The Tailoring Premium: How AI Design Unlocks Student Engagement and Learning *

Jaime Polanco-Jiménez† Kristof De Witte‡

September 23, 2025

Abstract

As schools increasingly adopt Artificial Intelligence, policymakers face a crucial trade-off between deploying inexpensive, general-purpose models and investing in tools tailored to the curriculum. We provide the first large-scale causal evidence on this choice. In a randomized control trial with 2,440 secondary students, we find that offering a curriculum-tailored chatbot increases immediate learning by 0.126 standard deviations (ITT), while a generic chatbot has no effect. This difference is driven entirely by student engagement: the tailored tool increased module completion by 15.5 percentage points. For students induced to complete the module by the tailored design, the effect is larger and more durable, increasing long-term knowledge retention by 0.23 standard deviations. We show that the learning gains from educational AI hinge on its design. Deep curricular integration is effective because it first secure the problem of student engagement.

JEL Codes: I21, C93, O33, J24

*Keywords:*Artificial Intelligence, Educational Technology, Student Engagement, Curricular Integration, Human Capital, Randomized Controlled Trial

*Authors acknowledge financial support from the Horizon Europe project BRIDGE (grant 101177154). Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

†Corresponding author. Leuven Economics of Education Research, KU Leuven, Belgium; Department of Economics at Pontificia Universidad Javeriana, Colombia (email: jaime.polancojimenez@kuleuven.be, jaime.polanco@javeriana.edu.co)

‡Full Professor at Leuven Economics of Education Research, KU Leuven, Belgium; UNU-Merit, Maastricht University, the Netherlands (email: kristof.dewitte@kuleuven.be).

1 Introduction

School systems worldwide are facing persistent teacher shortages (Pressley, 2021; Sutcher, Darling-Hammond, & Carver-Thomas, 2019), and are increasingly turning to Artificial Intelligence (AI) as a scalable solution (Wang et al., 2024). This pivot confronts policymakers with a crucial trade-off between adopting inexpensive, general-purpose AI (e.g., ChatGPT, Gemini, Claude, etc.) and investing in platforms deeply tailored to the curriculum. The optimal choice depends on the returns to such tailoring, a question for which there is a lack of causal evidence.

The design of effective educational interventions is a central question in economics. A large literature establishes that merely increasing instructional time is an inefficient way to build human capital; the quality of that time is the primary driver of student achievement (Aucejo & Romano, 2016; Jaume & Willén, 2019). Technology has long been proposed as a scalable solution to improve instructional quality, but the evidence is decidedly mixed. Early experimental studies often found that simply replacing live lectures with standard online formats had null or even negative effects on student learning, with particularly adverse impacts on lower-achieving students (Alpert, Couch, & Harmon, 2016; Cacault, Hildebrand, Laurent-Lucchetti, & Pellizzari, 2021; Figlio, Rush, & Yin, 2013). A key insight from this literature, however, is that the effectiveness of technology hinges on its implementation. The seminal work of Bai, Mo, Zhang, Boswell, and Rozelle (2016), for example, shows causally that the returns to educational technology are only realized when it is deeply integrated into the existing teaching program. This paper provides the first large-scale experimental evidence to test whether this principle of curricular integration holds in the context of modern AI.

We conduct a randomized controlled trial (RCT) with 2,440 Belgian secondary students to evaluate the impact of chatbot design on financial literacy, a critical form of human capital where knowledge gaps are wide and consequential (Lusardi & Mitchell, 2014). We randomly assign students within classrooms to one of three arms: traditional instruction

(Control), a Generic Chatbot with general knowledge, and a Tailored Chatbot. Our tailored tool combines two features—content-specificity and pedagogical adaptivity—and our design estimates their joint effect, providing a direct test of the returns to contextualization.

We find that curricular integration is the primary driver of student learning, an effect mediated through the student engagement. Our main Intent-to-Treat (ITT) estimate shows that the offer of a curriculum-tailored chatbot increased students' immediate learning by a significant 0.126 standard deviations. This learning effect is driven by the chatbot's ability to solve the first-order problem of participation: the tailored tool increased module completion by 15.5 percentage points, whereas a generic chatbot had no significant effect. For the "compliers" induced to complete the module by the tailored design, the effect is even larger and more durable, increasing long-term knowledge retention by 0.23 standard deviations. Exploratory analysis suggests a key channel for this success may be student self-perception, as the tailored chatbot appears to boost student self-confidence. While recent meta-analyses suggest that AI tools generally have a positive impact (Tlili, Saqer, Salha, & Huang, 2025; Wu & Yu, 2023), our findings qualify this by showing that poor design cannot motivate student participation.

This study makes three primary contributions to the literature on technology and education. First, we provide the first large-scale, causal estimates of returns to curricular integration for modern generative AI. While Bai et al. (2016) showed integration's importance for older ICT, the rise of powerful large language models raises the policy question of whether to "buy" generic AI or "build" tailored tools. Our comparison of a generic and tailored chatbot shows that contextualization is crucial in the AI era.

Second, we identify student engagement as the critical behavioral mechanism determining the effectiveness of educational AI, and we provide evidence on the psychosocial channels that drive this engagement. While recent studies find large, positive learning effects from AI (e.g., Henkel, Horne-Robinson, Kozhakhmetova, & Lee, 2024; Kestin, Miller, Klales, Milbourne, & Ponti, 2024), our results show this is contingent on design choices that foster participation. We find a tailored chatbot significantly increases completion while a generic one does not, and our analysis suggests this is because the tailored tool boosts student self-confidence while the generic tool undermines it. This demonstrates that the effect of AI on student self-perception is a first-order constraint on its ultimate productivity.

Third, we provide causal evidence on the durability of AI-driven learning. Much of the existing literature focuses on immediate learning gains measured directly after an intervention. By tracking students two months later, we can assess whether the knowledge gained is superficial or lasting. Our finding that the tailored chatbot produces a significant long-term retention effect of 0.23 standard deviations for compliers directly addresses the challenge of "fade-out" that plagues many educational interventions (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996) and shows that well-designed AI can foster durable human capital accumulation.

The remainder of this paper proceeds as follows. Section 2 details the institutional context. Section 3 describes our experimental design and data. Section 4 lays out our identification strategy. Section 5 presents our main findings, beginning with the ITT effect on learning, then documenting the engagement mechanism, and finally estimating the LATE on long-term retention for compliers. Section 6 concludes.

2 Financial Education in Flanders: A Case for Tailored AI

Financial literacy is a critical form of human capital with substantial, long-run consequences for household economic security (Lusardi & Mitchell, 2014). In response, a growing number of countries have integrated financial education into their secondary school curricula (OECD, 2020). Belgium's Flemish community offers a compelling setting to study the implementation of this mandate. Following a major 2019 reform, financial literacy became a key cross-curricular competence for all secondary students, as mandated by the region's educational modernization act (Vlaamse Regering, n.d.).

This policy, however, has produced a paradox: while Flemish students rank among the world's best on PISA financial literacy assessments, this high average masks a severe achievement gap linked to socioeconomic status (De Witte, De Beckker, & Holz, 2020). The most recent PISA data confirm this stark inequality. In 2022, the performance gap between socio-economically advantaged and disadvantaged students in Flanders was 104 score points, substantially larger than the OECD average of 87 points (OCDE, 2024). Even more telling, a student's economic, social, and cultural status (ESCS) explains 16.8% of the variance in financial literacy performance in Flanders—one of the strongest such relationships among developed economies and far exceeding the OECD average of 11.6%.[1] This evidence underscores that while average performance is high, the educational system in Flanders struggles to decouple academic achievement from students' family backgrounds, creating a clear policy imperative for interventions that can deliver high-quality, standardized instruction to all students.

This inequality stems from two core implementation challenges documented by De Witte et al. (2020). First, because financial literacy is a cross-curricular subject, it is often taught by non-specialists, creating a demand for standardized, high-quality instructional resources. Second, Flemish classrooms exhibit significant student heterogeneity across academic (ASO), technical (TSO), and vocational (BSO) tracks, making a one-size-fits-all approach ineffective.

These dual needs—for standardization to ensure quality and for personalization to address heterogeneity—create a fundamental tension for policymakers. This paper tests a technological solution explicitly designed to resolve this tension. Our 'Tailored Chatbot' intervention provides standardized, curriculum-aligned content to every student, addressing the quality-control problem. Simultaneously, it delivers adaptive, personalized instruction based on individual responses, addressing the need for differentiation. Our experiment, therefore, does not simply test a 'better' AI tool; it tests whether a specific design combin-

[1] The PISA index of economic, social and cultural status (ESCS) is a composite measure derived from student reports on parental occupations, parental education, and home possessions.

ing standardization and personalization can solve a core problem in educational delivery.

Our study is situated at the intersection of literatures on the educational production function, educational technology, and AI. A long line of research shows that simply increasing instructional time often yields surprisingly small returns (Hanushek, 2003; Jaume & Willén, 2019). Instead, the quality and efficiency of that time are the primary drivers of learning (Aucejo & Romano, 2016). This insight motivates our focus on outcomes beyond immediate test scores. We analyze learning efficiency—the knowledge gained per unit of time—and the durability of learning, or long-term knowledge retention. Demonstrating a lasting impact is particularly important, as many educational interventions tend to lose their effectiveness over time (Cooper et al., 1996).

Technology is often proposed as a solution to enhance instructional quality at scale, but evidence on its effectiveness is mixed. While some studies find positive effects, rigorous experimental evaluations often find null or even negative impacts from simply replacing in-person teaching with standard online formats (Cacault et al., 2021; Figlio et al., 2013). This suggests that implementation details are paramount. Indeed, the seminal work of Bai et al. (2016) shows causally that the effectiveness of ICT depends critically on its integration with the local curriculum.

The latest wave of EdTech, powered by AI, promises to overcome the limitations of older online tools by offering personalized and adaptive learning. Recent meta-analyses confirm that AI-powered tools can have a positive effect on student learning (e.g., Tlili et al., 2025; Wang et al., 2024; Wu & Yu, 2023). However, the rise of powerful, general-purpose AI models introduces a new dimension to the finding of Bai et al. (2016). The question is no longer simply whether to integrate technology, but how deeply. Is it sufficient to use a generic AI tool that understands a topic broadly (our T1 arm), or is the key to unlocking educational productivity to use an AI that is deeply tailored to the specific local curriculum (our T2 arm)? To our knowledge, no large-scale randomized trial has causally estimated the differential returns to generic versus curriculum-tailored AI. This study is designed to fill this critical gap.

This institutional context and literature motivate a clear set of testable hypotheses. First, consistent with recent meta-analyses (e.g., Wu & Yu, 2023), we expect both AI interventions to improve learning outcomes relative to traditional instruction. Second, and central to our contribution, we test the returns to contextualization. Motivated by the critical role of curricular integration (Bai et al., 2016), we hypothesize that the tailored AI (T2) will be significantly more effective than its generic counterpart (T1). Third, we predict the primary advantages of the tailored AI will be in improved learning efficiency and superior long-term knowledge retention, addressing the fade-out problem common to many interventions (Cooper et al., 1996), rather than in immediate test score gains. Finally, we explore mechanisms, positing that the tailored AI's success is mediated by its positive impact on non-cognitive outcomes, specifically by fostering greater student engagement (proxied by higher completion rates) and enhancing academic self-confidence (cf. Sales & Pane, 2020).

We also acknowledge two potential threats to the generalizability of our findings. First, our choice of topic, taxes, is one where students may have strong pre-existing beliefs. Second, the effectiveness of the chatbots could depend on students' prior attitudes toward technology. Our rich baseline data allow us to test these hypotheses directly. In Section **??**, we present a formal heterogeneity analysis and show that our main engagement effects are remarkably stable across these dimensions of student attitudes, strengthening the external validity of our conclusions.

3 Data and Experimental Context

We evaluate how the design of a generative AI chatbot influences student learning, and the role of engagement serving as a key mechanism, through a large-scale randomized controlled trial (RCT) conducted from January to May 2024 in the Flemish secondary school system in

Belgium.[2] Our study population consists of 2,440 students in their third grade of secondary school (typically aged 16-18) from 120 classrooms across 58 schools. The experiment was embedded within the standard curriculum on Economic and Financial Literacy, focusing on the complex Belgian personal income tax system.

Upon enrollment, students were randomly assigned individually within classrooms to one of three experimental groups. This design non-parametrically controls for unobserved heterogeneity across teachers, classrooms, and peer groups. Students in the control group (T0) followed the traditional learning path, using existing course materials. The first treatment group, the 'Generic Chatbot' (T1) group, received condensed instruction supplemented by a chatbot with general knowledge of taxation principles but no specific details of the Belgian tax code. Finally, students in the 'Tailored Chatbot' (T2) group interacted with an adaptive chatbot explicitly designed for the Flemish curriculum and the Belgian tax code. This tool combines two key features: content-specificity and pedagogical adaptivity, personalizing the learning path based on student responses. Our design, therefore, estimates the combined effect of these two 'tailoring' dimensions.

3.1 Data Collection and Variable Construction

We collected data via online questionnaires at three points in time: a pre-test (t=0), an immediate post-test (t=1), and a follow-up test two months later (t=2). Our analysis examines three categories of outcomes: learning outcomes, psychosocial outcomes, and a descriptive measure of efficiency.

Our primary learning outcomes are twofold. First, Gained Financial Literacy measures immediate learning, calculated as the difference between a student's post-test and pre-test scores. Second, Knowledge Retention evaluates how well learning persists, based on the student's score on the follow-up test. To ensure comparability, the Educatieve master in de economie FEB of KU Leuven developed the questions for all three test waves, creating

[2] This trial was pre-registered in the AEA RCT Registry on January 27, 2025, with ID AEARCTR-0015266. The pre-analysis plan is available at `https://doi.org/10.1257/rct.15266-1.0`.

a standard item bank of 10 multiple-choice questions that were validated for equivalent difficulty by subject-matter experts.

Second, we analyze the treatment's impact on a range of Psychosocial Outcomes. We administered a comprehensive battery of psychosocial constructs at both pre-test and post-test, allowing us to measure the change in these dimensions as an outcome. These instruments were adapted from seminal, validated scales in the educational psychology literature, including measures of Attitude & Motivation from the MSLQ (Pintrich, Smith, Garcia, & McKeachie, 1991) and Self-Confidence from the General Self-Efficacy Scale (Schaufeli, Salanova, González-Romá, & Bakker, 2002; Schwarzer & Jerusalem, 1995).

The pre-treatment collection of this rich data also serves three key functions for our identification strategy. The baseline measures of demographics, prior academic achievement, and these same psychosocial constructs are used to: (1) conduct a comprehensive balance check to validate our randomization; (2) enable a detailed diagnosis of the selection into attrition, which is our main mechanism; and (3) support a robust exploration of heterogeneous treatment effects across important student subgroups.

3.2 Baseline Balance of the Randomized Sample

We first verify that our randomization produced statistically equivalent groups across the full sample prior to the intervention. Table 1 presents the means and standard deviations of baseline characteristics for each experimental arm, along with p-values for the difference between each treatment group and the control group, estimated from regressions that control for school fixed effects. The balance for categorical variables is shown in Appendix Table 7. The tables show no statistically significant differences at conventional levels across predetermined characteristics. This comprehensive evidence confirms that the randomization was successful, providing a strong foundation for our causal analysis.

Table 1: Baseline Balance Check: Continuous Variables

Variable	(1) Control Mean (SD)	(2) Generic AI Mean (SD)	(3) Tailored AI Mean (SD)	(4) p-val (T1-Ctrl)	(5) p-val (T2-Ctrl)
Pre-Intervention Outcomes					
Financial Literacy Score (Pre-Test)	0.349	0.337	0.343	0.451	0.723
	(0.242)	(0.246)	(0.238)		
Psychosocial Scales (1-5)					
Attitude and Motivation	2.866	2.829	2.796	0.315	0.108
	(0.733)	(0.735)	(0.749)		
Learning & User Experience	2.781	2.781	2.749	0.998	0.452
	(0.888)	(0.879)	(0.845)		
Self-Regulation & Metacognition	2.712	2.647	2.669	0.104	0.298
	(0.821)	(0.780)	(0.781)		
Engagement & Commitment	2.548	2.491	2.507	0.127	0.301
	(0.776)	(0.711)	(0.779)		
Self-Confidence & Self-Efficacy	2.683	2.686	2.693	0.932	0.814
	(0.867)	(0.837)	(0.866)		
Emotional & Psychological Factors	2.884	2.857	2.897	0.421	0.763
	(0.684)	(0.667)	(0.706)		
Observations	799	870	771		

Notes: This table reports means of continuous baseline characteristics for the full randomized sample (N=2,440). Standard deviations are in parentheses. Columns 4 and 5 report p-values from OLS regressions of each baseline characteristic on treatment indicators for the Generic AI (T1) and Tailored AI (T2) groups, respectively, with the Control group as the omitted category. Regressions include school fixed effects. Standard errors are robust and clustered at the school level (58 clusters). No p-value is significant at the 10% level, providing strong evidence of successful randomization.

3.3 The Central Empirical Challenge: High and Differential Attrition

Although the full sample was balanced at baseline, a key aspect of our study is the high overall attrition rate that varies significantly across treatment groups, making it an important economic outcome. Figure 1 documents the participant flow. Out of 2,440 students who were randomized, only 616 (25.3%) provided complete post-test data. This attrition is not random. The completion rate for the tailored chatbot group (T2) was 33.5%, substantially higher than that of the control group (19.3%). In contrast, the generic chatbot group (T1) had the lowest completion rate at just 23.8%.

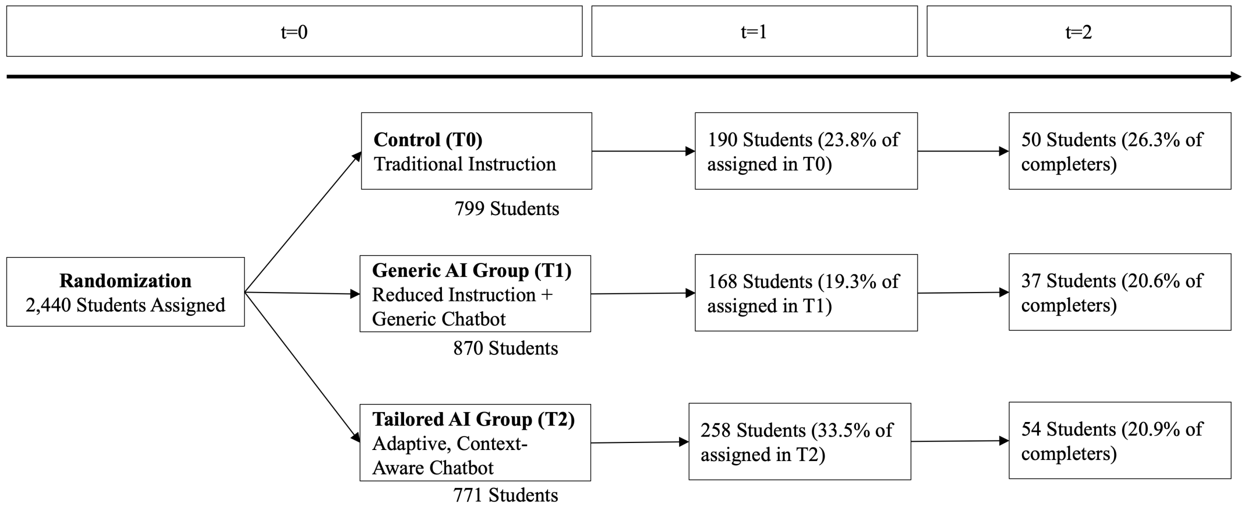


Figure 1: Experimental Design and Participant Flow

Notes: This figure shows the flow of participants through the randomized controlled trial. Numbers in boxes represent the count of participants at each stage. The sample at t=0 represents the full randomized sample. Completion rates at the post-test (t=1) are calculated relative to the number of students initially assigned to that arm. The follow-up completion rate at t=2 is calculated relative to the number who completed the post-test at t=1.

This differential attrition is a key finding of our paper. It suggests that simple comparisons on the sample of completers would lead to biased estimates and justifies an empirical strategy, outlined in Section 4, that uses Intent-to-Treat (ITT) estimates to obtain credible causal parameters.

3.4 Diagnosing Selection: How Chatbot Design Alters Who Persists

The chatbots do not just affect whether students complete the module; as shown in Table 2, they fundamentally alter who persists. This differential selection is a key finding of our study.

The Generic Chatbot (T1) acts as a harsh filter for ability. In this arm, completion is strongly predicted by prior academic performance; the gap in pre-test scores between completers and non-completers is large and statistically significant ($p < 0.001$). Furthermore, the strongest psychosocial predictor of persistence is a student's baseline self-confidence ($p =$

0.007). This evidence paints a clear picture: the generic tool is so unhelpful or frustrating that it retains only the most academically able and resilient students, while causing others to drop out.

Table 2: Diagnosing Selection: Baseline Characteristics of Completers vs. Non-Completers

Variable	**Control (T0)**		**Generic AI (T1)**		**Tailored AI (T2)**	
	Completer	Non-Comp.	Completer	Non-Comp.	Completer	Non-Comp.
Pre-Test Score	0.371 (0.240)	0.318 (0.253)	0.385 (0.236)	0.306 (0.254)	0.354 (0.230)	0.322 (0.252)
Attitude & Motivation	2.903 (0.674)	2.822 (0.740)	2.918 (0.797)	2.809 (0.719)	2.800 (0.693)	2.765 (0.748)
AI Attitude & Motivation	2.988 (1.000)	2.824 (0.925)	2.988 (0.943)	2.951 (0.962)	3.016 (0.938)	2.861 (0.916)
Learning Experience	2.974 (0.919)	2.704 (0.862)	2.895 (0.895)	2.756 (0.887)	2.750 (0.839)	2.726 (0.824)
Self-Regulation	2.723 (0.801)	2.698 (0.823)	2.646 (0.759)	2.634 (0.784)	2.672 (0.726)	2.645 (0.812)
Engagement & Commitment	2.584 (0.771)	2.528 (0.777)	2.519 (0.761)	2.478 (0.696)	2.513 (0.735)	2.476 (0.794)
Self-Confidence	2.738 (0.811)	2.663 (0.890)	2.855 (0.971)	2.633 (0.784)	2.689 (0.795)	2.696 (0.896)
Emotional Factors	2.926 (0.655)	2.886 (0.687)	2.900 (0.663)	2.838 (0.674)	2.870 (0.707)	2.887 (0.713)
Observations	235	564	201	669	300	471

Note: Table reports means with standard deviations in parentheses. It diagnoses the selection into module completion by comparing the baseline characteristics of students who completed the post-test versus those who did not, within each treatment arm. Statistical significance of the differences is discussed in the text. Formal statistical tests for the differences discussed in the text are reported in Appendix Table 8.

In sharp contrast, the Tailored Chatbot (T2) fundamentally changes this selection dynamic and democratizes participation. It largely neutralizes selection on prior academic ability; the difference in pre-test scores between completers and non-completers becomes small and only marginally significant ($p = 0.082$). With ability no longer the primary hurdle, the main factor predicting completion becomes a student's baseline attitude toward using AI for learning ($p = 0.032$). This result is central to our paper's narrative: when an educational tool is well-designed, the barrier to its use shifts from student ability to a student's simple willingness to engage with the technology itself.

4 Empirical Strategy

Our empirical strategy is designed to identify the causal effects of different chatbot designs on student outcomes in the presence of the high and differential attrition documented in Section 3. Our approach establishes a clear hierarchy of evidence. We begin with our primary and most credible estimator, the Intent-to-Treat (ITT) effect. We then specify an Instrumental

Variable (IV) model to quantify the mechanism of student engagement and to estimate the effect of program completion for the relevant subgroup of compliers.

4.1 The Effect of Treatment Assignment: Intent-to-Treat (ITT) Analysis

The ITT measures the causal effect of being *assigned* to a treatment group, regardless of whether the treatment is actually completed. This is a highly policy-relevant parameter because it captures the overall impact of offering a program, including any effects on student engagement (Angrist & Pischke, 2009). We estimate the ITT on the full randomized sample (N=2,440) using the following specification:

$$Y_{is} = \alpha + \delta_1 Z_{is,1} + \delta_2 Z_{is,2} + \mathbf{X'_{is}}\gamma + \mu_s + \eta_{is} \tag{1}$$

where Y_{is} is the outcome for student i in school s for instance, it represents the gained financial literacy for student i in school s, calculated as their post-test score minus their pre-test score. $Z_{is,k}$ is a dummy variable equal to 1 if the student was assigned to treatment arm k, $\mathbf{X_{is}}$ is a vector of baseline student characteristics, and μ_s are school fixed effects.

A primary challenge in estimating Equation 1 is that the outcome Y_i is missing for all students who attrited. Our main specification addresses this by imputing a knowledge gain of zero for all attritors. This approach provides a conservative lower-bound estimate of the true average treatment effect under a plausible assumption about student learning (Duflo, Glennerster, & Kremer, 2007; Kling, Liebman, & Katz, 2007). Specifically, we assume that no student was actively harmed (i.e., ended up with less knowledge) by exposure to an incomplete learning module relative to their own baseline. Given the nature of the educational content, we argue that an outcome of zero net learning represents the most plausible worst-case scenario for a disengaged student. This ensures our ITT estimate is not artificially inflated by the high attrition rate. To confirm our results are not driven by this single choice, we demonstrate their robustness through a comprehensive sensitivity analysis that includes alternative imputation schemes and non-parametric Lee (2009) bounds.

The Lee (2009) outlines potential effects under worst-case selection with a plausible monotonicity assumption. A monotonic treatment effect from a curriculum module, even if incomplete or frustrating, is unlikely to cause a net knowledge loss compared to baseline. Modules aim to increase human capital; the worst case for a disengaged student is learning nothing, with zero knowledge gain. By assigning a value of zero, we adopt this most pessimistic scenario for every attrition case. This ensures that our ITT estimate is a highly conservative lowerbound on the true average treatment effect, which is a common approach in experimental analysis (Angrist & Pischke, 2009; Duflo et al., 2007; Kling et al., 2007). To confirm that our results are not dependent on this specific assumption, we also estimate non-parametric bounds on the Average Treatment Effect (ATE) following Tauchmann (2014), which provides sensitivity case bounds under a similar monotonicity assumption about selection.

4.2 The Effect of Treatment Completion: Instrumental Variable (IV) Strategy

.

While the ITT provides a policy-relevant population average, we are also interested in the causal effect of *actually completing* the chatbot modules. A simple comparison of completers to non-completers would be biased by student self-selection, as documented in Section 3.4.

To identify the effect of completion, we first must conceptualize the ideal experiment. Ideally, we would randomly compel a group of students to complete a given module while a control group uses traditional materials, and then compare their outcomes. Since we cannot force completion, we instead use the random assignment to the offer of each tool as a valid instrument for its use.

Because our treatments are mutually exclusive (a student is in control, generic chatbot T1, or tailored chatbot T2), we cannot estimate a single model with multiple endogenous treatments. Our approach is to estimate two separate Local Average Treatment Effects (LATEs): one comparing the Tailored AI (T2) to the Control group (T0), and a second comparing the Generic AI (T1) to T0 (**?**).

For the tailored chatbot, we estimate the following two-stage least squares (2SLS) model on the subsample of students assigned to either the Control (T0) or Tailored AI (T2) arms:

$$\textit{First Stage:} \quad D_i = \pi_0 + \pi_1 Z_{is,T2} + \mathbf{X_i'}\omega + \mu_s + \nu_{is} \tag{2}$$

$$\textit{Second Stage:} \quad Y_i = \beta_0 + \tau_{T2}\hat{D}_i + \mathbf{X_i'}\lambda + \mu_s + \eta_{is} \tag{3}$$

where D_i is an indicator for module completion for student i in school s. The instrument, $Z_{is,T2}$, is a dummy variable equal to 1 if student i was assigned to the Tailored AI arm (and 0 if assigned to the control). $\mathbf{X_i}$ is a vector of baseline controls, and μ_s are school fixed effects. The first stage (Equation 2) estimates the effect of assignment on completion—our engagement mechanism. The second stage (Equation 3) yields our coefficient of interest, τ_{T2}, which identifies the LATE of completion for the tailored module. An analogous system is estimated on the T0 and T1 subsample to identify τ_{T1}.

4.2.1 Identification and Threats to Validity in the IV Framework

.

The causal interpretation of τ_{LATE} rests on a set of key identifying assumptions (Imbens & Angrist, 1994). First, the instrument relevance condition requires that random assignment significantly affects the probability of module completion ($\pi_1 \neq 0$). We will formally test this by examining the first-stage coefficient and reporting the F-statistic to diagnose weak instrument concerns.

Second, we assume monotonicity, which stipulates that the offer of a chatbot does not cause any student to refuse completion who would have otherwise completed the module in the control group. This is highly plausible in an educational setting; it is difficult to imagine a student who is motivated to complete a standard module but refuses when offered an AI tool designed to help them.

Third, and most critically, the exclusion restriction requires that random assignment ($Z_{is,k}$) affects learning outcomes (Y_i) only through its effect on module completion (D_i). We must rule out direct psychological effects of assignment. For example, one might worry about: (1) a "discouragement effect," where a low-achieving student assigned to a "fancy" AI tool feels intimidated and performs worse even without using it; or (2) an "excitement effect," where a student is excited by the assignment and engages more with the teacher, affecting learning through a non-module channel. We argue these effects are second-order compared to the intensive treatment of completing the learning module itself. Furthermore, our within-classroom randomization design non-parametrically controls for any general Hawthorne effects common to all students in the experiment, strengthening the credibility of the exclusion assumption. Any remaining direct psychological effect would most likely bias our LATE estimates downwards if, for example, it discourages control-group students who feel they missed out on a tool, making our findings conservative estimates of the true effect of completion.

Under these conditions, the IV estimate τ_k identifies the Local Average Treatment Effect: the average causal effect of completing the module for the specific subgroup of students who were induced to complete it by their assignment to treatment arm k. These students are known as "compliers." In our context, the compliers for the tailored tool are students who would not complete the module under traditional instruction but are persuaded to do so when offered the tailored chatbot. Our attrition analysis in Section 3.4 suggests these are not just the highest-ability students, but a broader, more representative group whose participation is conditional on the tool's design. Understanding the effect for this marginal group is highly policy-relevant, as they are precisely the students an effective intervention must reach.

5 Results

We present our findings in a sequence that reflects our causal strategy. We begin by establishing our main causal estimate: the Intent-to-Treat (ITT) effect of chatbot assignment on student learning outcomes. We immediately subject this finding to rigorous robustness checks. We then unpack this learning effect by quantifying the impact of chatbot design on student engagement, which serves as our key mechanism. Finally, we estimate the Local Average Treatment Effect (LATE) to understand the magnitude of learning gains for the specific students whose engagement was secured by the tailored chatbot.

5.1 The Impact of Chatbot Assignment on Student Learning

Our primary causal estimates are the Intent-to-Treat (ITT) effects on our two main learning outcomes: immediate Gained Financial Literacy and long-term Knowledge Retention. To ensure our estimates are robust to the severe attrition documented in Section 4, we present our most conservative ITT point estimates alongside non-parametric Lee (2009) bounds.

Table 3 presents our main findings. Our preferred ITT point estimate (Column 1) indicates that offering the tailored chatbot increased the 'Gained Financial Literacy' score by 0.126 standard deviations.

Table 3: The Causal Effect of Chatbot Assignment on Learning: ITT Estimates and Lee Bounds

	Gained Financial Literacy (SD)		Knowledge Retention (SD)	
	(1) ITT (Impute 0)	(2) Lee Bounds	(3) ITT (Impute 0)	(4) Lee Bounds
Panel A: Generic AI (T1) vs. Control				
Treatment Effect	-0.004 (0.035)	[-0.015, 0.412] (0.152, 0.169)	-0.002 (0.003)	[-0.101, 0.507] (0.224, 0.230)
95% Conf. Interval		*[-0.265, 0.690]*		*[-0.469, 0.886]*
Panel B: Tailored AI (T2) vs. Control				
Treatment Effect	0.126*** (0.037)	[-0.443, 0.474] (0.129, 0.117)	0.026* (0.009)	[-0.106, 0.590] (0.202, 0.203)
95% Conf. Interval		*[-0.655, 0.666]*		*[-0.438, 0.923]*
Observations	2,440	2,440	2,440	2,440

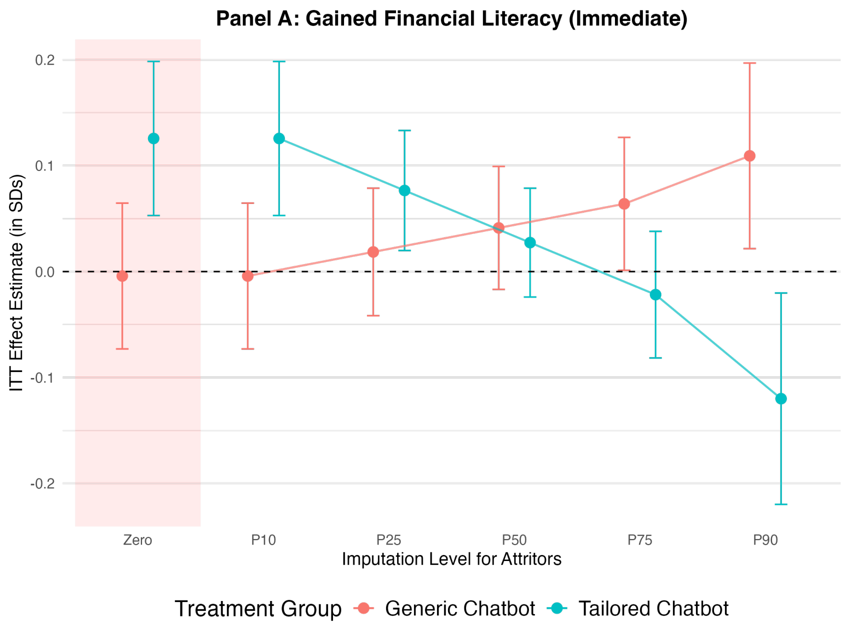
Notes: This table reports the causal effect of treatment assignment on learning for the full randomized sample (N=2,440). Outcomes are standardized. Columns 1 and 3 report ITT point estimates from OLS regressions where outcomes for attritors are imputed to be zero. Robust standard errors, clustered by school, are in parentheses. Columns 2 and 4 report non-parametric Lee (2009) bounds on the Average Treatment Effect. The first row for the bounds shows the point estimates for the lower and upper bound. The second row shows the standard error for each bound estimate in parentheses. The third row reports the 95% Imbens-Manski confidence interval for the full bound set in italics and square brackets. Models in Columns 1 and 2 do not include baseline controls; models in Columns 3 and 4 include a full set of baseline controls and fixed effects. *** p¡0.01, ** p¡0.05, * p¡0.1.

To demonstrate that the result is not a unique effect of the imputation or the level of attrition. We address this in two ways. First, the Lee bounds analysis (Column 2) provides a formal, conservative test. The point estimates for the bounds on the Average Treatment Effect (ATE) for the tailored chatbot are [-0.443, 0.474], and the 95% confidence interval around this set is [-0.655, 0.666]. As expected with high attrition, this interval is wide and contains zero, meaning we cannot rule out a null effect based on this test alone.

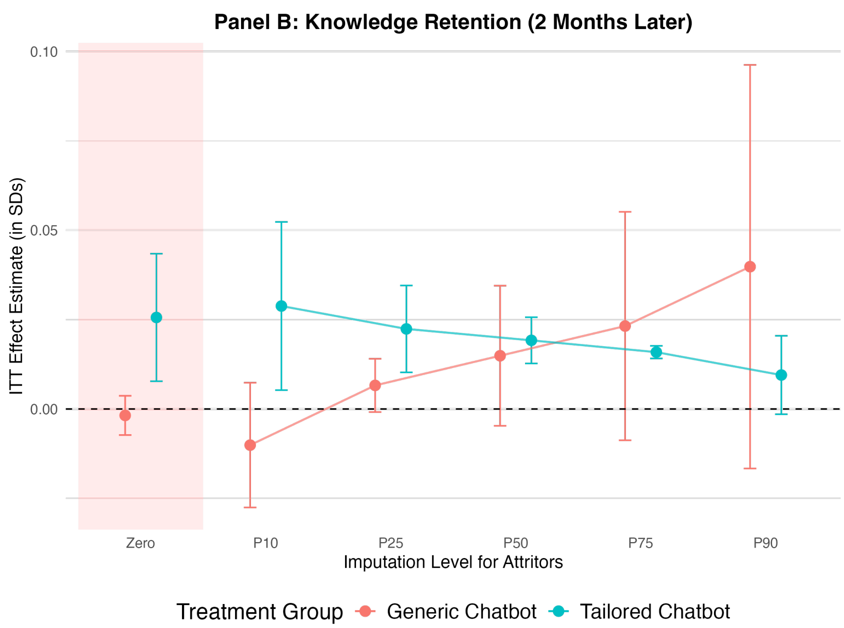
Therefore, our second and more informative robustness check is a comprehensive sensitivity analysis, visualized in Figure 2. This figure provides powerful visual confirmation of our core finding. As shown in Panel A, the 95% confidence interval for the tailored chatbot's effect on immediate learning remains positive and statistically significant for our preferred 'Zero' imputation and for the 'P10' imputation. The effect only attenuates to statistical insignificance under more optimistic assumptions about how attritors would have performed.

The evidence for long-term retention is more modest but follows a similar pattern. The ITT point estimate for the tailored chatbot (Column 3) is a small but statistically significant

0.026 standard deviations. Panel B of Figure 2 shows this positive effect is remarkably stable, remaining statistically significant across several conservative imputation scenarios.



(a) Panel A: Gained Financial Literacy (Immediate)



(b) Panel B: Knowledge Retention (2 Months Later)

Figure 2: Sensitivity of ITT Estimates to Imputation Assumptions

Notes: This figure plots Intent-to-Treat (ITT) point estimates and their corresponding 95% confidence intervals for the effect of chatbot assignment on our two primary learning outcomes. Each point on the x-axis represents a different method for imputing missing outcomes for attritors. "Zero" imputes a score of zero, while "P10" through "P90" impute the 10th to 90th percentiles of the observed outcome distribution from the completer sample. The plots visually demonstrate the robustness of the tailored chatbot's positive effects under conservative assumptions.

Taken together, this comprehensive analysis provides strong, credible evidence that the offer of the tailored chatbot has a genuine, positive causal effect on student learning, both immediately and, more modestly, in the longer term. In contrast, we find that the generic chatbot performs as well as traditional instruction.

5.2 The Mechanism: A Non-Monotonic Effect on Student Engagement

Having established a learning effect for the tailored chatbot, we now investigate the primary mechanism: student engagement. Table 4 presents the ITT effect of treatment assignment on the probability of completing the module. This analysis serves as both a quantification of the engagement mechanism and the necessary first stage for our LATE model.

Table 4: The Effect of Chatbot Design on Student Engagement

	(1)	(2)
Dependent Variable:	Completed Post-Test (0/1)	
Assigned to Generic AI (T1)	0.033	-0.007
	(0.081)	(0.076)
Assigned to Tailored AI (T2)	0.155**	0.178**
	(0.079)	(0.077)
Control Group Mean (Constant)	19.3%	
Model Specification		
Baseline Controls	Yes	No
Fixed Effects	Yes	Yes
Observations	2,314	2,440
R-squared	0.951	0.944

Notes: This table reports estimates of the Intent-to-Treat (ITT) effect on the probability of completing the post-test. The sample is the full set of randomized students. The coefficients represent the effect of treatment assignment relative to the control group (omitted). Column (1) includes a full set of baseline controls and individual- and school-level fixed effects. Column (2) is a parsimonious specification with fixed effects only. Robust standard errors, clustered by student, are in parentheses. The control group completion rate was 19.3 *** p¡0.01, ** p¡0.05, * p¡0.1.

The results reveal that curricular tailoring is the critical determinant of engagement. In our preferred specification with controls (Column 1), assignment to the tailored AI (T2) significantly increased the completion rate by 15.5 percentage points relative to the control group's baseline rate of 19.3%. This represents a nearly 80% increase in participation. In contrast, the generic AI (T1) had no statistically significant effect on student completion. Simply providing access to a general-purpose AI tool was not enough to improve student participation over traditional methods.

The failure of the generic tool appears to be rooted in the intense positive selection it induces. As we document in Appendix D, an exploratory analysis of the non-random completer sample shows that those who persisted in the generic arm appeared spuriously more 'efficient'—a likely artifact of only the most able students completing the frustrating module. This reinforces that the tool's primary effect was to filter out, rather than engage, the typical student.

A key question is whether this average effect of the tailored chatbot is driven by specific student subgroups. To test the robustness of this engagement effect, we explore whether it is concentrated among specific student subgroups. We conduct a comprehensive heterogeneity analysis by interacting the treatment assignments with a wide range of baseline characteristics (see Appendix F for full results).

The analysis reveals a remarkable consistency in the effectiveness of the tailored chatbot. As shown in Appendix Table 12, the interaction terms between assignment to the tailored chatbot (T2) and nine different characteristics—including gender, school track, baseline performance, and prior attitudes—are all small and statistically insignificant. This provides strong evidence that the tailored chatbot's ability to increase engagement is a general phenomenon across the student population, not one confined to a particular group. The primary takeaway is the robust, broad-based positive engagement effect of curricular tailoring.

5.3 The Effect on the Margin: Estimating the LATE for Compliers

Having established that the offer of the tailored chatbot boosts learning, we now use an Instrumental Variable (IV) strategy to estimate the causal effect of *completing* the module for the specific students whose behavior was changed by the intervention. This Local Average Treatment Effect (LATE) is policy-relevant as it quantifies the returns to a successful engagement strategy for students on the margin. We estimate the effect for each treatment arm in separate 2SLS regressions against the control group.

Table 5 presents the 2SLS estimates. The results for the two chatbots could not be more different.

Table 5: The Causal Effect of Module Completion on Learning (2SLS LATE Estimates)

	(1)	(2)
Dependent Variable:	Gained Learning (SD)	Knowledge Retention (SD)
Panel A: Generic AI (T1 vs. T0) - Not Interpreted as LATE		
Completed Generic AI	1.267	-1.328
	(4.832)	(6.883)
First-Stage Coeff. (π_1)	0.060 (0.230) *	0.060 (0.230)*
First-Stage F-statistic	4.38	4.38
Observations (T0 + T1)	1,669	1,669
Panel B: LATE for Completed Tailored AI (T2 vs. T0)		
Completed Tailored AI	0.1476***	0.333$^{\cdot}$
	(0.0305)	(0.202)
First-Stage Coeff. (π_1)	0.0956 (0.0266)***	0.0956 (0.0266)***
First-Stage F-statistic	18.10	18.10
Observations (T0 + T2)	1,570	1,570

Notes: This table reports LATE estimates from separate 2SLS regressions. Panel A uses the sample of students assigned to the Control (T0) and Generic AI (T1) groups. Panel B uses the sample of students assigned to the Control (T0) and Tailored AI (T2) groups. The endogenous variable is an indicator for module completion. The instrument is an indicator for assignment to the respective treatment group. The first-stage coefficient for the instrument is reported. The second-stage coefficient for the Generic AI in Panel A is not interpreted as a LATE due to the negative first stage. All regressions include baseline controls and school fixed effects. Standard errors, clustered at the school level, are in parentheses. $^{***} p < 0.01$, $^{**} p < 0.05$, $^{*} p < 0.1$.

For the Generic AI (Panel A), the instrument is weak. The first-stage F-statistic of 4.38 is well below the conventional threshold of 10 for instrument relevance. This formally confirms that assignment to the generic tool was not a powerful motivator for this group. Because the first stage is statistically and economically weak, we lack a well-defined group of "compliers," and we therefore cannot interpret the second-stage 2SLS estimate as a meaningful LATE.

In contrast, the Tailored AI (Panel B) serves as a strong and valid instrument. Assignment to T2 has a large, positive, and statistically significant effect on completion, with a first-stage F-statistic of 18.10. This allows for a credible estimation of the LATE for the students induced to complete the module by the tailored design.

For these compliers, the effects are substantial. Completing the tailored module caused an immediate and statistically significant increase in learning of 0.148 standard deviations ($p < 0.01$). This suggests that for students on the margin of disengagement, a well-designed, curriculum-integrated tool can be highly productive.

When we turn to long-term knowledge retention, the payoff for engaging the marginal student appears even larger. We estimate a large positive LATE of 0.33 standard deviations. While this point estimate is not statistically significant at conventional levels ($p = 0.101$), its magnitude is economically significant and suggests that the learning gains for this complier group may be highly durable, even if we lack the statistical power in our follow-up sample to make this conclusion definitive.

Taken together, our IV analysis provides a powerful complement to the ITT results. It formally demonstrates that the generic tool fails as an engagement device, while the tailored tool's true value lies in producing significant—and potentially lasting—learning gains for the very students who are on the margin of participation.

5.4 Exploratory Evidence on Other Mechanisms

We now explore other potential mechanisms by examining outcomes measured only on the selected sample of completers. These results are descriptive, not causal, and must be interpreted with extreme caution due to the severe selection bias documented in Section 3.4. They are presented to generate hypotheses for future research.

First, we examine the change in students' psychosocial attitudes from pre- to post-test. As shown in Table 6, the most striking finding is the opposing effect on self-confidence. Assignment to the tailored AI is associated with a significant increase in self-confidence among completers, while assignment to the generic AI is associated with a significant decrease. This suggests a potential channel for the engagement effect: the tailored tool may build students' belief in their own ability, while the generic tool undermines it.

Second, we descriptively examine the Learning Efficiency measure for the completer sample. We find that among the selected group of students who finished, those assigned to the generic AI appear more "efficient." This is almost certainly driven by strong positive selection: our attrition analysis shows that only the most able and confident students persisted with the frustrating generic tool, and it is plausible they completed the task quickly. This descriptive result, when combined with our causal findings on attrition, provides further evidence of the generic tool's failure.

Table 6: Exploratory Analysis of Psychosocial Outcomes for Module Completers

	(1)	(2)
Dependent Variable:	Change in Psychosocial Score (Post - Pre)	
	Generic AI (T1)	Tailored AI (T2)
Panel A: Attitude & Motivation	0.044	-0.136
	(0.098)	(0.084)
Panel B: Self-Confidence & Self-Efficacy	-0.224	0.208*
	(0.115)	(0.099)
Panel C: Learning & User Experience	0.058	0.050
	(0.111)	(0.096)
Panel D: Engagement & Commitment	0.002	0.070
	(0.103)	(0.089)
Panel E: Self-Regulation & Metacognition	-0.060	0.127
	(0.090)	(0.078)
Observations	640	

Notes: This table reports coefficients from separate OLS regressions on the **selected subsample of module completers only**. **These are not causal estimates** and are presented for descriptive and hypothesis-generating purposes. Each panel reports the estimated treatment effect on the change in the specified psychosocial construct (measured on a 1-5 Likert scale). The coefficients represent the effect of assignment to each treatment group relative to the control group. All regressions include a full set of baseline controls and school fixed effects. Robust standard errors, clustered by school, are in parentheses.
*** p¡0.01, ** p¡0.05, * p¡0.1.

While not a causal estimate, exploratory analysis on the completer sample suggests a potential mechanism: self-confidence. Completers in the tailored arm reported increased self-confidence, while those in the generic arm reported a decrease. This suggests that good design may foster a virtuous cycle of engagement and self-efficacy, a hypothesis that warrants future experimental investigation.

6 Conclusion

This paper provides the first large-scale experimental evidence on the returns to curricular integration for AI in education. We demonstrate that the design of educational technology is not a secondary detail but a first-order determinant of its success. This effect operates through the critical mechanism of student engagement. By randomly assigning 2,440 students to traditional instruction, a generic chatbot, or a curriculum-tailored chatbot, we show that deep curricular integration is essential for fostering both student participation and durable learning.

Our primary causal finding is that the offer of a tailored chatbot produced a robust increase in immediate student learning of 0.126 standard deviations at the population level. This effect is driven entirely by the chatbot's ability to solve the fundamental challenge of student engagement. The tailored tool increased module completion by 15.5 percentage points, while the generic tool had no effect on participation. The learning gains were most pronounced for students on the margin of engagement. For these "compliers," completing the tailored module led to a large and potentially durable increase in long-term knowledge retention of 0.33 standard deviations, demonstrating that a well-designed tool can convert engagement into lasting human capital.

Our findings offer three crucial insights for policy and the economics of education. First, they serve as a critical qualification to the burgeoning literature on AI in education. The promise of inexpensive, "one-size-fits-all" AI solutions may be illusory. Our results show that without deep curricular integration, these tools fail to engage students and, consequently, fail to produce learning. The null effect of our generic chatbot stands as a stark warning against the indiscriminate adoption of general-purpose AI in the classroom.

Second, our results highlight a path forward for addressing persistent educational challenges. The finding that the tailored chatbot's positive engagement effect is remarkably consistent across students of different academic backgrounds suggests that well-designed technology can serve as a powerful "democratizing" force, making quality instruction accessible to a broad range of learners. In contexts where specialist teachers are scarce, a high-quality, tailored chatbot can be a vital tool for delivering standardized and effective instruction.

Finally, this study establishes a fundamental principle for the age of AI in education. Technology's potential is not realized by its mere existence, but by design choices that solve the first-order behavioral problem of student engagement. For policymakers, the lesson is clear: the "tailoring premium" is real, and investing in deep curricular integration is essential to convert participation into productive and lasting human capital. While our study is limited in that it cannot disentangle the specific returns to content-specificity versus pedagogical adaptivity, it sets a clear agenda for future research. The most pressing question is no longer *if* AI can work, but *how* to design it to solve the fundamental problem of engaging students in a way that generates productive and lasting learning.

References

Alpert, W. T., Couch, K. A., & Harmon, O. R. (2016, May). A randomized assessment of online learning. *American Economic Review*, *106*(5), 378–82. Retrieved from `https://www.aeaweb.org/articles?id=10.1257/aer.p20161057` doi: 10.1257/aer.p20161057

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Aucejo, E. M., & Romano, T. F. (2016). Assessing the effect of school days and absences on test score performance. *Economics of Education Review*, *55*, 70–87. doi: 10.1016/j.econedurev.2016.08.007

Bai, Y., Mo, D., Zhang, L., Boswell, M., & Rozelle, S. (2016). The impact of integrating ict with teaching: Evidence from a randomized controlled trial in rural schools in china. *Computers And Education*, *96*, 1–14. doi: 10.1016/j.compedu.2016.02.005

Cacault, M. P., Hildebrand, C., Laurent-Lucchetti, J., & Pellizzari, M. (2021). Distance learning in higher education: Evidence from a randomized experiment. *Journal of the European Economic Association*, *19*(4), 2322–2372. doi: 10.1093/jeea/jvaa060

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research*, *66*(3), 227–268. doi: 10.3102/00346543066003227

De Witte, K., De Beckker, K., & Holz, O. (2020, June 18). Financial education in flanders (belgium). In K. De Witte, O. Holz, & K. De Beckker (Eds.), *Financial education* (pp. 67–85). Germany: Waxmann Verlag GMBH.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics. In *Handbook of development economics* (Vol. 4, pp. 3895–3962). Elsevier.

Figlio, D., Rush, M., & Yin, L. (2013). Is it live or is it internet? experimental estimates of the effects of online instruction on student learning. *Journal of Labor Economics*, *31*(4), 763–784. doi: 10.1086/669930

Hanushek, E. (2003). The failure of input-based schooling policies. *Economic Journal*, *113*(485), F64-F98. Retrieved from https://EconPapers.repec.org/RePEc:ecj:econjl:v:113:y:2003:i:485:p:f64-f98

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.* (Working Paper, available at arXiv:2402.09809)

Honey, P., & Mumford, A. (1986). *The manual of learning styles.* Peter Honey. Retrieved from https://books.google.be/books?id=4TV-twAACAAJ

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475. Retrieved from https://doi.org/10.2307/2951620 doi: 10.2307/2951620

Jaume, D., & Willén, A. (2019). The long-run effects of teacher strikes: Evidence from argentina. *Journal of Labor Economics*, *37*(4), 1097–1139. doi: 10.1086/703138

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2024). Ai tutoring outperforms active learning. *Research Square*. (Preprint) doi: 10.21203/rs.3.rs-3965934/v1

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*(1), 83–119.

Lee, D. S. (2009, 07). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, *76*(3), 1071-1102. Retrieved from https://doi.org/10.1111/j.1467-937X.2009.00536.x doi: 10.1111/j.1467-937X.2009.00536.x

Lusardi, A., & Mitchell, O. S. (2014, March). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature*, *52*(1), 5–44. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jel.52.1.5 doi: 10.1257/jel.52.1.5

OCDE. (2024). *Resultados pisa 2022 (volumen iv): ¿qué tan inteligentes financieramente son los estudiantes?* París: Publicaciones de la OCDE. Retrieved from https://

doi.org/10.1787/5a849c2a-es doi: 10.1787/5a849c2a-es

OECD. (2020). *OECD/INFE 2020 International Survey of Adult Financial Literacy.* Retrieved 2023-10-27, from https://www.oecd.org/content/dam/oecd/en/publications/reports/2020/06/oecd-infe-2020-international-survey-of-adult-financial-literacy_bbad9b27/145f5607-en.pdf (Accessed: 2023-10-27)

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). A manual for the use of the motivated strategies for learning questionnaire (mslq) [Computer software manual]. Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning, University of Michigan.

Pressley, T. (2021). Factors contributing to teacher burnout during covid-19. *Educational Researcher*, *50*(5), 325–327. doi: 10.3102/0013189X211004683

Sales, A. C., & Pane, J. F. (2020). *Student log-data from a randomized evaluation of educational technology: A causal case study* (Tech. Rep.). RAND Corporation. (Published in Journal of Research on Educational Effectiveness, 13:2, 237-259) doi: 10.1080/19345747.2019.1678257

Schaufeli, W. B., Salanova, M., González-Romá, V., & Bakker, A. B. (2002). The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. *Journal of Happiness Studies*, *3*(1), 71–92. doi: 10.1023/A:1015630930326

Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. In J. Weinman, S. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. causal and control beliefs* (pp. 35–37). Windsor, UK: NFER-NELSON.

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2019). Understanding teacher shortages: An analysis of teacher supply and demand in the united states. *Education Policy Analysis Archives*, *27*(35). doi: 10.14507/epaa.27.3626

Tauchmann, H. (2014). Lee (2009) treatment-effect bounds for nonrandom sample selection. *The Stata Journal*, *14*(4), 884-894. Retrieved from https://doi.org/10.1177/1536867X1401400411 doi: 10.1177/1536867X1401400411

Tlili, A., Saqer, K., Salha, S., & Huang, R. (2025, jan 17). Investigating the effect of artificial intelligence in education (AIEd) on learning achievement: A meta-analysis and research synthesis. *Information Development*. Retrieved from http://dx.doi.org/10.1177/02666669241304407 doi: 10.1177/02666669241304407

Vlaamse Regering. (n.d.). *Voorontwerp van decreet tot wijziging van de codex secundair onderwijs van 17 december 2010, wat betreft de modernisering van de structuur en de organisatie van het secundair onderwijs* (Voorontwerp van decreet No. VR 2018 0202 DOC.0094/3BIS). Retrieved 2024-05-21, from https://www.klasse.be/73458/nieuw-model-studieaanbod-secundair/

Wang, X., Huang, R. T., Sommer, M., Pei, B., Shidfar, P., Rehman, M. S., … Martin, F. (2024, may 15). The Efficacy of Artificial Intelligence-Enabled Adaptive Learning Systems From 2010 to 2022 on Learner Outcomes: A Meta-Analysis. *Journal of Educational Computing Research*, *62*(6), 1348–1383. Retrieved from http://dx.doi.org/10.1177/07356331241240459 doi: 10.1177/07356331241240459

Wu, R., & Yu, Z. (2023, may 3). Do AI chatbots improve students learning outcomes? Evidence from a metaanalysis. *British Journal of Educational Technology*, *55*(1), 10–33. Retrieved from http://dx.doi.org/10.1111/bjet.13334 doi: 10.1111/bjet.13334

A Appendix Data: Baseline Balance Check

Table 7: Baseline Balance Check: Categorical Variables

Variable	(1) Control (%)	(2) Generic AI (%)	(3) Tailored AI (%)	(4) p-value (χ^2 test)
Gender				0.695 ($\chi^2(4) = 2.22$)
Female	50.45	48.41	47.00	
Male	49.55	51.59	52.99	
School Type				0.635 ($\chi^2(8) = 6.20$)
General Secondary (ASO)	68.03	65.82	64.27	
Technical Secondary (TSO)	29.30	31.65	32.00	
Vocational Secondary (BSO)	2.04	1.93	2.44	
Secondary Education in the Arts (KSO)	0.38	0.36	0.51	
Other	0.255	0.24	0.77	
Secondary School Field of Study				0.593 ($\chi^2(12) = 10.26$)
Arts & Sports	1.665	2.182	0.907	
Care & Social Studies	10.371	8.848	9.974	
Economics & Business	20.871	20.970	20.337	
Humanities & Languages	22.663	20.606	20.596	
Vocational & Applied Skills	1.152	1.091	1.425	
STEM	36.748	38.061	40.285	
Others	6.530	8.242	6.477	
Last Dutch Grade				0.290 ($\chi^2(8) = 9.66$)
Under 50%	2.80	2.41	2.70	
50% to 59%	13.76	14.44	12.08	
60% to 69%	40.26	42.48	40.62	
70% to 79%	34.90	29.84	35.09	
Over 80%	8.28	10.83	9.51	
Last Math Grade				0.440 ($\chi^2(8) = 7.94$)
Under 50%	7.52	6.86	8.61	
50% to 59%	23.06	22.86	20.69	
60% to 69%	30.06	32.49	32.01	
70% to 79%	23.95	23.47	26.35	
Over 80%	15.41	14.32	12.34	
Language at home				0.634 ($\chi^2(4) = 2.56$)
Dutch	83.43	83.27	81.10	
French	5.61	6.38	6.43	
Other	10.95	10.34	12.47	
Highest Parents' Educational Level				0.554 ($\chi^2(6) = 4.92$)
Higher education degree	68.92	68.95	68.89	
Secondary education	18.22	16.73	15.68	
No secondary education	3.44	3.01	4.24	
Unknown	9.43	11.31	11.18	
Frequency of Asking Teachers for Help				0.643 ($\chi^2(6) = 6.17$)
Always	0.30	0.48	0.39	
Never	11.46	12.64	13.24	
Often	8.79	8.42	7.58	
Rarely	38.60	38.15	36.50	
Sometimes	41.15	40.31	42.29	
Learning Style (Honey & Mumford, 1986)				0.094 ($\chi^2(6) = 10.82$)
Activist	28.79	33.09	32.13	
Pragmatist	25.99	23.35	24.04	
Reflector	19.36	18.41	22.37	
Theorist	25.86	25.15	21.47	
Observations	799	870	771	

Notes: This table reports the fraction of students in various categorical groups at baseline for the full randomized sample (N=2,440). Column 4 reports the p-value from a Pearson's χ^2 test for the independence of the variable and treatment assignment status across all three groups. No test yields a p-value significant at the 10% level, confirming successful randomization across observable categorical characteristics. For brevity, some categorical variables from the original table have been omitted but show similar balance.

B Appendix: Additional Tables

B.1 Statistical Tests for Attrition Analysis

Table 8 provides the statistical foundation for the attrition analysis presented in Section 3.4. It reports the p-values from two-sample t-tests comparing the means of baseline characteristics for students who completed the post-test versus those who did not, within each of the three experimental arms.

Table 8: Statistical Tests for Differences in Baseline Characteristics Between Completers and Non-Completers

	P-value of Difference (Completer vs. Non-Completer)		
Baseline Variable	Control (T0)	Generic AI (T1)	Tailored AI (T2)
Pre-Test Score	0.008***	<0.001***	0.082*
Attitude & Motivation	0.164	0.111	0.532
AI Attitude & Motivation	0.048**	0.653	0.032**
Learning Experience	<0.001***	0.077*	0.717
Self-Regulation	0.713	0.848	0.646
Engagement & Commitment	0.387	0.527	0.533
Self-Confidence	0.283	0.007***	0.913
Emotional Factors	0.473	0.284	0.758

Notes: This table reports the p-values from two-sample t-tests comparing the means of baseline characteristics for students who completed the post-test versus those who did not, within each treatment arm. The table supports the analysis in Section 3.4.
*Significance codes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.*

C Appendix: Diagnostic Analysis, The Impact of Selection Bias on Naive Estimates

In this section, we provide a diagnostic analysis to illustrate the severe bias that arises from failing to account for non-random attrition. While our main analysis relies on ITT estimates with imputation and Lee Bounds on the full sample, examining the naive estimates on the subsample of completers is instructive. It reveals the potential of the interventions for the selected group of students who use them and highlights the critical importance of our primary empirical strategy.

Table 9 presents two sets of ITT estimates. Panel A reports our conservative lower-bound estimates on the full sample with zero-imputation for our validly imputable learning outcomes. Panel B reports naive OLS estimates on the unrepresentative subsample of students who completed the post-test or follow-up test.

Table 9: The Impact of AI Assignment on Learning Outcomes: ITT Estimates

	Full Sample (Imputed)			**Completer Sample**		
Dependent Variable:	(1) Gain Score (SD)	(2) Learning Eff. (SD)	(3) Retention (SD)	(4) Gain Score (SD)	(5) Learning Eff. (SD)	(6) Retention (SD)
Panel A: Treatment Effects						
Assigned to Generic AI (T1)	-0.0024	0.0333***	-0.2835	0.0625***	0.2584***	0.2685
	(0.0064)	(0.0102)	(0.4970)	(0.0178)	(0.0555)	(0.2214)
Assigned to Tailored AI (T2)	0.0356***	0.0260	-0.0256	0.0127***	0.0465**	0.3872***
	(0.0062)	(0.0251)	(0.2640)	(0.0048)	(0.0182)	(0.0643)
Panel B: Model Specification						
Observations	2,440	2,440	2,440	616	616	141
R-squared	0.035	0.258	0.325	0.300	0.235	0.153
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Baseline Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports Intent-to-Treat (ITT) estimates from OLS regressions. All dependent variables are standardized to have a mean of 0 and a standard deviation of 1 relative to the control group. Columns (1)-(3) use the full randomized sample (N=2,440). Outcomes for students who attrited are imputed to be zero. These are our primary, conservative lower-bound estimates. Columns (4)-(6) use the subsample of students who completed the relevant survey (post-test for Gain Score and Learning Efficiency, N=616; follow-up test for Retention, N=141). These estimates are conditional on completion and are presented to illustrate the impact of selection. All regressions include a full set of baseline controls (pre-test score, gender, parental education, prior grades) and school fixed effects. Robust standard errors, clustered by school, are in parentheses. Significance codes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The contrast between Panel A and Panel B is stark and demonstrates the importance of our identification strategy. The most dramatic difference appears for knowledge retention. In our primary analysis (Panel A, Column 2), we find no significant population-level effect of the tailored AI on retention. However, the naive estimate on the selected completer sample (Panel B, Column 4) shows a massive and highly significant effect of 0.387 standard deviations.

This large discrepancy reveals two things. First, it highlights the **potential** of the tailored chatbot: for the motivated and persistent students who complete the module, the tool is exceptionally effective at fostering durable, long-term learning. This is a crucial finding for understanding the pedagogical power of the intervention. Second, it underscores the severity of the selection bias. The large effect in Panel B is realized only by a non-random subset of students. Our more conservative and credible ITT and LATE analyses in the main text correctly account for this selection to estimate the true causal effects for the broader population and for the marginal student, respectively. The divergence between these panels provides a powerful, non-parametric illustration of why such methods are essential.

D Appendix: Exploratory Analysis of Time Use for Completers

.

To further characterize the selection process, we constructed a descriptive measure of "learning efficiency" for the non-random subsample of students who completed the post-test. We define this as a student's standardized knowledge gain divided by the time they spent on the module (logged by the online platform).

$$\text{Learning Efficiency}_i = \frac{\text{Standardized Gained Financial Literacy}_i}{\text{Time Spent on Module (minutes)}_i}$$

This measure is subject to severe selection bias and cannot be interpreted causally. A simple OLS regression on the completer sample (Table 10) shows that assignment to the Generic AI (T1) is associated with a higher learning efficiency than assignment to the Tailored AI (T2).

We do not interpret this as evidence that the generic tool is a superior teaching instrument. On the contrary, this result is the mechanical consequence of the powerful selection mechanism documented in Section 3.4. The Generic AI was so ineffective at engaging students that only the most able and confident persisted. It is highly plausible that these top-performing students were able to complete the task quickly, leading to a spuriously high "efficiency" score. This descriptive finding, therefore, serves as further confirmation of the intense positive selection induced by the poorly designed generic tool, reinforcing the main narrative of our paper.

Table 10: Descriptive Analysis of Learning Outcomes for Module Completers

Dependent Variable:	(1) Gained Learning (SD)	(2) Knowledge Retention (SD)	(3) Learning Efficiency (SD)
Assigned to Generic AI (T1)	0.194·	0.168	0.212***
	(0.077)	(0.061)	(0.024)
Assigned to Tailored AI (T2)	0.115*	0.311*	0.062**
	(0.032)	(0.049)	(0.018)
Observations	616	141	616

Notes: This table reports results on the selected subsample of module completers only. These are not causal estimates of the LATE and are likely biased due to the severe, non-random attrition documented in Section 3.4. They are presented for descriptive purposes to explore patterns of learning conditional on completion. The dependent variables are standardized. Robust standard errors, clustered by school, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

This descriptive analysis, when combined with our main causal findings, paints a coherent picture. The generic chatbot appears to facilitate superficial, short-term learning for a small, elite group of students, and this learning fades quickly. The tailored chatbot, by successfully engaging a wider and more representative group of students, fosters learning that is less immediately pronounced but significantly more durable. This suggests that the true value of a well-designed educational tool lies not in producing high scores for the best students, but in producing lasting knowledge for the many.

E Appendix: Heterogeneity Analysis of Learning

This appendix presents an exploratory analysis of whether the Intent-to-Treat (ITT) effect on our imputed 'Gained Financial Literacy' outcome varies across different student subgroups. As discussed in the main text, interpreting these interaction effects is complex, as they may capture a combination of differential learning effects and differential engagement effects. We therefore present this analysis to guide future research rather than to draw firm causal conclusions about heterogeneous learning.

Table 11 reports the coefficients for the interaction terms from nine separate OLS regressions. In each regression, the treatment assignment indicators are interacted with a different baseline student or school characteristic.

Table 11: Exploratory Heterogeneity of ITT Effects on Imputed Gained Learning

Interaction Variable:	(1) Female	(2) TSO/BSO Track	(3) Low Score	(4) High Score	(5) Tax Perception	(6) AI Attitude	(7) Asks for Help	(8) Non-Dutch	(9) Teacher Shortage
Interaction Terms									
T1 × Subgroup	-0.025	0.018	0.002	0.022	-0.012	-0.003	-0.008	-0.005	-0.015
	(0.018)	(0.020)	(0.020)	(0.017)	(0.020)	(0.011)	(0.027)	(0.020)	(0.018)
T2 × Subgroup	0.013	0.010	0.023	-0.021	0.007	0.003	0.042	0.001	-0.011
	(0.019)	(0.023)	(0.024)	(0.018)	(0.020)	(0.010)	(0.039)	(0.026)	(0.035)
Observations	2,440	2,440	2,440	2,440	2,440	2,301	2,440	2,434	2,440

Notes: This table reports only the coefficients on the interaction terms from nine separate OLS regressions. The dependent variable in all models is the imputed 'Gained Financial Literacy' score (with zero imputation for attritors). Each column interacts the treatment assignment indicators (T1 and T2) with the "Subgroup" variable listed in the column header. The main effects of treatment and the subgroup characteristic are included in all models but not reported for brevity. All models include school fixed effects. Robust standard errors, clustered by school, are in parentheses.

The results of this exploratory analysis are consistent with our findings on engagement heterogeneity. The estimated learning effect of the tailored chatbot (T2) appears to be remarkably stable across all nine dimensions. None of the interaction terms between T2 assignment and the various student characteristics are statistically significant at conventional levels. This provides suggestive evidence that the modest, positive learning effect of the tailored chatbot is a general phenomenon and not one that is concentrated in or driven by a particular subgroup of students.

F Appendix: Heterogeneity Analysis of the Engagement Effect

This appendix provides the results for the heterogeneity analysis of the engagement effect, which are summarized in Section 5.2. Table 12 reports the estimated interaction effects between treatment assignment and nine different baseline student and school characteristics.

Table 12: Heterogeneous Effects of Chatbot Assignment on Student Engagement (Module Completion)

	(1) Gender	(2) School Track	(3) Baseline Score	(4) Tax Perception	(5) AI Attitude	(6) Asks for Help	(7) Parental Ed.	(8) Home Language	(9) Teacher Shortage
Interaction Terms									
T1 × Subgroup	-0.034	-0.057		0.050	0.078	0.146	-0.095	0.200	-0.139
	(0.546)	(0.157)		(0.468)	(0.222)	(0.690)	(0.335)	(0.466)	(0.351)
T2 × Subgroup	0.128	-0.024		-0.246	-0.006	0.562	-0.014	-0.016	-0.002
	(0.332)	(0.172)		(0.387)	(0.256)	(0.829)	(0.551)	(0.594)	(0.451)
T1 × Low Score			-0.107						
			(0.408)						
T2 × Low Score			-0.036						
			(0.571)						
T1 × High Score			-0.305						
			(0.799)						
T2 × High Score			-0.100						
			(0.733)						
Observations	2,440	2,440	2,440	2,440	2,301	2,440	2,440	2,434	2,440

Notes: Each column reports coefficients for the interaction terms from a separate OLS regression. The dependent variable is an indicator for completing the post-test. The "Subgroup" refers to the characteristic in the column header. The main effects of treatment and the subgroup characteristic are included in all models but not reported for brevity. All models include controls for baseline score and gender (where not the interaction variable) and school fixed effects. Robust standard errors, clustered by school, are in parentheses.

G Appendix: Imputation Sensitivity Analysis

This appendix provides the full regression results for the imputation sensitivity analysis that is summarized in the main text in Section 5.1 and visualized in Figure 2. These tables report the Intent-to-Treat (ITT) estimates for our two primary learning outcomes across a wide range of assumptions about the performance of students who attrited. This analysis demonstrates the robustness of our main findings to these alternative assumptions.

Table 13 presents the sensitivity analysis for our immediate learning outcome, 'Gained Financial Literacy'. The results confirm the robustness of our core finding. Our main lower-bound estimate in Column (1) shows a significant effect of 0.126 standard deviations for the tailored chatbot. This positive and significant effect persists when we impute the 10th percentile for attritors (Column 2) and remains marginally significant at the 25th percentile (Column 3). The effect only fully attenuates to statistical insignificance under the optimistic assumption that attritors would have performed at or above the median of the completer group.

Table 13: Gained Learning (SD): ITT Estimates Across Full Range of Imputations

	Imputation Level for Missing Scores					
	(1) Zero	(2) P10	(3) P25	(4) P50	(5) P75	(6) P90
Assigned to Generic AI (T1)	-0.004	-0.004	0.019	0.041	0.064*	0.109*
	(0.035)	(0.035)	(0.031)	(0.030)	(0.032)	(0.045)
Assigned to Tailored AI (T2)	0.126***	0.126***	0.077**	0.027	-0.022	-0.120*
	(0.037)	(0.037)	(0.029)	(0.026)	(0.031)	(0.051)
Observations	2,440					
Baseline Controls	No					
School Fixed Effects	Yes					

Notes: This table reports Intent-to-Treat (ITT) estimates from OLS regressions of the standardized ‘Gained Financial Literacy‘ score on treatment assignment. The sample is the full randomized population. Each column corresponds to a different method for imputing missing scores for attritors. Column (1) imputes zero. Columns 2-6 impute the 10th to 90th percentiles of the observed ‘Gained Financial Literacy‘ score distribution from the completer sample. All models include school fixed effects. Robust standard errors, clustered by school, are in parentheses.
*** p¡0.01, ** p¡0.05, * p¡0.1.

Table 14 presents the corresponding sensitivity analysis for our long-term ‘Knowledge Retention‘ outcome. These results show that the small, positive effect of the tailored chatbot is remarkably stable. The effect remains positive and statistically significant across a wide range of imputation scenarios, from the most pessimistic (Zero) to the moderately optimistic (P75). This reinforces the conclusion that the tailored intervention had a genuine, durable, albeit modest, impact on student learning.

Table 14: Knowledge Retention (SD): ITT Estimates Across Full Range of Imputations

	Imputation Level for Missing Scores					
	(1) Zero	(2) P10	(3) P25	(4) P50	(5) P75	(6) P90
Assigned to Generic AI (T1)	-0.002	-0.010	0.007	0.015	0.023	0.040
	(0.003)	(0.009)	(0.004)	(0.010)	(0.016)	(0.029)
Assigned to Tailored AI (T2)	0.026**	0.029*	0.022*	0.019**	0.016***	0.010
	(0.009)	(0.012)	(0.006)	(0.003)	(0.001)	(0.006)
Observations	2,440					
Baseline Controls	Yes					
School Fixed Effects	Yes					

Notes: This table reports Intent-to-Treat (ITT) estimates from OLS regressions of the standardized ‘Knowledge Retention‘ score on treatment assignment. The sample is the full randomized population. Each column corresponds to a different method for imputing missing scores for attritors. Column (1) imputes zero. Columns 2-6 impute the 10th to 90th percentiles of the observed ‘Knowledge Retention‘ score distribution from the completer sample. All models include a full set of baseline controls and school fixed effects. Robust standard errors, clustered by school, are in parentheses. *** p¡0.01, ** p¡0.05, * p¡0.1.

H Appendix: Experimental Materials

Pre-Test Questionnaire

- A1. What is your full first and last name? (For example: Johnson John)
- A2. You are a ...
  - Boy
  - Girl
  - X
- A3. Municipality/City of your school:
- A4. Name of your school:
- A5. Which study program do you follow? (e.g. Economics-Mathematics, Economics-Modern Languages or Business Studies)
- A6. In which type of education are you in school?
  - General Secondary Education (ASO)
  - Technical Secondary Education (TSO)
  - Vocational Secondary Education (BSO)
  - Art Secondary Education (KSO)
  - Other
- B1. What was your last grade for Dutch at the end of last school year?
  - Less than 50%
  - 50% or more, but less than 60%
  - 60% or more, but less than 70%

  - 70% or more, but less than 80%
  - More than 80%

- B2. What was your last grade for mathematics at the end of last school year?
  - Less than 50%
  - 50% or more, but less than 60%
  - 60% or more, but less than 70%
  - 70% or more, but less than 80%
  - More than 80%

- B3. Which language do you speak most at home?
  - Dutch
  - French
  - Other

- B4. What is the highest educational level of your parents (mother or father) who live at your home?
  - No secondary education/secondary school not completed
  - Diploma of secondary education/secondary school
  - College/university degree or higher
  - I don't know

- B5. How many (step)brothers and (step)sisters still live at home?
  - 0
  - 1
  - 2

- 3 or more

- B6. How motivated are you to perform well at school?
  - Very unmotivated
  - Unmotivated
  - Neutral
  - Motivated
  - Very motivated

- B7. How often do you ask your teachers for help with schoolwork or studying?
  - Never
  - Rarely
  - Sometimes
  - Often
  - Always

- C1. How much time do you spend on social media and the internet (such as watching videos, surfing, or chatting) on a typical day?
  - 0-1 hour
  - 1-2 hours
  - 2-4 hours
  - More than 4 hours

- C2. How do you feel about the following activities?
  - I use AI assistants a lot (such as ChatGPT or Gemini)

- I already know a lot about financial concepts such as taxes, budgeting, saving and investing
- I find lessons about financial concepts interesting.
- AI tools can help me study.

- D1. Answer the following question based on your preferences:
  - I like to learn by doing experiments and trying things out myself.
  - I am not afraid to take risks and try new things when I learn.
  - I like to think carefully about things before I do them.
  - I learn best when I have time to think about my experiences.
  - I want to understand how things work and why things are the way they are.
  - I like to analyze information and put the pieces together to figure things out.
  - I want to learn things that I can actually use in real life.
  - I like clear instructions and know exactly what to do.
- E1. Jan pays €1000 tax on an income of €5000. What is his average tax rate (tax percentage)?
  - 10%
  - 20%
  - 25%
  - 50%
  - I don't know
- E2. Answer the following question based on your preferences:
  - Do you think taxes are fair in your country?

  - Do you think people in your country know much about taxes?
  - Taxes are essential for funding public services.
  - In general, I feel comfortable performing calculations with numbers
  - I expect AI to help me learn about taxes.

- E3. Peter has an income of €2200 per month. Bart earns €3800 per month. Calculate the pay gap
  - 173
  - 58
  - 43
  - 73
  - I don't know

- E4. Which tax system leads to the most equal income distribution?
  - Degressive tax system
  - Proportional tax system (flat tax)
  - Progressive tax system
  - None of the above
  - I don't know

- E5. A freelancer earns €57,000 gross per year. In a tiered progressive tax system with the following brackets, what is the tax payable (round to whole euros)?
  - Bracket Income bracket (gross per year) Tax rate (%)
  - 1 €0-€20,000 25
  - 2 €20,000-€40,000 40

- 3 over €40,000 53

- €25010
- €26790
- €22010
- €30210
- I don't know

- E6. Ann earns €43,000 gross per year. How much would she have left if the tax rate is 30%?
  - 30100
  - 26667
  - 14333
  - 12900
  - I don't know

- E7. Which factor has the LEAST direct influence on the calculation of income tax?
  - Professional costs
  - Number of dependent children
  - The national average wage
  - Tax-free amount
  - I don't know

- E8. Mattice has a gross annual income of €43,000 and pays €26794 in taxes. What is the average tax rate?
  - 62%

- 23%
- 160%
- 165%
- I don't know

"'latex

- E9. A self-employed person with an income of €50,000 is considering taking on an extra assignment worth €10,000. Which of the following statements is most correct regarding the impact of this additional income on her tax burden?
  - In a globally progressive system, the extra assignment would always result in a higher net income.
  - In a tiered progressive system, the tax rate on the additional income would be identical to that on the initial income.
  - In a degressive system, the total average tax rate on the income would fall after the extra assignment.
  - I don't know.
- F1. Answer the following question based on your preferences:
  - In general, I enjoy learning new subjects, even if they are not directly among my interests.
  - I think knowledge about financial matters can be useful in the future.
  - I am usually open to extra teaching material or tools to help me learn.
  - I like to discover new ways to learn.
  - I expect AI can help me learn about taxes.
  - I like working with computers.

  - I like working with AI tools.

- F2. Answer the following question based on your preferences:
  - I like to try out new digital tools if they can be useful for my studies.
  - I usually don't find it difficult to work with new (online) tools.
  - If I have to use a new digital tool, I am usually willing to put in some extra time to learn it.

- F3. Answer the following question based on your preferences:
  - I often make a plan or schedule before I start my schoolwork.
  - While learning, I pay attention to whether I really understand the material and adjust my approach if not.
  - If I don't immediately understand something, I try to find out what I can do better or differently.

- F4. Answer the following question based on your preferences:
  - I usually find it important to fully commit to my schoolwork.
  - I can usually concentrate well when I am working on an assignment.
  - I often feel like finding out more about the topics covered in class.

- F5. Answer the following question based on your preferences:
  - If something is complicated, I believe I can understand it if I try my best.
  - In general, I feel confident when I start a new challenge for school.

- F6. Answer the following question based on your preferences:
  - I sometimes feel nervous if I don't know what to expect from a subject or lesson topic.

- I look forward to the challenge of learning something new, even though it may be difficult.

Post-Test Questionnaire

- A1. What is your full name? (Example: Jansen Jan)
- A2. To which group were you assigned?
  - Learning path group 1
  - Learning path group 2
  - Learning path group 3
- A3. You are a ...
  - Boy
  - Girl
  - X
- A4. Municipality/City of your school:
- A5. Name of your school:
- A6. Which study program do you follow? (e.g. Economics-Mathematics, Economics-Modern Languages or Business Studies)
- A7. In which form of education do you follow lessons?
  - General Secondary Education (ASO)
  - Technical Secondary Education (TSO)
  - Vocational Secondary Education (BSO)
  - Art Secondary Education (KSO)

    - Other

- A8. Where did you follow the digital lesson?

    - In the regular class with my economics teacher
    - In the regular class but not with my usual teacher
    - In study
    - At home

- B1. Lisa pays €1800 in taxes on an income of €7200. What is her average assessment rate (in percentage)?

    - 30%
    - 25%
    - 20%
    - 35%
    - I don't know

- B2. A junior employee earns €2100 per month. A senior manager earns €5100 per month. Calculate the wage gap (rounded to the nearest whole number).

    - 243
    - 58
    - 143
    - 41
    - I don't know

- B3. A self-employed person earns €38,000 gross per year. With a tiered progressive tax system with the following brackets, what is the tax payable (round to the nearest whole number):

- Bracket Income bracket (gross per year) Tax rate (%)
  - 1 0-€22,000 25
  - 2 €22,000-€42,000 40
  - 3 above €42,000 50
  - €22800
  - €6180
  - €11900
  - €15200
  - I don't know

- B4. Which tax system leads to the most equal income distribution?
  - Degressive tax system
  - Proportional tax system (flat tax)
  - Progressive tax system
  - None of the above
  - I don't know

- B5. Thomas has a gross annual salary of €48,000. If the tax rate is 32%, how much does he have left after taxes?
  - €32640
  - €15360
  - €40000
  - €53500
  - I don't know

- B6. Which factor has the LEAST direct influence on an individual's income tax?
  - The professional costs
  - The average national wage
  - Number of dependent children
  - The tax-free sum
  - I don't know
- B7. Paul has a gross annual income of €52,000 and pays €19,240 in taxes. What is his average assessment rate?
  - 37%
  - 63%
  - 270%
  - 165%
  - I don't know
- B8. A self-employed person with an income of €50,000 is considering taking on an extra assignment worth €10,000. Which of the following statements is most correct regarding the impact of this additional income on her tax burden?
  - In a globally progressive system, the extra assignment would always result in a higher net income.
  - In a tiered progressive system, the tax rate on the additional income would be identical to that on the initial income.
  - In a degressive system, the total average tax rate on the income would fall after the extra assignment.
  - I don't know.

- B9. Answer the following question based on your preferences:
  - Do you think taxes are fair in Belgium?
  - Do you think people know a lot about taxes?
  - Taxes are essential for the financing of public services.
  - In general, I have no problem performing calculations with numbers
  - I expect AI can help me learn about taxes.
- C1. Answer the following question based on your preferences:
  - I like to learn new knowledge about taxes.
  - If I understand more about taxes, this will be useful to me in the future.
  - I find it interesting to learn how tax rules work.
  - The digital lesson I took made me motivated about the subject of taxes.
- C2. Answer the following question based on your preferences:
  - The digital lesson I took helps me to better understand the subject matter of taxes.
  - I feel that the digital lesson I took supported my learning process about taxes.
  - The digital lesson I took is simple and intuitive to use.
  - I would like to use such a digital lesson more often in the future for other subjects.
  - I found that there was a lot of repetition in the digital lesson.
  - I was able to follow the instruction well in the digital lesson.
- C3. Answer the following question based on your preferences:
  - I often make a plan or schedule before I start my schoolwork.

- While learning, I pay attention to whether I really understand the material and adjust my approach if not.
- If I don't immediately understand something, I try to find out what I can do better or differently.

- C4. Answer the following question based on your preferences:
  - I feel enthusiastic when I can learn more about the theme of taxes.
  - I am completely absorbed in the activities related to tax subjects.
  - I have a lot of energy when I have to learn about taxes.
  - I was able to learn a lot in the digital lesson.
  - I was able to concentrate well in the digital lesson.
- C5. Answer the following question based on your preferences:
  - When I made a mistake in reasoning, I was able to figure out how this came about in the digital lesson.
  - If I notice that I do not understand something about taxes, I could find an answer to my questions in the digital lesson.
  - Thanks to the digital lesson, I think about how I can apply what I have learned about taxes in daily life.
- C6. Answer the following question based on your preferences:
  - I feel enthusiastic when I can learn more about the theme of taxes.
  - I am completely absorbed in the activities related to tax subjects.
  - I have a lot of energy when I have to learn about taxes.
  - I was able to learn a lot in the digital lesson.
  - I was able to concentrate well in the digital lesson.