Fan et al. (2024) provide evidence of how AI-guided tutor tools can affect self-regulatory learning processes. In a randomised laboratory experiment, four types of learning support were compared during a writing task: (i) a generative AI-based chatbot (ChatGPT), (ii) an expert human tutor, (iii) analytical writing tools, and (iv) a group without additional support. The aim was to analyse differences in intrinsic motivation, self-regulation processes, and task performance. The university students were randomly assigned to each of the four groups, and data were collected on motivation, self-regulatory behaviour, and academic performance. The group working with ChatGPT showed a significant improvement in the quality of the final text, demonstrating that the tool can have immediate positive effects on performance. However, this improvement did not translate into long-term knowledge gains or greater transfer capacity. No significant differences were observed in intrinsic motivation between groups, suggesting that the use of AI does not necessarily lead to increased internal commitment to the task.

Patterns of self-regulated learning differed by type of support received. Students who used ChatGPT showed a lower frequency of metacognitive strategies such as planning, monitoring, and self-evaluation. The authors interpreted this trend as a form of "metacognitive laziness". When students received well-structured answers immediately, they tended to delegate cognitive effort to the tool, reducing their active involvement in the learning process.

From a theoretical standpoint, this phenomenon is connected to the notion of cognitive offloading (Risko and Gilbert 2016), in which individuals externalise mental tasks to reduce cognitive load. While this strategy can be helpful in contexts of overload, it can also weaken internal reasoning abilities when it becomes a habitual approach. In the case of ChatGPT, its ease of use and apparent authority can reduce the perceived threshold of difficulty for students, thereby limiting their willingness to review, question, or rework the information they receive. The use of generative AI may also lower students' perceived challenge, thus restricting the activation of more demanding analytical processes—so-called "System 2" processes in cognitive psychology (Alter et al. 2007).

These results reinforce the importance of designing pedagogical strategies that incorporate AI as a support, rather than a substitute, for students' metacognitive efforts. Fan et al. (2024) recommend, for example, that teachers clearly define which tasks should be carried out with AI help and which require a more autonomous approach. They also propose establishing explicit scaffolding to foster critical reflection on model-generated responses, thus promoting a culture of active and conscious AI use in the classroom. For an AI-based personalised tutor to truly contribute to sustainable learning, it should be integrated into an educational context that reinforces intrinsic motivation, critical thinking, and self-regulation. Otherwise, we risk generating an illusion of competence, where the student improves in specific tasks but loses autonomy and transfer capacity—key elements for lifelong learning.

Oakley et al. (2025) introduce another relevant and little-explored dimension to the debate on generative AI in education: its impact on memory processes and long-term learning consolidation. Drawing on an interdisciplinary review based on cognitive neuroscience, they argue that excessive use of external aids, such as AI tutors, can weaken declarative and procedural memory systems, which are fundamental for the development of internal schemata, expert intuition, and flexible thinking. Oakley et al. (2025) do not present new empirical evidence but synthesise recent findings from the literature on learning and memory. They argue that reliance on tools like ChatGPT can foster cognitive offloading, as discussed—the delegation of mental tasks to external devices—and hinder the formation of robust schemata. This practice compromises the deep encoding necessary for lasting learning, as it limits the use of the declarative system and makes it more challenging to transition to the procedural system, where knowledge is automated and becomes intuitive.

Oakley et al. (2025) link this concern to the reversal of the Flynn effect—the decline in IQ scores in developed countries since the 1970s—suggesting that underuse of internal memory and excessive externalisation of knowledge may be weakening the cognitive structures necessary for complex reasoning and transfer. At a theoretical level, the authors connect this problem with the role of metacognitive effort and the activation of System 2 analytical thinking, which is often inhibited when AI provides complete, frictionless solutions. Consequently, they warn that passive use of AI in educational contexts could compromise the development of deep skills and create an illusion of competence without real understanding.

In line with Fan et al. (2024), who showed a reduction in metacognitive self-regulation among students who used ChatGPT without guidance, Oakley et al. (2025) stress that the real educational value of AI does not lie in providing answers, but in its potential to promote mental effort, active retrieval, and the formation of meaningful connections. Therefore, Oakley et al. (2025) propose that the integration of AI tutors into learning should be accompanied by explicit instructional design that stimulates active student participation, reinforces internal memory, and avoids over-reliance on external resources.

One of the most comprehensive and contextualised proposals for incorporating generative AI (and AI-guided tutors) into developing education systems is put forward by Levy Yeyati et al. (2025). These authors propose a framework for integrating tools based on generative models into classrooms in Latin America, under principles of complementarity, gradualism, and equity. Their study emphasises that any integration of AI in education should consider the structural conditions of the systems, including access inequalities, gender gaps, teachers' training limitations, and the lack of connectivity in many schools. Levy Yeyati et al. (2025) analyse qualitative and quantitative evidence, drawing on data from the computational thinking program and the Ceibal Gender Dashboard (Uruguay) to show usage patterns and adoption inequalities. For example, boys tend to show greater participation in robotics activities. At the same time, female teachers—most of the teaching workforce—have lower rates of AI use, partly due to cultural barriers, self-perceptions, and limited access to training. In response, the authors recommend specific interventions such as training programs aimed at women, gender-sensitive adoption strategies, and scalable hybrid models that combine teacher supervision with generative chatbot assistance. The study concludes that, if applied with pedagogical and institutional care, generative AI can help reduce inequalities, strengthen teacher preparation, and increase student engagement. In terms of implementation, Levy Yeyati et al. (2025) advocate for a progressive integration approach. Their framework is based on the principle that AI should complement, not replace, teachers. The authors suggest starting with teacher-focused uses, such as lesson planning or material generation, and only introducing student-directed applications once appropriate training has been provided. This gradual approach preserves the central role of the teacher and allows the educational community to develop critical ownership of the tools. Their perspective is aligned with a vision of AI as a lever for equity. They stress that its value does not lie in treating everyone the same, but in allowing more personalised responses for those who are usually overlooked: students with learning gaps, those with less verbal participation, or those who require more time to process information. In short, the proposal by Levy Yeyati et al. (2025) reinforces that AI can help improve learning and reduce inequalities only if it is deployed within a robust, context-adapted, and teacher-centred pedagogical approach. The key is not to automate teaching, but to create institutional and training conditions that enable teachers to harness AI's potential to teach better.

5 Digital Divide, Equity, and Barriers to Inclusive CAL and AI-guided Tutors Adoption in Italy and OECD Countries

Having established the potential of CAL and AI-guided tutors to improve learning outcomes, we now examine the critical question of whether these benefits are equitably distributed, or whether existing digital divides risk deepening educational inequalities. The preceding sections have underscored both the transformative potential and significant risks associated with scaling up CAL and AI-guided tutoring systems in schools. Section 4 highlighted how the expansion of these technologies—while promising for personalised learning—carries the risk of amplifying existing inequalities if access to digital infrastructure and quality resources is not ensured for all students. In this context, Italy's position within the broader European landscape offers a critical case study for understanding the structural barriers and policy priorities necessary for an equitable digital transition in education.

Artificial intelligence tools can facilitate the adaptation of content, support students with special educational needs, and provide access to advanced digital resources. However, evidence compiled by the OECD shows that these benefits are far from universal: persistent inequalities in digital infrastructure and resource quality remain a central obstacle to the widespread and equitable integration of AI in education (OECD, 2024).

Table 3 presents the overall index of educational material shortages as reported by school principals in PISA 2022, along with a specific breakdown of shortages in digital resources, including both quantity and quality. The countries included in the table focus on the prominent OECD members geographically close to Italy, as well as those with special comparative relevance in the European and transatlantic context (France, Germany, Italy, Portugal, Spain, the United Kingdom, and the United States).

The Educational Resources Shortage Index (EDUSHORT), used in PISA 2022 (and previous editions), is constructed from principals' responses to question SC017, which asks to what extent various factors hinder the school's ability to provide instruction. Answers are provided on a four-level scale: "not at all," "very little," "to some extent," and "a lot." The EDUSHORT index combines four items referring to both the quantity and quality of educational materials and physical infrastructure and is standardised so that a value of 0 represents the OECD average. Educational materials include textbooks, ICT equipment, a library, laboratory materials, and other resources. Physical infrastructure encompasses the school building, grounds, heating and cooling systems, lighting, and acoustics. Negative values indicate fewer shortages than the average (indicating better resourcing), while positive values signal worse conditions. In this context, Italy (–0.21) shows a resource endowment above the OECD average (as is the case also for France, Spain, or the UK), whereas Portugal (0.24) reports greater shortages.

A key element for interpreting these results is their evolution over time. According to PISA 2022, in approximately half of the participating education systems, school principals reported fewer shortages of educational materials in

Table 3 The extent to which educational and digital resource shortages affect learning in PISA 2022. Results reported by school principals

Country	Educational Resources Shortage Index (EDUSHORT)		Lack of digital resources (%)				Inadequate/low-quality digital resources (%)			
	Mean	Std. dev.	None	Very little	To some extent	A lot	None	Very little	To some extent	A lot
France	–0.40 (0.06)	0.88 (0.04)	46.4 (3.8)	30.5 (3.2)	19.7 (2.9)	3.5 (1.1)	46.2 (3.6)	31.2 (3.5)	16.5 (2.8)	6.1 (1.4)
Germany	–0.07 (0.07)	1.00 (0.04)	27.2 (3.3)	34.5 (3.4)	27.4 (3.5)	10.9 (2.3)	25.2 (3.3)	37.8 (3.4)	26.7 (3.4)	10.3 (2.5)
Italy	–0.21 (0.07)	0.93 (0.05)	50.8 (4.0)	35.6 (3.6)	12.1 (2.4)	1.5 (0.8)	48.6 (4.1)	37.1 (3.9)	12.7 (2.4)	1.7 (1.0)
Portugal	0.24 (0.06)	0.98 (0.05)	30.4 (3.4)	40.4 (3.5)	23.4 (3.3)	5.8 (1.4)	22.7 (2.9)	37.8 (3.1)	29.9 (3.1)	9.6 (1.8)
Spain	–0.29 (0.04)	1.06 (0.04)	48.3 (2.4)	24.7 (2.1)	20.4 (1.9)	6.5 (1.2)	45.6 (2.3)	30.0 (2.2)	18.2 (1.6)	6.2 (1.1)
United Kingdom	–0.32 (0.06)	0.87 (0.04)	38.3 (4.0)	42.7 (4.4)	14.3 (2.7)	4.7 (1.6)	39.1 (4.0)	39.7 (4.3)	17.5 (3.1)	3.7 (1.4)
United States	–0.66 (0.09)	0.92 (0.07)	76.5 (3.9)	16.9 (3.2)	5.1 (2.2)	1.5 (1.1)	73.6 (4.1)	17.0 (3.2)	8.0 (2.7)	1.4 (0.9)
OECD Average	–0.17 (0.01)	0.97 (0.01)	47.2 (0.5)	28.8 (0.5)	16.1 (0.4)	7.8 (0.3)	45.5 (0.5)	29.9 (0.5)	17.1 (0.4)	7.5 (0.3)

Source: OECD (2023), Table II.B1.5.17 PISA 2022 Results (Volume II): Learning During – and From – Disruption

2022 compared to 2018. This improvement was particularly significant in countries such as Ireland, Indonesia, Croatia, Spain, and, notably, Italy. However, shortages of educational staff were perceived as more acute in most countries.

Beyond the global index, Table 3 disaggregates digital resource information along two specific dimensions:

(i) the lack of digital resources, such as computers, tablets, internet access, or school digital platforms.
(ii) the presence of inadequate or low-quality digital resources.

In both cases, the table shows the percentage of students whose principals report that instruction is hindered “not at all,” “very little,” “to some extent,” or “a lot.”

Specifically, Table 3 is based on indicators from question SC017 of the PISA 2022 School Questionnaire. Items SC017Q09JA and SC017Q10JA focus on:

- SC017Q09JA: Lack of digital resources (e.g., computers, tablets, internet, Learning Management Systems such as Google Classroom, Moodle, or school digital platforms).
- SC017Q10JA: Inadequate or low-quality digital resources (same examples as above).

In the Italian context, 50.8% of students are enrolled in schools where principals report that lack of digital resources does not hinder instruction at all, and 35.6% where it is reported as “very little”. Only 13.6% of Italian students are in schools where digital shortages hinder teaching “to some extent” or “a lot”—well below the OECD average (23.9%). Similarly, regarding the quality of digital resources, 48.6% of students are in schools with no perceived problems. Only 14.4% are in schools where quality issues are reported as “to some extent” or “a lot.” This places Italy among the countries with the lowest reported barriers to digital resource quality, comparable to the situation in France and the UK, and significantly different from the levels observed in Portugal or Germany. Portugal shows that almost a third of its students attend schools with moderate or severe digital quality problems. The United States stands out as a benchmark, with over 76% of students in schools reporting no lack of digital resources and over 73% reporting no quality deficiencies, underscoring the North American advantage in both access and quality.

The variability of resource allocation remains an issue: although Italy as a whole is above average in digital resources, the SD (0.93) indicates a moderate but not insignificant degree of inequality across Italian schools, ranking it less than Spain but more than the UK or France. This analysis suggests that Italy is relatively well-positioned within the European context in terms of the digital foundations required for the inclusive adoption of CAL or AI-guided Tutors in education. Nonetheless, ensuring that all schools and students—including those in rural or disadvantaged areas—have access to adequate digital tools and infrastructure is essential for preventing the deepening of educational divides as CAL and AI-guided learning becomes more widespread.

The deployment of CAL and AI in education has the potential to amplify existing gaps between schools, students, and communities. Factors such as device availability, connectivity quality, teacher training, and the ability of institutions to integrate emerging technologies determine whether all students can benefit equally. The "unregulated" adoption of CAL and AI tools can accelerate polarisation: schools with more resources can access and implement innovations earlier and more effectively. At the same time, less advantaged institutions are left behind, both in opportunities and outcomes. In addition to these material challenges, other critical issues emerge, such as the need to reinforce educational integrity against commercial pressures and the importance of equipping teachers with the skills for responsible AI use.

An equitable educational environment in Italy and other OECD countries requires not only physical access to technology but also strong institutional support, ongoing professional development for teachers, and a clear ethical framework to protect the most vulnerable students. One of the primary obstacles to the adoption of inclusive CAL and AI use in education is the persistence of material inequalities in access to basic digital resources. Far from having been mitigated in recent years, these inequalities continue to disproportionately affect the most vulnerable students, as shown by PISA 2022.

Table 4 illustrates the percentage of students in Italy whose principals report shortages of digital resources—such as computers, internet connectivity, or

Table 4 Percentage of students in schools with digital resource shortages, by school socioeconomic status and ownership

Country	All students (1)	Disadvantaged schools (2)	Average SES schools (3)	Advantaged schools (4)	Difference (disadvantaged – advantaged) (5)	Public schools (6)	Private schools (7)	Difference (private – public) (8)
France	23.2 (3.0)	22.0 (5.6)	28.4 (4.8)	13.8 (4.5)	–8.2 (6.9)	21.8 (3.3)	27.9 (6.7)	6.1 (7.4)
Germany	38.3 (3.6)	39.8 (7.3)	41.3 (5.0)	31.6 (7.1)	–8.2 (10.7)	39.3 (3.6)	9.5 (10.3)	–29.8* (10.7)
Italy	13.6 (2.5)	14.3 (5.2)	13.6 (3.3)	13.0 (5.6)	–1.3 (7.6)	13.3 (2.6)	21.3 (10.7)	8.0 (10.9)
Portugal	29.2 (3.2)	26.7 (5.8)	34.4 (4.8)	21.5 (6.1)	–5.2 (8.3)	32.0 (3.7)	13.6 (5.7)	–18.3* (7.4)
Spain	27.0 (2.0)	31.6 (4.8)	30.3 (2.8)	16.1 (3.3)	–15.5* (5.9)	29.5 (2.4)	21.6 (3.0)	–7.9* (3.7)
United Kingdom	19.0 (3.1)	26.6 (7.5)	19.7 (3.9)	12.7 (5.1)	–13.9 (8.8)	26.8 (4.8)	15.1 (3.9)	–11.7 (6.3)
United States	6.6 (2.4)	8.0 (6.4)	5.6 (2.6)	5.8 (3.9)	–2.1 (7.4)	5.2 (1.8)	—	—
OECD Average	23.9 (0.4)	27.8 (0.9)	22.4 (0.6)	18.3 (0.9)	–9.5* (1.3)	26.0 (0.6)	13.4 (1.0)	–13.5* (1.2)

Note: School socioeconomic profile is defined using the PISA ESCS index: disadvantaged schools are in the lowest quartile and advantaged schools are in the top quartile within each country. Statistically significant differences are marked in bold with an asterisk (*)

Source: OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, Table II.B1.5.19, https://doi.org/10.1787/a97db61c-en learning management platforms—broken down by the school's socioeconomic status and by sector (public or private). The PISA index of Economic, Social and Cultural Status (ESCS) is a composite measure summarising the socioeconomic and cultural environment of the student's family. This index is constructed from three principal dimensions. First, the highest level of education reached by either parent, coded according to the ISCED-2011 international classification and converted into years of schooling (PAREDINT). Second, the highest occupational status among parents (HISEI), derived from ISCO-08 codes and assigned to the international socio-economic status index (ISEI), reflecting social position linked to occupation beyond direct income. Third, the HOMEPOS index, which records the presence of educational, technological, and cultural assets in the home (such as the number of books, a computer for school use, internet connection, own desk, calculator, literature books, reference books, washing machine, dishwasher, and other country-specific items). The variables are combined into an index using principal component analysis and adapted culturally in each country. The final socioeconomic index is standardised so that 0 represents the OECD average, and SD is equal to 1. Thus, a positive value indicates a more advantaged context than the OECD average, while a negative value indicates a more disadvantaged background.

In all countries analysed, including Italy, schools serving socioeconomically disadvantaged students report digital resource shortages more frequently than those serving more advantaged populations. Table 4 reports the percentage of students attending schools where principals state that instruction is hindered "to

Table 5 Percentage of students in schools with inadequate or low-quality digital resources, by school socioeconomic status and ownership

Country	All students (1)	Disadvantaged schools (2)	Average SES schools (3)	Advantaged schools (4)	Difference (disadvantaged – advantaged) (5)	Public schools (6)	Private schools (7)	Difference (private – public) (8)
France	22.6 (3.0)	26.7 (6.1)	25.8 (4.4)	11.8 (4.8)	–14.9 (7.2)	23.1 (3.4)	20.8 (5.9)	–2.3 (6.7)
Germany	37.0 (3.3)	40.9 (7.5)	38.4 (5.0)	31.8 (7.3)	–9.1 (11.0)	38.4 (3.3)	0.0 (c)	–38.4 (3.3)
Italy	14.3 (2.6)	15.1 (6.0)	14.5 (3.5)	13.3 (4.9)	–1.8 (7.6)	14.3 (2.7)	14.2 (9.7)	–0.1 (9.9)
Portugal	39.5 (3.4)	33.9 (6.6)	44.0 (4.5)	36.0 (6.7)	2.1 (9.5)	43.9 (3.7)	14.2 (5.8)	–29.7 (7.1)
Spain	24.4 (1.8)	26.6 (4.5)	29.3 (2.7)	12.7 (2.4)	–13.9 (5.1)	28.9 (2.5)	14.9 (2.4)	–14.0 (3.4)
United Kingdom	21.2 (3.2)	27.9 (8.2)	23.9 (4.6)	12.5 (4.8)	–15.4 (9.2)	27.9 (5.5)	18.2 (3.9)	–9.6 (6.8)
United States	9.4 (2.9)	9.2 (6.5)	11.7 (4.2)	3.5 (3.2)	–5.8 (7.2)	8.1 (2.5)	—	—
OECD Average	24.6 (0.5)	28.0 (1.0)	23.1 (0.6)	19.5 (0.9)	–8.5 (1.3)	26.2 (0.5)	15.0 (1.0)	–12.0 (1.1)

Note: Socioeconomic profile defined as above. Statistically significant differences are in bold with an asterisk (*)

Source: OECD (2023), PISA 2022 Results (Volume II), Table II.B1.5.20

some extent" or "a lot" by a lack (Table 4) or poor quality (Table 5) of digital resources. Both indicators enable us to assess how digital infrastructure affects equity in access to learning opportunities, particularly in the context of the increasing use of CAL and AI tools in education.

In Italy, 14.3% of students in disadvantaged schools are in institutions where digital resource shortages hinder instruction, compared to 13.0% in advantaged schools. While the gap between disadvantaged and advantaged schools is smaller than the OECD average (–1.3% points in Italy, compared to −9.5 for the OECD), these differences still highlight a structural inequality that affects the ability of Italian schools to incorporate technology tools on an equitable basis. Looking at ownership, 13.3% of students in public schools and 21.3% in private schools in Italy face digital resource shortages. This pattern differs from other OECD countries, where public schools are often at a clear disadvantage.

Turning to the quality of digital resources (Table 5), 15.1% of students in disadvantaged schools in Italy face inadequate or low-quality digital resources, compared to 13.3% in advantaged schools—a difference of −1.8 points. For public versus private schools, the difference is virtually nil (14.3% in public schools, 14.2% in private schools). These data confirm that, while Italy exhibits lower levels of digital deprivation than many of its OECD peers, challenges remain in specific segments of the system.

According to principals' reports, Italy exhibits a significant rural-urban gap in the availability of digital resources in schools. Full breakdowns by school location (rural, town, city) are reported in Appendix Tables A1–A2. The percentage of students attending schools located in rural areas or villages (fewer than 3,000 people), with inadequate or poor-quality digital resources, reaches 47.5%, compared to just 9.6% in schools located in cities (over 100,000 people). This results in a striking gap of −37.9% points between rural and urban schools. In contrast, the OECD average for rural schools is 31.7%, and for city schools, it is 23.1%, with a much smaller gap of −8.8% points. These figures indicate that, in Italy, the rural-urban divide in digital resources is not only significantly above the OECD average, but is also substantially larger than other gaps, such as those related to school socio-economic profile ot type of school (public-private), or the concentration of immigrant students.

Tables 4 and 5 show that Italian schools serving more disadvantaged students are more likely to lack digital resources or to have resources of lower quality. This raises the question of whether these material limitations also influence students' learning outcomes. This issue is especially relevant in the current context, where the effective and equitable integration of CAL and AI tools in education depends on a minimum digital infrastructure.

PISA 2022 data show that, on average, students attending schools where principals report shortages of educational resources—including digital tools—achieve lower results in mathematics, the main focus area of PISA 2022, across OECD countries. The OECD report highlights the connection between resource scarcity in educational institutions and students' performance in mathematics. Specifically, it highlights the average difference in test scores between students in schools where principals state that the ability to provide instruction is hindered—

"to some extent" or "a lot"—by the lack or low quality of resources (whether material or human), and those attending schools without such limitations ("very little" or "not at all"). Building on this general analysis of resource shortages, it is essential to focus on the specific dimension at the heart of this article: digital resources. Their availability and quality are not only critical for the ordinary functioning of schools but also represent the fundamental basis required for the effective and equitable deployment of CAL and AI-based tools in the education system.

PISA 2022 show that students attending schools with fewer digital resource shortages tend to achieve better mathematics results, on average, across OECD countries. While the adverse effects of these shortages tend to disappear when controlling for the socioeconomic profile of both students and schools, the evidence underscores that more vulnerable contexts not only have less access to resources, but that these deficits are directly associated with lower academic performance. Resources are not evenly distributed; they fall disproportionately on those already at a disadvantage. Furthermore, even though principals in Italy in 2022 expressed less concern about shortages of educational materials compared to 2018, internal variability between schools within the country remains high.

6 Examining Equity in Access To ChatGPT in Italy: Evidence from Google Trends

Italy provides a compelling case to examine these dynamics in practice: despite having relatively strong digital infrastructure at the national level, substantial disparities persist across regions and schools. We use new data to assess whether technological innovations in education are bridging or reinforcing these gaps. In this section, we empirically analyse equity in the use of artificial intelligence for educational purposes in Italy. Specifically, we conduct an econometric analysis to identify the relevant differences in the intensity of ChatGPT usage across time and regions. It is important to note that Google Trends data on ChatGPT searches do not directly capture educational usage. Instead, we interpret them as a proxy for the adoption of AI tools for educational purposes, reflecting broader patterns of engagement with generative AI across regions.

Bacher-Hicks et al. (2021) utilise internet search data to study, in real-time, how US households sought online learning resources when schools closed during the COVID-19 pandemic. They conclude that national search intensity for online learning resources doubled compared to pre-pandemic levels. However, areas with higher income, better internet access, and fewer rural schools experienced significantly larger increases in search intensity. As a result, the authors suggest that the pandemic likely widened academic achievement gaps, as schools and families interacted differently with online resources to compensate for lost classroom time.

Our econometric analysis focuses on ChatGPT, the most widely used AI tool in Italy for work and education. The regression specification follows the model in Bacher-Hicks et al. (2021):

$$IS_{rmt} = \sum\nolimits_{m=1}^{12-1} \beta_m\, Month_m + \sum\nolimits_{t=1}^{4-1} \beta\,_t Year_t + \sum\nolimits_{r=1}^{20-1} \beta_r\, Region_r + \beta_g \ln(GDP\, per\, capita_r) + \sum\nolimits_{a=1}^{3-1} \beta_{gy} \ln(GDP\, per\, capita_r) * Year_t \tag{1}$$

Where IS_{rmt} is the search intensity for ChatGPT in region *r*, month *m* and year *t*. Here, IS denotes the Index of Searches (search intensity). *Month* and *Year* are sets of month and year dummies, respectively, and *Region* is a set of region fixed effects. $\ln(GDP\, percapita_r)$ is the log of GDP per capita of *Region r* in 2023 (relative to the Italian average, from EUROSTAT), which remains constant across the three years of observation, and $\ln(GDP\, percapita_r) * Year_t$ is its interaction with year dummies to capture heterogeneous adoption patterns over time. The regional GDP per capita for 2023 is the most recent data available from Eurostat.

As reported in Model 1 of Table 6, Molise is omitted from the regressions, as it is the region with the lowest observed intensity of ChatGPT searches for education and employment. Similarly, August is omitted as the reference month, reflecting its role as the period with the lowest use of ChatGPT for both education and employment. In Models 2 and 3, Molise and Aosta Valley (the second region with the lowest search intensity) are excluded due to collinearity, which arises because GDP per capita is time-invariant across 2022–2025. The dummies for each of the years 2023, 2024, and 2025 capture changes in ChatGPT search intensity relative to all other searches, with 2022 serving as the reference year.[1]

First, the results indicate substantial territorial disparities. In Model 1, almost all Italian regions display large and statistically significant positive coefficients relative to Molise, the region with the lowest observed intensity of ChatGPT searches for education and employment. For instance, Campania (+ 14.96), Sicily (+ 12.98), Lazio (+ 12.86) and Apulia (+ 12.82) show markedly higher search activity compared to the baseline. Aosta Valley, by contrast, records a coefficient that is not statistically significant. At the same time, Basilicata, Liguria and Umbria exhibit positive but comparatively minor values. This pattern suggests that the diffusion of ChatGPT was weakest in Molise and, to a lesser extent, Aosta Valley.

Second, the results reveal apparent seasonality in ChatGPT searches. The summer months and the end of the academic year—June, July, and August—show significantly negative coefficients relative to December, the baseline month. This pattern aligns with the academic calendar, featuring breaks in the summer and over Easter, when demand for educational tools typically decreases. In contrast, autumn months such as October and November exhibit somewhat less harmful or even positive coefficients, which may be linked to the start of the school year and increased academic activity.

Third, the models show robust growth in AI-related searches for education and employment since 2022. The year dummies for 2023, 2024, and 2025 are increasingly positive and statistically significant, indicating an upward trajectory in ChatGPT adoption over time.

[1] Data for 2025 cover the period up to May only.

Table 6 Effect on the intensity of ChatGPT use by region, month and year (Italy). *Model with regional, temporal, and GDP per capita variables* (Standard errors in parentheses. Reference categories: Molise in Model 1, Molise and Aosta Valley in Models 2 and 3, August, 2022)

	ChatGPT Searches	ChatGPT Searches	ChatGPT Searches
	Model 1(Region, Year, Month)	Model 2 + ln(GDP per capita)	Model 3 + ln(GDP*Year)
Abruzzo	6.65*** (1.01)	6.47*** (0.90)	6.47*** (0.89)
Apulia	12.82*** (1.01)	12.97*** (1.15)	12.97*** (1.14)
Basilicata	2.04** (1.01)	2.01** (0.98)	2.01** (0.97)
Calabria	11.15*** (1.01)	11.41*** (1.28)	11.41*** (1.26)
Campania	14.96*** (1.01)	15.13*** (1.17)	15.13*** (1.15)
Sardinia	6.34*** (1.01)	6.36*** (1.02)	6.36*** (1.01)
Emilia-Romagna	10.45*** (1.01)	9.90*** (0.95)	9.90*** (0.94)
Friuli-Venezia Giulia	6.69*** (1.01)	6.29*** (0.88)	6.29*** (0.87)
Lazio	12.86*** (1.01)	12.35*** (0.93)	12.35*** (0.92)
Liguria	4.86*** (1.01)	4.46*** (0.89)	4.46*** (0.88)
Lombardy	11.47*** (1.01)	10.77*** (1.07)	10.77*** (1.06)
Marche	9.28*** (1.01)	9.04*** (0.88)	9.04*** (0.87)
Molise (ref.)	—	—	—
Piedmont	11.36*** (1.01)	11.00*** (0.88)	11.00*** (0.87)
Sicily	12.98*** (1.01)	13.17*** (1.19)	13.17*** (1.17)
Tuscany	11.11*** (1.01)	10.71*** (0.88)	10.71*** (0.87)
Trentino-Alto Adige	8.64*** (1.01)	7.97*** (1.05)	7.97*** (1.03)
Umbria	5.11*** (1.01)	4.97*** (0.91)	4.97*** (0.90)
Aosta Valley (ref.)	0.62 (1.01)	—	—
Veneto	11.53*** (1.01)	11.05*** (0.91)	11.05*** (0.90)
Year 2022 (ref.)	—	—	—
Year 2023	6.62*** (0.77)	6.62*** (0.77)	–2.52 (11.61)
Year 2024	13.43*** (0.77)	13.43*** (0.77)	14.94 (11.62)
Year 2025	34.68*** (0.87)	34.68*** (0.87)	89.60*** (12.45)
January	0.19 (0.86)	0.19 (0.86)	0.19 (0.85)
February	0.48 (0.87)	0.48 (0.87)	0.48 (0.86)
March	1.00 (0.85)	1.00 (0.85)	1.00 (0.84)
April	0.92 (0.86)	0.92 (0.86)	0.92 (0.85)
May	5.15*** (0.87)	5.15*** (0.87)	5.15*** (0.86)
June	5.20*** (0.85)	5.20*** (0.85)	5.20*** (0.84)
July	1.42 (0.92)	1.42 (0.92)	1.42 (0.91)
August (ref.)	—	—	—
September	4.92*** (0.92)	4.92*** (0.92)	4.92*** (0.91)
October	7.24*** (0.90)	7.24*** (0.90)	7.24*** (0.89)
November	9.24*** (0.90)	9.24*** (0.90)	9.24*** (0.89)
December	8.54*** (0.86)	8.54*** (0.86)	8.54*** (0.85)
ln(GDP)	—	1.14 (1.85)	2.77 (2.94)
ln(GDP)*2023	—	—	2.03 (2.57)
ln(GDP)*2024	—	—	–0.34 (2.57)
ln(GDP)*2025	—	—	–12.19*** (2.76)
Constant	–17.22*** (1.22)	–22.13** (8.51)	–29.43** (13.34)
Adjusted R^2	0.628	0.628	0.637
N	2,800	2,800	2,800

*$p<0.10$, **$p<0.05$, ***$p<0.01$

Second, the results reveal clear seasonal dynamics. With August set as the reference month—the period of lowest search intensity—virtually all other months display positive and statistically significant coefficients. May (+5.15), June (+5.20), September (+4.92), October (+7.24), and November (+9.24) stand out with strong increases, reflecting peaks in educational and employment-related activity around the start and end of the academic year. By contrast, January through April show coefficients close to zero and not statistically significant, consistent with a gradual recovery of activity following the winter break. This pattern underscores the alignment of ChatGPT searches with the academic and work calendar.

Third, the year effects point to robust growth in the adoption of ChatGPT over time. Relative to the baseline year 2022, the coefficients for 2023 (+6.62), 2024 (+13.43), and 2025 (+34.68) are increasingly positive and highly significant in Model 1, suggesting an upward trajectory in the use of generative AI for education and employment. Even when controlling for regional income in Models 2 and 3, the results confirm a strong expansion, with 2025 showing the largest increase in search intensity. These findings document both the seasonality and the rapid diffusion of ChatGPT across Italy in the first years following its release.

In Model 2 (Table 6), the log of regional GDP per capita for 2023 is introduced as a control variable. The coefficient is positive but small and statistically insignificant (+1.14), which indicates that once regional, temporal, and seasonal fixed effects are accounted for, differences in regional income do not explain variation in ChatGPT search intensity. Still, the sign of the coefficient suggests that higher-income regions may have had slightly higher adoption, consistent with the expectation that wealthier areas are early adopters of new technologies. The absence of a significant effect suggests that disparities in ChatGPT adoption in Italy cannot be explained by income levels, but are more strongly shaped by regional characteristics and structural factors beyond GDP.

Model 3, shown in Table 6, explores this issue further by interacting ln(GDP per capita) with year dummies to test whether adoption trajectories diverged systematically between richer and poorer regions over time. The coefficient of log of GDP per capita remains positive, although not statistically significant. The interaction term for 2023 is positive (+2.03) but not statistically significant, suggesting that in the initial phase of adoption higher-income regions were somewhat more active, though not in a robust way. Similarly, the 2024 interaction is close to zero (–0.34), reinforcing the absence of systematic differentiation in the middle phase. By contrast, in 2025 the interaction term turns negative and strongly significant (–12.19), indicating that the relative growth of ChatGPT searches in wealthier regions slowed down compared to poorer ones. This dynamic points to a process of convergence: while higher-income areas likely led the way in the early stages of adoption, lower-income regions subsequently accelerated their uptake, reducing the initial digital divide.

Taken together, these findings from Models 2 and 3 reinforce the idea that ChatGPT adoption in Italy is characterised by rapid growth, strong seasonal cycles, and pronounced territorial differences, but also by an underlying tendency towards convergence. Initial inequalities in access and use—driven by socioeconomic factors and regional disparities—appear to diminish as the technology becomes more widely diffused, suggesting that barriers to adoption weaken over time.

6.1 Discussion

When carefully and contextually implemented, computer-assisted learning (CAL) and AI-guided tutoring can deliver personalised instruction and timely feedback, with especially strong results in mathematics. The magnitude and persistence of these gains, however, hinge on integration into everyday teaching practice—curricular alignment, teacher training, and school-level routines. A central unresolved issue concerns the underlying mechanism of impact: learning gains may reflect the software's pedagogical value or simply additional time spent on task. Recent experimental designs that hold total instructional time constant or compare extra lessons with software use suggest genuine value added from CAL, particularly when teachers are directly involved (Büchel et al. 2022; Hirata, 2022), but more evidence is needed to pin down the marginal contribution relative to expanded traditional instruction.

Durability is another open question. Short-run effects are often sizeable, yet some fade in the medium term while specific competencies persist (Hirata, 2022). This is likely to depend on whether implementations elicit retrieval, explanation, and self-monitoring rather than passive practice. Subject heterogeneity also matters. The most robust gains appear in mathematics; in reading and writing, effects are more mixed, consistent with evidence that language-rich tasks place greater demands on design and teacher mediation (Escueta et al. 2020). These considerations also extend to AI-guided tutors, where interaction design is pivotal. Interfaces that scaffold with hints and guided questions foster active reasoning, whereas answer-giving configurations risk cognitive offloading and weaker transfer, with heterogeneous effects across students (Bastani et al. 2024; Fan et al. 2024).

Intensity of use does not map linearly into learning. While practice time is predictive up to a point—as shown in mastery-based deployments paired with coaching—returns can diminish with unstructured or excessive exposure (Oreopoulos et al. 2024; Bettinger et al. 2023). Identifying the dosage that is appropriate for the grade, subject, and learner profile should guide classroom routines and programme design. Evidence also points to the importance of human-in-the-loop models. CAL and AI are most effective when embedded in teacher-led instruction, with educators monitoring progress and providing targeted support; teacher coaching and real-time, tutor-facing AI can raise instructional quality at relatively low cost (Büchel et al. 2022; Oreopoulos et al. 2024; Wang et al. 2024). Hybrid approaches that combine algorithmic personalisation with human tutoring show promising impacts and improved scalability relative to purely human models (Bhatt et al. 2024; Thomas et al. 2024).

Finally, equity and governance remain preconditions for success at scale. Unequal access to devices and connectivity, variability in school capacity, and uneven teacher preparation can widen gaps even when average effects are positive. Appropriate safeguards for data protection, transparency about limitations, and bias monitoring are essential to ensure benefits accrue to the students who need them most. These points align with our Italian evidence: the diffusion patterns of generative-AI use correlate with structural divides, and thoughtful policy is necessary to ensure that implementation reaches disadvantaged schools and regions. In sum, the promise of CAL and AI-guided tutors will be realised where design elicits active learning, institutional support is sustained, teachers are equipped to integrate tools effectively, and governance addresses risks.

7 Conclusions

This article has reviewed the state of the art in the use of computer-assisted learning (CAL) and AI-guided tutors, drawing on recent causal evidence from large-scale experimental and quasi-experimental studies. The results demonstrate that both CAL and new AI-driven approaches can generate meaningful learning gains, especially in mathematics and for students who are most at risk of falling behind. However, their effectiveness depends critically on thoughtful implementation—particularly the integration of adaptive technology with structured pedagogical support and sustained teacher engagement. Hybrid models that combine algorithmic personalisation with human tutoring appear especially promising for reconciling scalability with educational quality.

At the same time, our analysis highlights the risks and unresolved questions that accompany the rapid expansion of educational technology. These include potential cognitive offloading, the need for long-term impact evaluations, and the risk of deepening inequalities if access to high-quality digital resources remains uneven. The evidence from Italy and other OECD countries suggests that material shortages—such as a lack of devices or connectivity—remain concentrated in disadvantaged schools and rural areas, limiting the potential for technology to foster equity unless these gaps are addressed through sustained investment and policy attention.

The main research gaps identified in the CAL literature concern whether learning gains derive from the software itself or from additional instructional time; whether positive effects are sustained in the medium and long term; and how effective CAL is in subjects beyond mathematics, such as reading and writing. There is also limited evidence on the optimal intensity of use, since more time with CAL does not necessarily lead to more learning, and on how these programs complement traditional teaching methods. Recent evidence suggests that teacher-supervised CAL is more effective than programs monitored by assistants, underscoring the importance of integrating it into classroom practice. Addressing these questions is essential to maximise the benefits of CAL in diverse educational contexts.

Beyond these research gaps, policy implications are also central. Educational policies should support the long-term sustainability of CAL by investing in adaptive software capable of personalising instruction across multiple subjects, not just mathematics. They should also promote continuous teacher training so that educators are equipped to integrate CAL effectively with traditional pedagogical practices. Collaboration between software developers and education professionals can help design tools that genuinely respond to classroom needs. In terms of personalisation and equity, CAL tools are particularly relevant to ensure that no student is left behind, offering individualised support that adapts to the diverse learning levels and needs detected by teachers.

Our empirical analysis of ChatGPT adoption in Italy—based on Google Trends searches used as a proxy for educational use—indicates that, although initial digital divides existed between regions of different income levels, the spread of generative AI tools has become more equitable over time. However, this measure remains only a general proxy of AI adoption in education, not a direct observation of classroom practices. Moreover, important regional divides persist, with southern and smaller regions showing systematically lower search intensity than northern regions, such as Veneto. These structural differences underscore the need for targeted investment to ensure that the benefits of technological adoption are more evenly distributed across the country.

In sum, the transformative potential of CAL and AI-guided tutors in education will only be realised if their deployment is accompanied by robust institutional support, ongoing research into their mechanisms and long-term effects, and a deliberate focus on digital equity. Future educational policy should prioritise not just access to devices, but the development of adaptive, high-quality content and the professional development of teachers, ensuring that no student is left behind as digital transformation accelerates in schools.

Appendix

Appendix Table A1 Percentage of students in schools whose principal reported a lack of digital resources, by school location (PISA 2022)

Country	All students (1)	Rural area or village (fewer than 3 000 people) (2)	Town (3 000 to 100 000 people) (3)	City (over 100 000 people) (4)	City – Rural (5)
France	23.2 (3.0)	34.9 (17.3)	23.5 (3.5)	17.5 (6.4)	–17.5 (18.0)
Germany	38.3 (3.6)	35.7 (27.3)	35.1 (3.8)	45.6 (7.5)	9.9 (28.1)
Italy	13.6 (2.5)	48.8 (33.0)	15.5 (3.1)	8.0 (3.7)	–40.8 (33.2)
Portugal	29.2 (3.2)	19.3 (18.2)	30.4 (3.6)	27.1 (6.7)	7.8 (19.0)
Spain	27.0 (2.0)	23.6 (7.2)	31.5 (3.1)	20.9 (3.1)	–2.7 (8.1)
United Kingdom	19.0 (3.1)	39.1 (14.0)	19.3 (4.0)	15.1 (4.9)	–24.0 (15.2)
United States	6.6 (2.4)	9.6 (9.9)	7.0 (2.9)	5.8 (3.9)	–3.8 (10.6)
OECD average	23.9 (0.4)	30.4 (2.2)	24.9 (0.7)	22.7 (0.8)	–7.9 (2.3)

Notes: Results are based on principals' reports to PISA 2022 School Questionnaire item SC017 ("To what extent is your school's capacity to provide instruction hindered by the following?"). "Lack of digital resources" corresponds to SC017Q09JA (e.g., computers, Internet access, learning-management systems). Percentages refer to students in schools where instruction is hindered **"to some extent" or "a lot"**

Source: OECD (2023), *PISA 2022 Results (Volume II): Learning During – and From – Disruption*, Table II.B1.5.19

Appendix Table A2 Percentage of students in schools whose principal reported inadequate or poor-quality digital resources, by school location (PISA 2022)

Country	All students (1)	Rural area or village (fewer than 3 000 people) (2)	Town (3 000 to 100 000 people) (3)	City (over 100 000 people) (4)	City – Rural (5)
France	22.6 (3.0)	37.0 (17.4)	23.3 (3.4)	15.1 (6.0)	–21.9 (17.9)
Germany	37.0 (3.3)	54.3 (21.0)	36.3 (3.9)	37.1 (7.7)	–17.2 (22.4)
Italy	14.3 (2.6)	47.5 (33.1)	15.9 (3.0)	9.6 (4.1)	–37.9 (33.3)
Portugal	39.5 (3.4)	19.3 (18.2)	42.7 (4.0)	32.5 (6.4)	13.2 (19.2)
Spain	24.4 (1.8)	35.7 (7.3)	26.8 (2.7)	20.1 (3.1)	–15.6 (8.4)
United Kingdom	21.2 (3.2)	40.6 (14.0)	21.3 (3.9)	17.8 (5.2)	–22.8 (15.2)
United States	9.4 (2.9)	9.6 (9.9)	13.3 (4.2)	4.4 (3.6)	–5.3 (10.5)
OECD average	24.6 (0.5)	31.7 (2.1)	25.7 (0.7)	23.1 (0.8)	–8.8 (2.3)

Notes: Results are based on principals' reports to PISA 2022 School Questionnaire item SC017 ("To what extent is your school's capacity to provide instruction hindered by the following?"). "Inadequate or poor-quality digital resources" corresponds to SC017Q10JA (e.g., computers, Internet access, LMS). Percentages refer to students in schools where instruction is hindered "to some extent" or "a lot

Source: OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, Table II.B1.5.20

Data Availability All data used in this study are publicly available. PISA microdata can be accessed and downloaded from the OECD (https://www.oecd.org/pisa/data/), and Google Trends regional search intensity data are accessible at https://trends.google.com.

References

Abbey C, Green E, Mo D, Lai F, Bai Y, Zhang L, Bianchi N, Ma Y, Feng Y, Clark T, Fafchamps M, Yang S (2024) The effectiveness of EdTech on student learning outcomes in china: A systematic review and meta-analysis. Comput Educ Open 6:100161

Alter AL, Oppenheimer DM, Epley N, Eyre RN (2007) Overcoming intuition: metacognitive difficulty activates analytic reasoning. J Exp Psychol Gen 136:569–576. https://doi.org/10.1037/0096-3445.136.4.569

Bacher-Hicks A, Goodman J, Mulhern C (2021) Inequality in household adaptation to schooling shocks: Covid-induced online learning engagement in real time. J Public Econ 193:104345. https://doi.org/10.1016/j.jpubeco.2020.104345

Barrow L, Markman L, Rouse CE (2009) Technology's edge: the educational benefits of computer-aided instruction. Am Econ J Econ Policy 1:52–74. https://doi.org/10.1257/pol.1.1.52

Bastani H, Bastani O, Sungu A, Ge H, Kabakcı Ö, Mariman R (2024) Generative AI can harm learning. The Wharton School Research Paper. SSRN. https://ssrn.com/abstract=4895486

Beg SA, Fitzpatrick AE, Lucas A (2023) Managing to learn. NBER Working Paper No. 31757. https://doi.org/10.3386/w31757

Bettinger E, Fairlie R, Kapuza A, Kardanova E, Loyalka P, Zakharov A (2023) Diminishing marginal returns to computer-assisted learning. J Policy Anal Manage 42:552–570. https://doi.org/10.1002/pam.22442

Bhatt R, Kulkarni M, Muralidharan K, Singh A, Sood A, Yadav S (2024) Can technology facilitate scale? Evidence from a randomized evaluation of high dosage tutoring. NBER Working Paper No. 32510. https://www.nber.org/papers/w32510

Büchel K, Jakob M, Kühnhanss C, Steffen B, Brunetti A (2022) The effect of personalized learning on students' learning outcomes and motivation: evidence from a randomized controlled trial. J Labor Econ 40:151–194

Bulman G, Fairlie R (2016) Technology and education. In: Hanushek EA, Machin S, Woessmann L (eds) Handbook of the economics of education, 5th edn. Elsevier, Amsterdam, pp 239–280

De Simone L, Ogundeyi A, Adesanya A, Bello M, Adeniran A, Bello A, Daramola A, Jhingran D (2025) From chalkboards to chatbots. Generative AI for education: Experimental evidence from Nigeria. World Bank Policy Research Working Paper No. 10747

Dynarski M, Agodini R, Heaviside S, Novak T, Carey N, Campuzano L, Pendleton A (2007) Effectiveness of reading and mathematics software products: findings from the first student cohort. Natl Cent Educ Eval Reg Assist, Inst Educ Sci, US Dept Educ

Escueta M, Nickow AJ, Oreopoulos P, Quan V (2020) Upgrading education with technology: insights from experimental research. J Econ Lit 58:897–996

Fan Y, Tang L, Le H, Shen K, Tan S, Zhao Y, Shen Y, Li X, Gašević D (2024) Beware of metacognitive laziness: effects of generative artificial intelligence on learning motivation, processes, and performance. Br J Educ Technol 56:489–530. https://doi.org/10.1111/bjet.13544

Gray-Lobe G, Keats A, Kremer M, Mbiti I, Ozier O (2022) Can education be standardized? Evidence from Kenya. Becker Friedman Institute Working Paper 2022-68. https://bfi.uchicago.edu/working-paper/2022-68/

Guryan J, Ludwig J, Bhatt MP, Cook PJ, Davis JMV, Dodge KA, Pollack HA (2023) Not too late: improving academic outcomes among adolescents. Am Econ Rev 113:738–765. https://doi.org/10.1257/aer.20210434

Heckman JJ, Stixrud J, Urzua S (2006) The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. J Labor Econ 24:411–482. https://doi.org/10.1086/504455

Henkel O, Horne-Robinson H, Kozhakhmetova N, Lee A (2024) Effective and scalable math support: experimental evidence on the impact of an AI-math tutor in Ghana. In: André E et al (eds) Artificial intelligence in Education. Posters and late breaking Results, workshops and Tutorials, industry and innovation Tracks, Practitioners, doctoral consortium and blue Sky (AIED 2024). Springer, pp 373–381

Hirata, Guilherme (2022) Play to learn: The impact of technology on students' math performance. Journal of Human Capital, 16(3), 437–459. https://doi.org/10.1086/719771

Levy Yeyati E, Robano V, Pereiro E, Porto C, Koleszar V (2025) Generative AI in education: A framework for leveraging digital tools in Latin American classrooms. School of Government Working Paper No. 202050327, Universidad Torcuato Di Tella

Muralidharan K, Singh A, Ganimian AJ (2019) Disrupting education? Experimental evidence on technology-aided instruction in India. Am Econ Rev 109:1426–1460

Oakley B, Johnston M, Chen K-Z, Jung E, Sejnowski T (2025) The memory paradox: why our brains need knowledge in an age of AI. The future of artificial intelligence: Economics, Society, risks and global policy. Springer Nature. in press

OECD (2023c) PISA 2022 results (I): the state of learning and equity in education. OECD Publishing. https://doi.org/10.1787/53f23881-en

OECD (2023a) Putting AI to the test: Large language models and their relevance for education. OECD Publishing. https://www.oecd.org/en/publications/putting-ai-to-the-test_2c297e0b-en.html. Accessed 11 July 2025

OECD (2023b) Is education losing the race with technology? AI's progress in maths and reading. OECD Publishing https://www.oecd.org/en/publications/is-education-losing-the-race-with-technology_73105f99-en.html. Accessed 11 July 2025

OECD (2024) PISA 2022 Results (Volume II): Learning During – and From – Disruption. OECD Publishing. https://doi.org/10.1787/a97db61c-en

Oreopoulos P, Gibbs C, Jensen M, Price J (2024) Teaching teachers to use computer assisted learning effectively: Experimental and quasi-experimental evidence. NBER Working Paper No. 32388

Risko EF, Gilbert SJ (2016) Cognitive offloading. Trends Cogn Sci 20:676–688. https://doi.org/10.1016/j.tics.2016.07.002

Rockoff J (2015) Evaluation report on the School of One i3 expansion. Unpublished manuscript. https://www8.gsb.columbia.edu/researcharchive/articles/26253

Rodriguez-Segura D (2022) EdTech in developing countries: A review of the evidence. World Bank Res Obs 37:171–203

Thomas DR, Lin J, Gatz E, Gurung A, Gupta S, Norberg K, Fancsali SE, Aleven V, Branstetter L, Brunskill E, Koedinger KR (2024) Improving student learning with hybrid human-AI tutoring: A three-study quasi-experimental investigation. Proc Learn Anal Knowl Conf 14:404–415. https://doi.org/10.1145/3636555.3636896

Van Klaveren C, Vonk SJJ, Cornelisz I (2017) The effect of adaptive versus static practicing on student learning: evidence from a randomized field experiment. Econ Educ Rev 58:175–187

Wang RE, Ribeiro AT, Robinson CD, Loeb S, Demszky D (2024) Tutor CoPilot: A human-AI approach for scaling real-time expertise. Stanford University Working Paper. https://osf.io/8d6ha