

Task Expansion with Generative AI: Experimental Evidence on Middle-Skilled, Early-Career Workers*

Hanna Brosch^a Christina Langer^b Philipp Lergetporer^c Sandra Mueller^d
Harald Pfeifer^{d,e} Nele Tschoepe^d

^aUniversity of Erlangen-Nuremberg, ^bifo Institute, Stanford Digital Economy Lab, ^cTechnical University of Munich, ^dFederal Institute for Vocational Education and Training, ^eROA at Maastricht University

Abstract

Does generative artificial intelligence (AI) expand the set of tasks that middle-skilled, early-career workers can successfully perform? In a field experiment with 673 final-year IT apprentices at twelve German vocational schools, we randomly assign access to AI for completing occupation-specific tasks. To identify task expansion, we study tasks both within the apprenticeship curriculum and deliberately beyond it. AI raises productivity substantially across all task types, with gains tending to be larger on tasks where performance without AI is lowest. Crucially, productivity increases by 26–31 percentage points on tasks beyond apprentices’ formal training, providing direct evidence of AI-enabled task expansion. These productivity gains do not come at the cost of reduced immediate task comprehension. Treatment effects are generally larger among apprentices with above-median AI literacy.

Keywords: generative artificial intelligence, field experiment, productivity, task expansion, middle-skilled workers, apprenticeships

JEL: J24, O33, C93, I21

*We are grateful to Uschi Backes-Gellner, Torsten Bartel, Stefan Bauernschuster, Eric Bettinger, Erik Brynjolfsson, Andreas Haupt, Marta Serra-Garcia, Albrecht Glitz, Brian Jabarian, Jonas Jessen, Steffen Trutz, Simon Wiederhold, and seminar participants at the Universitat Pompeu Fabra, Technical University of Munich, Annual Meeting of the Economics of Education Committee, International Workshop on Applied Economics of Education (IWAE), Stanford Early-Career Workers Conference, SASE Annual Conference, TUM Campus Heilbronn Workshop on Field Experiments in Economics and Business, and the MPI / ECONtribute Workshop on Human Capital and Education for helpful suggestions and comments. We thank Florian Sakiri for his outstanding support in setting up the IT infrastructure and implementing the experiment as well as George Purkins for excellent data engineering support. We also thank Jonas Belardi, Lukas Bertram, Michael Dörsam, Leon Fridner, Rui Liu, Lars Niederweis, Gudrun Schönfeld, Alexander Steckelberg, Stefanie Steeg, Caroline Wehner, and Felix Wenzelmann for excellent research assistance. We gratefully acknowledge financial support from the Dieter-Schwarz Foundation and the Dr. Hans-Riegel Foundation. Ethics approval was obtained from the Technical University of Munich (2025-43-NM-BA). The experiment presented in this study was preregistered in the AEA RCT Registry (AEARCTR-16936).

1. Introduction

Generative artificial intelligence (AI) may reshape the labor market not only by raising workers’ productivity in tasks they already perform, but by changing which tasks they can perform at all. A growing literature documents rapid workplace adoption of AI (Humlum and Vestergaard, 2025; Bick et al., 2026b) and substantial effects on performance within given tasks and occupations (Noy and Zhang, 2023; Brynjolfsson et al., 2025; Awuah et al., 2025; Dell’Acqua et al., 2026a; Jabarian and Henkel, 2026). Much less is known about whether AI expands workers’ task sets. This margin may be particularly consequential for middle-skilled workers: if AI enables them to perform tasks that previously required more highly skilled workers, it could alter the allocation of work across skill groups and expand their employment and career opportunities (Autor, 2024). We study this task-expansion margin for middle-skilled, early-career workers.

We do so in the context of the German dual apprenticeship system. Apprenticeships are the main pathway into middle-skilled work in Germany – about 60% of the workforce holds an apprenticeship qualification (IAB, 2017). We focus on final-year apprentices, who are close to entering the labor market as fully qualified middle-skilled workers.

The setting provides a particularly clear way to measure task expansion because curricula and examinations are standardized nationwide, allowing us to distinguish tasks within apprentices’ formal training from tasks requiring skills beyond it. We focus on the two largest IT apprenticeships, IT-specialists in Application Development and System Integration. IT work is highly exposed to AI and consists largely of computer-based tasks. It is also among the domains in which AI can most readily raise productivity (Peng et al., 2023; Cui et al., 2026), while human competence remains essential because workers must still read, understand, and supervise code as AI takes over a larger share of the workflow (Shen and Tamkin, 2026).

We conduct a preregistered field experiment with 673 final-year IT apprentices at twelve vocational schools across six German federal states. Within schools, we randomly assign apprentices at the individual level either to individual ChatGPT Pro accounts or to a no-internet control condition.¹ After completing a common pre-treatment baseline task, apprentices solve three occupation-specific tasks: one within the apprenticeship curriculum and two deliberately beyond it, requiring skills associated with early-career

¹The control condition mirrors the standard conditions under which apprentices complete vocational examinations, so the experiment compares AI-assisted performance with the unaided performance on which occupational competence is formally assessed.

software- or network-engineering work, depending on their specialization.² For each task, we measure productivity, comprehension of the underlying material, and subjective task experience. Finally, we record browser histories, task-level timestamps, and ChatGPT interaction logs to measure AI use and construct detailed measures of how apprentices use and interact with AI.

Four findings stand out. First, AI access raises productivity – measured as the share of maximum attainable points achieved – by 26–44 percentage points across the three post-treatment tasks. These effects are large, roughly doubling task performance. Importantly, productivity increases by 26–31 percentage points on the two tasks deliberately beyond the apprenticeship curriculum. This indicates that AI not only raises performance within apprentices’ existing task set but also enables them to perform more successfully on work beyond their formal training. The largest treatment effect arises on the within-curriculum task, consistent with evidence that AI can generate large productivity gains when combined with workers’ existing domain expertise (Dell’Acqua et al., 2026a; Tambe, 2026). Thus, AI generates both substantial task deepening within apprentices’ existing domain and task expansion beyond it. We also find that treatment effects tend to be larger on tasks where performance without AI is lower.³

Second, the average productivity effects also reflect changes in whether apprentices take on a task at all. AI most strongly increases the probability of submitting any substantive answer on the task in each specialization where performance without AI is lowest. This suggests that AI can encourage apprentices to engage with challenging tasks they might otherwise not attempt. Among apprentices who submit an answer, those with AI access also achieve higher scores across all post-treatment tasks.

Third, at least in the immediate term, the productivity gains do not come at the expense of understanding. Immediately after each post-treatment task, the apprentices answer four comprehension questions without AI assistance. Across tasks, we find no evidence of reductions in comprehension. Instead, AI access improves subsequent task comprehension on several tasks. This contrasts with recent evidence that unguarded AI can act as a “crutch” that harms learning: high school students given an unrestricted AI tutor later performed worse on unaided math exams (Bastani et al., 2025), and software engineers who fully delegated coding tasks to AI achieved productivity gains at the cost

²We validate this task progression using reviews by vocational-school teachers, apprentices’ assessments of task relevance, control-group performance, and associations with pre-apprenticeship education.

³This task-level pattern is consistent with prior evidence that AI compresses performance gaps across workers by disproportionately helping lower performers (Noy and Zhang, 2023; Brynjolfsson et al., 2025).

of learning (Shen and Tamkin, 2026). It also speaks to Autor et al. (2026), who find in a three-month experiment with patent lawyers that AI raises assisted performance most for junior lawyers, while only senior lawyers perform better without AI afterwards. The differing findings suggest that productivity gains from AI need not come at the cost of task understanding across settings. Our results are also consistent with Brynjolfsson et al. (2025), who find that customer support agents retain productivity gains even during AI outages. We also find that AI improves task experience, reducing frustration and increasing enjoyment, with effects tending to be larger on tasks where performance without AI is lower.

Fourth, treatment effects are generally larger among apprentices with above-median self-reported AI literacy, but do not vary systematically with pre-apprenticeship education. The heterogeneity by AI literacy is concentrated in items capturing apprentices’ ability to use AI effectively rather than their understanding of how the technology works. This pattern is consistent with Mehmood et al. (2026), who show that targeted training in the use of an AI tool increases its adoption and productivity effects among judges. Taken together, these patterns suggest that the dimensions of human capital relevant for benefiting from AI are not fully captured by conventional educational attainment.

In additional analyses, we use detailed behavioral data from browser histories, task-level timestamps, and ChatGPT interaction logs to characterize how apprentices interact with AI. Apprentices most commonly paste task material into ChatGPT, either directly or together with further instructions, while requests for explanations and prompts that transform task material are less frequent. More intensive prompting and the direct incorporation of AI-generated content into submitted solutions are associated with higher productivity. Finally, our main findings are robust to adjustments for multiple hypothesis testing, alternative outcome definitions, specifications with and without controls, and sample restrictions addressing differential effort and potential experimenter-demand effects.

By providing experimental evidence on AI-enabled task expansion, our paper contributes to three strands of the literature. First, we contribute to the literature on AI as a tool to narrow skill gaps and enable task expansion. Autor (2024) argues conceptually that AI could restore middle-class jobs by extending expert capabilities to broader segments of the workforce. The empirical evidence remains limited and largely concerns heterogeneous effects within a given set of tasks.⁴ In a randomized online experiment

⁴Recent descriptive survey evidence shows that IT workers report reallocating time toward more complex tasks when using AI tools (Zisler and Backes-Gellner, 2026).

conducted outside firms, Cruces et al. (2026) ask adults with heterogeneous educational backgrounds to complete a general, non-domain-specific business problem-solving task and find that AI closes about 75% of the education-based productivity gap. In another online experiment, Haslberger et al. (2025) ask UK working-age adults to complete general text-based tasks involving workplace communication, argument evaluation, and reading comprehension. They find productivity gains but no equalization across educational, occupational, or demographic groups. Otis et al. (2026) conduct a field experiment with Kenyan entrepreneurs and find no average effect of an AI business assistant on revenues and profits: high-performing entrepreneurs benefit, whereas low-performing entrepreneurs perform worse. Kanazawa et al. (2026) study an AI demand-prediction tool for taxi drivers and find that productivity gains accrue only to low-skilled drivers, narrowing the skill gap between high- and low-skilled drivers by 14%. We extend this literature by providing experimental evidence on task expansion itself. Rather than asking whether AI changes performance gaps across workers performing a given task, we test whether AI enables middle-skilled workers to successfully perform tasks beyond their formal training.

Second, we contribute to the recent literature on AI and early-career workers. Brynjolfsson et al. (2026) document a 19% relative employment decline among early-career workers in occupations most exposed to AI, and Lichtinger and Hosseini Maasoum (2025) find that AI adoption is seniority-biased in U.S. resume and job-posting data.⁵ These studies identify early-career workers as particularly vulnerable to AI-driven labor-market shifts, but they study these shifts from the demand side. We complement them from the supply side: we ask whether AI expands the set of tasks that early-career workers can perform and whether these gains come at the expense of task understanding.

Third, we contribute to a fast-growing experimental literature on the productivity effects of AI and task-level complementarity. Field and online experiments show that AI raises workers’ output quality and speed across a range of tasks, including writing (Noy and Zhang, 2023), customer service (Brynjolfsson et al., 2025), software engineering (Peng et al., 2023; Cui et al., 2026; Gambacorta et al., 2026; Paradis et al., 2024; Yeverechyahu et al., 2026), product innovation (Dell’Acqua et al., 2026b), and consulting (Dell’Acqua et al., 2026a). These studies typically focus on specific occupational settings, such as management consultants, software engineers, or customer-service agents, and in several cases on workers within a single firm or organization, while holding workers’ task sets fixed. We instead study tasks that fall both within and beyond workers’ formal training.

⁵Lambert and Schindler (2026) propose an alternative explanation, finding that remote work accounts for most of the decline in entry-level hiring in AI-exposed occupations.

This allows us to identify whether AI not only raises productivity on existing tasks but expands the set of tasks workers can successfully perform.

The remainder of the paper is structured as follows. Section 2 describes the institutional background. Section 3 presents the study design, and Section 4 introduces the empirical model. Section 5 reports our results. Section 6 concludes.

2. Institutional Background

Germany’s dual apprenticeship system provides a natural setting for studying middle-skilled, early-career workers. Apprentices acquire a nationally recognized occupational qualification through a combination of firm-based training and vocational schooling. This qualification places them between workers without formal vocational training and university-educated professionals: apprentices are trained for skilled occupations, but not for the higher-skilled professional roles typically requiring tertiary education. In this sense, apprenticeship graduates correspond closely to the international category of middle-skilled workers at the center of current debates about AI and the labor market (Autor, 2024). Studying whether AI enables them to perform tasks associated with more highly skilled occupations therefore speaks to a broader question that extends beyond the German apprenticeship system.

Around 60% of the workforce holds an apprenticeship qualification, and roughly 1.2 million apprentices train each year across more than 320 nationally recognized occupations (IAB, 2017; BIBB, 2026). Apprentices typically combine three to four days of firm-based training with one to two days at a public vocational school. Nationally standardized curricula and examination requirements ensure that apprentices within an occupation acquire a comparable set of skills across firms and regions (Langer and Wiederhold, 2023). Apprenticeships last between two and three-and-a-half years and are primarily undertaken by graduates of lower- and intermediate-track secondary schools, who enter training directly after school. This means apprentices are typically between the ages of 16 and 20 when they start their apprenticeship. Final-year apprentices are therefore simultaneously workers and learners, allowing us to study AI at an early career stage within a clearly defined occupational skill framework.

Within the apprenticeship system, IT occupations offer a particularly informative setting for studying the effects of AI. We focus on the two largest IT apprenticeships: the IT Specialist in Application Development (Fachinformatiker Anwendungsentwicklung) and the IT Specialist in System Integration (Fachinformatiker Systemintegration). Together, these two occupations account for roughly 4% of all new apprenticeship contracts in Ger-

many, and System Integration alone is the second most popular apprenticeship among men (BIBB, 2026). Three features make these occupations especially well suited to our study. First, they involve tasks highly exposed to AI, including writing and reviewing code, designing data models, building and maintaining IT infrastructure, and resolving security and reliability issues. Second, AI is not yet part of the official apprenticeship curriculum, but 84% of our sample report using it regularly at school or at work. Third, the standardized curriculum allows us to benchmark tasks against an objective curricular standard — the interim and final examination content — and to define tasks “beyond” the apprenticeship scope by reference to early-career software- and network-engineering work. These higher-skilled occupations, which typically require tertiary education, provide a natural benchmark for tasks beyond the apprenticeship curriculum.

This combination of high AI exposure, standardized assessment, and a clear benchmark for tasks beyond the apprenticeship curriculum allows us to test whether AI expands the tasks that middle-skilled, early-career workers can perform.

3. Study Design

This section describes the recruitment and sample, the randomization and experimental procedure, the task design and its validation, and the outcome measures. The study was preregistered in the AEA RCT Registry (AEARCTR-16936).

3.1. Recruitment and Data Collection

Since vocational education in Germany is regulated at the state level, field experiments in vocational schools require approval from the relevant state education ministries. We obtained approval from nine of the sixteen federal states and recruited twelve vocational schools across six states.⁶

We collected data between October 2025 and January 2026 during regular school hours. Depending on the number of final-year IT classes, we conducted up to three consecutive sessions per school and day. To standardize treatment conditions across sites, we installed 60 pre-configured laptops at each school. Schools and apprentices were not informed about the research question or the treatment–control contrast to avoid experimenter-demand effects. Instead, the sessions were framed as an assessment of occupation-relevant task performance.

⁶One additional school participated in a pilot session in September 2025. Appendix C.1 provides further details on the recruitment process.

3.2. Sample

The experimental sample consists of 673 final-year apprentices from twelve vocational schools across six German states, drawn from 23 experimental sessions. Appendix C.2 reports details on the sample construction and exclusion criteria. Column 1 of Table 1 summarizes the sample: apprentices are on average 23.5 years old, predominantly male (9% female), and split between the two largest IT specializations (30% Application Development, 70% System Integration).⁷ Roughly 60% hold a university-of-applied-sciences entrance qualification or a higher qualification. Appendix Table A.1 provides further descriptive statistics.

Table 1: Sample Characteristics and Balancing

	All (1)	Treatment (2)	Control (3)	p-value (4)
Experimental variables				
Task order (within, then beyond)	0.53	0.54	0.52	0.496
Descriptives				
Age	23.47	23.27	23.67	0.230
Female	0.09	0.10	0.07	0.128
University entrance qualification	0.59	0.59	0.59	0.901
Apprenticeship				
Apprenticeship length in years	2.80	2.79	2.80	0.478
Specialization				
<i>Application Development</i>	0.30	0.31	0.28	0.458
<i>System Integration</i>	0.70	0.69	0.72	0.458
Average grade (interim exam)	2.80	2.78	2.82	0.619
Writes final exam during field phase	0.14	0.16	0.12	0.152
Outcomes in baseline tasks				
Share of received points	0.41	0.42	0.40	0.342
Points achieved per minute	1.89	1.94	1.85	0.318
Comprehension: Share correct	0.87	0.87	0.86	0.625
Frustrating task experience	0.01	0.02	0.00	0.821
Fun task experience	-0.05	-0.09	-0.00	0.265
Observations	673	347	326	673

Notes: This table shows group means for the whole sample (column 1), the treatment group (column 2) and control group (column 3). Column 4 shows the p-value of a t-test.

⁷The comparatively high mean age is influenced by a few outliers (median: 22.42) and the fact that only apprentices aged 18 or older were eligible to participate.

3.3. Treatment and Randomization

Treatment and Control Conditions. We estimate the causal effect of access to AI tools on task performance and comprehension in occupation-relevant IT tasks. Treatment-group apprentices were given unrestricted access to individual ChatGPT Pro accounts and the broader internet during the experimental tasks. Browser-tracking data show that use of other websites was limited, and the results are unchanged when excluding apprentices who accessed them.⁸ Control-group apprentices were assigned to complete the same tasks without internet access. The control condition mirrors the standard conditions under which apprentices complete vocational examinations, so control-group performance reflects the unaided performance on which occupational competence is formally assessed. Since AI access was provided through ChatGPT Pro rather than a pinned model, the default model changed during the field period as more advanced models were released. Thus, treated apprentices encountered GPT-4o, GPT-5.1, or GPT-5.2 depending on the date of their session. Appendix D shows that neither average task performance nor treatment effects vary systematically across these model windows.

Experimental laptops were centrally configured to enforce treatment conditions and track internet usage. Treatment-group apprentices worked in an open browser, whereas the browsers of control-group apprentices were configured to block internet access. A browser-based website-tracking extension recorded browsing activity during all open-browser sessions. These data, together with ChatGPT interaction logs, allow us to verify treatment compliance and characterize AI usage patterns.⁹

Randomization. Treatment assignment was randomized at the individual apprentice level within schools on the day of the experiment. Apprentices were randomly assigned to separate treatment and control classrooms through a physical card draw conducted immediately before the session. This within-school randomization ensures that the treatment

⁸Treatment-group apprentices received a direct link to an authenticated ChatGPT Pro account. For technical reasons, restricting their access to the broader internet was not feasible. We therefore used a browser-based website-tracking extension to record their browsing activity during the experimental session. Only 54 treatment-group apprentices (16.6%, excluding 22 people with missing website tracking data) visited a website other than ChatGPT and the survey; nearly half of them (22) with only one visit. The most frequently visited categories were infrastructural websites, i.e., school-specific Wifi-connection portals (15 apprentices, 27 visits), IT information portals (6 apprentices, 38 visits), entertainment websites (6 apprentices, 15 visits), and other AI tools (2 apprentices, 27 visits). Our results remain robust when excluding these individuals, as shown in Appendix Tables D.3 to D.8.

⁹For $N = 76$ control participants, the browser-level internet restriction could not be implemented because of local network problems, and these participants worked in an open browser instead. The website-tracker extension was active during these sessions, so any unintended access to AI tools or other external resources could be identified ex post.

and control groups are drawn from the same school and experimental session, so the estimated treatment effects are not driven by differences across schools or classes. Conducting the two conditions in separate classrooms prevented interaction between treatment- and control-group apprentices during the experiment, thereby minimizing treatment spillovers and behavioral responses arising from awareness of the other treatment condition. Appendix C.4 documents the classroom setup and randomization procedure.

Balancing. Table 1 reports treatment-control differences across predetermined apprentice characteristics and Baseline-task outcomes. None of these differences is statistically significant at conventional levels, indicating that the randomization produced well-balanced treatment and control groups. The table also provides direct evidence on balance in unaided task proficiency: all apprentices completed the same Baseline task before treatment-group access to ChatGPT was activated, and treatment and control apprentices do not differ significantly in terms of points achieved, points achieved per minute, and comprehension.

3.4. Experimental Tasks

Task Setup. Apprentices completed four tasks that differed in their curricular alignment (Baseline, Within, Beyond 1, and Beyond 2) through a standardized interface administered via the online survey platform LimeSurvey. The interface imposed a sequential structure and did not allow participants to navigate freely across tasks or questionnaire blocks. For each task, apprentices first completed the task itself, then answered short task-experience questions, and finally completed a block of comprehension questions targeting the task content.

All participants first completed the Baseline task without AI assistance, providing a common pre-treatment measure of performance. Treatment-group apprentices then received browser access to ChatGPT, whereas the control group continued without internet access. The three post-treatment tasks were occupation-specific and comprised one task within the apprenticeship curriculum (Within) and two beyond it (Beyond 1 and Beyond 2). The order of the Within and Beyond tasks was randomized at the participant level to balance potential fatigue and learning effects.

Treatment-group apprentices retained internet access during the comprehension blocks but were instructed to complete them without AI or other aids. Compliance with this instruction was generally high (see Section 5.2). Excluding apprentices who may have used

AI during the comprehension blocks leaves the results virtually unchanged (Appendix Table A.10).¹⁰

Participant Motivation and Incentives. The experiment did not use task-contingent monetary incentives, consistent with regular IT work, where individual tasks are typically not compensated through piece rates. Moreover, German vocational-school regulations prohibit monetary payments for studies conducted on school premises. We therefore relied on non-monetary incentives aligned with the school and occupational setting. Vocational teachers emphasized the relevance of the tasks for apprentices’ final examinations and professional qualification, and the experimenter highlighted the opportunity to practice exam-relevant tasks. At the end of the session, apprentices reported whether they had taken the study seriously. Only 12% (fully) disagreed with the statement “Overall, I took the study seriously and put effort into the tasks,” suggesting high levels of engagement.¹¹

Task Design. The Application Development subsample completed tasks covering algorithmic problem-solving, software architecture, and machine learning, whereas the System Integration subsample completed tasks covering subnet planning, IT security architecture, and hybrid cloud migration (see Table 2). Each task consisted of one to three subtasks. Full English translations of all task materials are provided in Appendix F.

Tasks were derived from a range of sources and benchmarked against the standard apprenticeship curriculum. The Baseline task corresponds to the level of the apprenticeship’s interim exam, while the Within tasks correspond to the level of the apprenticeship’s final exam. Both draw on examination materials, vocational textbooks, and established e-learning platforms. The Beyond 1 and Beyond 2 tasks were developed to approximate the skill requirements of entry-level software- or network-engineering work, which typically requires a university degree. Importantly, all tasks and scoring rubrics were reviewed and validated by vocational-school teachers from the target occupations.¹²

Curricular Alignment of Tasks. The tasks are designed to differ systematically in their curriculum alignment, allowing us to test whether middle-skilled workers are able to per-

¹⁰Since ChatGPT access was provided through its public web interface, treated participants could in principle access ChatGPT in a separate browser tab before treatment access was formally activated; the tracking logs suggest this happened for exactly one participant.

¹¹The distribution of this measure does not differ between the treatment and control groups, and our results are virtually unchanged when we restrict the sample to apprentices who report having taken the study seriously (see Appendix D).

¹²Appendix Table A.2 provides a further validation by showing that task performance is systematically associated with apprentices’ interim-exam grades.

form tasks that extend beyond the scope of their formal apprenticeship training. The Baseline and Within tasks cover material that apprentices have already learned or are currently learning in the apprenticeship curriculum, whereas the Beyond tasks approximate the skill requirements of entry-level software-engineering positions for Application Development apprentices and entry-level network-engineering positions for System Integration apprentices. The task design worked as intended. Table 2 reports apprentices’ assessments of task relevance separately for the Application Development and System Integration subsamples. In both subsamples, the Baseline and Within tasks are perceived as highly relevant to the apprenticeship, whereas perceived relevance declines as tasks extend beyond the curriculum. This pattern is particularly pronounced for the Beyond 2 machine-learning task in Application Development, which receives the lowest relevance rating across all tasks: only 3% of apprentices consider it relevant to their apprenticeship, compared with 63% for the Baseline task. Overall, these assessments are consistent with the intended progression in curricular alignment, indicating that the tasks capture meaningful variation in their alignment with the formal apprenticeship curriculum.

Table 2: Task Structure and Perceived Task Characteristics (Control Group Only)

	Baseline	Within	Beyond 1	Beyond 2
Time	10 min	20 min	20 min	20 min
Application Development				
Task content	ER model analysis	Algorithmic problem-solving	Software architecture & resilience	Machine learning
<i>Curricular alignment</i>				
Task relevance (share)	0.63	0.40	0.11	0.03
<i>Task difficulty</i>				
Share of points achieved	0.47	0.20	0.28	0.04
System Integration				
Task content	ER model analysis	IPv4 subnet planning	Zero-Trust security concept	Hybrid cloud concept
<i>Curricular alignment</i>				
Task relevance (share)	0.46	0.72	0.38	0.35
<i>Task difficulty</i>				
Share of points achieved	0.38	0.23	0.44	0.58

Notes: The table summarizes the structure and perceived characteristics of the experimental tasks by subsample. The Baseline task was completed before treatment access was activated for the treatment group; the remaining tasks were completed thereafter. “Task relevance” refers to the share of apprentices who rated the task as relevant or very relevant for activities performed in vocational school. “Share of points achieved” reports the average share of achievable points obtained in the respective task. Both measures are reported as control-group means.

Task Difficulty. While the progression in required skill level allows us to measure task expansion, this concept is distinct from task difficulty. Required skill level refers to the knowledge and competencies a task demands, whereas difficulty reflects how challenging the task is for apprentices in practice. We measure task difficulty by realized performance in the control group, captured by the share of achievable points obtained. The two concepts do not necessarily coincide. In Application Development, for example, apprentices perform better on the Beyond 1 task than on the Within task (0.28 versus 0.20), while performance drops sharply for Beyond 2 (0.04). The distinction is even more apparent in System Integration, where apprentices achieve the highest performance on the Beyond 2 task (0.58), despite it requiring the highest skill level.

3.5. Task-Based Outcome Measures

Productivity. Our primary outcome is task productivity, measured as (i) the share of points achieved on a task and (ii) points achieved per minute. The first measure cap-

tures the quality of the submitted solution, whereas the second additionally accounts for how quickly the solution was produced. The Baseline task has a maximum score of 20 points and a time limit of 10 minutes; the three post-treatment tasks each have a maximum score of 40 points and a time limit of 20 minutes. For both measures, apprentices’ responses were evaluated using task-specific scoring rubrics and corresponding reference solutions. Vocational-school teachers from the target occupations reviewed and verified these materials to ensure that they accurately reflected the relevant occupational content and curriculum level (Appendix F). The rubrics and reference solutions were prepared with AI assistance prior to this expert validation.

Open-ended responses were graded criterion-by-criterion using GPT-5.4 mini, which evaluated each answer based on the task materials, the apprentice’s response, and rubric-specific scoring instructions, without information about treatment assignment or AI use.¹³ Because LLM-based grading may also be sensitive to answer length, we verify that assigned points are robust to answer length. We further find that estimated treatment effects are similar when responses are graded using alternative LLMs.¹⁴

Numerical subtasks were graded by comparing responses to predefined correct answers. This applied in particular to the System Integration Within task (IPv4 subnet-planning), where correct answers consist of exact numerical values (e.g., /20). Total task scores were constructed as the sum of criterion-level points. Points per minute were calculated by dividing task scores by task completion time recorded by the survey tool.

Comprehension. After completing each task, apprentices answered a short set of multiple-choice comprehension questions covering the core concepts of the task. Most questions assessed concrete understanding of the task solution, including the reasoning underlying key steps and results. The Baseline task was followed by three comprehension questions, while the remaining tasks were followed by four questions each. Each question had exactly one correct answer, and “I don’t know” was included as a response option to discourage

¹³Using ChatGPT to evaluate responses that may themselves contain ChatGPT-generated text raises the concern that the grader could favor AI-generated answers. Zubaer et al. (2025) directly examine this possibility by comparing GPT-4 assessments with those of trained human examiners and find no evidence that GPT-4 systematically ranks or scores GPT-generated answers more favorably than human-written answers. In our setting, this concern is further mitigated by grading each response criterion-by-criterion using predefined rubrics and fixed point maxima, while keeping the grader blind to treatment assignment and applying the identical procedure to all responses.

¹⁴Appendix Table A.3 provides an overview of the average treatment effects when graded with different models. Overall, treatment effects are similar across grading models, regardless of model type (smaller vs. larger models), model vendor (ChatGPT vs. Gemini), and whether grading is aggregated across all models (Schoenegger et al., 2024).

guessing. Our comprehension outcome is the share of questions answered correctly. The full set of comprehension questions is reported in Appendix H.

Task Experience. After each task, apprentices reported their subjective task experience on five-point Likert scales. We focus on two measures: task enjoyment (“I enjoyed working on this task”) and frustration (“The task was frustrating”). Each measure is standardized using the control-group mean and standard deviation. For analyses by subsample, we standardize each measure using the corresponding subsample-specific control-group distribution.¹⁵

4. Empirical Strategy

Treatment Effect Estimation. We estimate treatment-control differences separately for the four tasks (Baseline, Within, Beyond 1, and Beyond 2). For the three post-treatment tasks, these differences identify the intent-to-treat (ITT) effect of AI access.¹⁶ For each task, we report estimates for the full sample and separately for the Application Development and System Integration subsamples, using the following specification:

$$Y_i = \beta_0 + \beta_1 \mathbf{1}[\text{treat}_i] + X_i' \gamma + \mu_{s(i)} + \varepsilon_i, \quad (1)$$

where Y_i is the outcome of interest for apprentice i , $\mathbf{1}[\text{treat}_i]$ is an indicator equal to one if apprentice i was assigned to the treatment condition—i.e. was assigned to receive access to ChatGPT after the Baseline task (see Section 3.3)—and $\mu_{s(i)}$ denotes fixed effects for the experimental session in which apprentice i participated. The vector X_i contains apprentice-level controls. Because we estimate separate regressions for each outcome and task, each regression contains one observation per apprentice. We report robust standard errors and, following Abadie et al. (2023), cluster them at the level of treatment assignment, which in our design is the individual apprentice.

Because treatment is randomly assigned, control variables are not required for identification. They can, however, improve precision. We therefore report both unadjusted estimates, which omit the control vector X_i and session fixed effects $\mu_{s(i)}$, and adjusted

¹⁵In a deviation from the preregistration, we do not use the preregistered index combining task enjoyment, frustration, and boredom (“I found the task boring”). Instead, we report enjoyment and frustration separately. We include these two outcomes in the multiple-hypothesis-testing adjustment described in Appendix Table D.1. Treatment effects on boredom are reported in Appendix Table E.1.

¹⁶Due to high compliance, the ITT estimate is close to a TOT estimate. Only 15 apprentices are never-takers.

estimates that include session fixed effects and controls for apprenticeship specialization, gender, age, randomized task order, high pre-apprenticeship education (a university-of-applied-sciences entrance qualification or higher qualification), and whether the apprentice’s final examination took place during the field period.¹⁷

Heterogeneity. In addition to average ITT effects, we examine heterogeneity along two prespecified dimensions. First, we split the sample by AI literacy, measured using six items in the post-task survey. Apprentices rated their agreement with each statement on a five-point scale.¹⁸ Second, we examine heterogeneity by pre-apprenticeship education by interacting treatment assignment with an indicator for whether the apprentice holds at least a university-of-applied-sciences entrance qualification.

5. Results

We start by presenting the effects of AI access on productivity across the experimental tasks in Section 5.1. Section 5.2 then examines whether these productivity gains come at the expense of task comprehension, while Section 5.3 analyzes effects on task experience, i.e., frustration and enjoyment. Section 5.4 investigates heterogeneity in treatment effects by AI literacy and pre-apprenticeship education. Section 5.5 provides descriptive evidence on apprentices’ interactions with AI during the tasks, and section 5.6 presents robustness checks.¹⁹

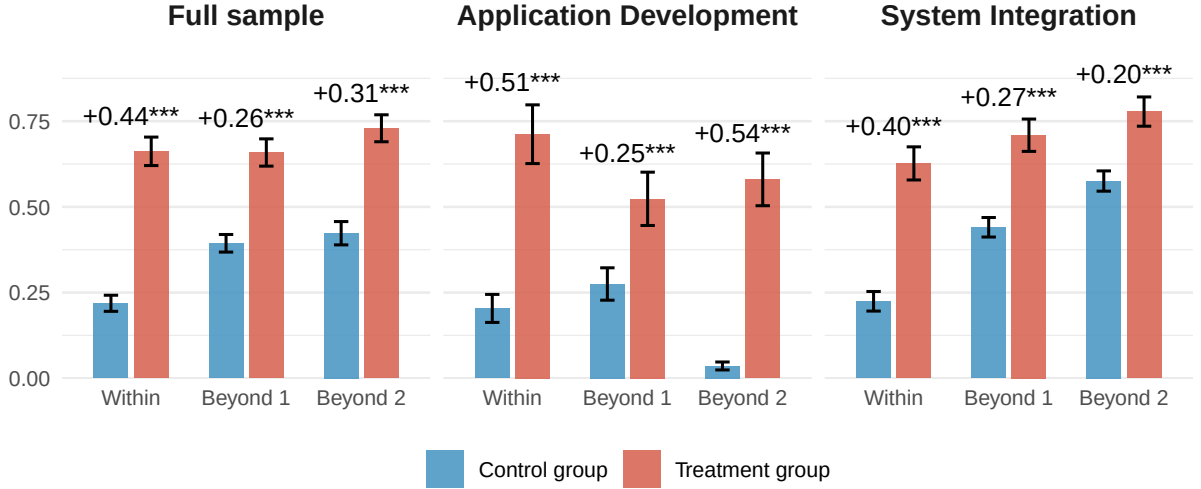
¹⁷To retain observations with missing covariate values, we impute missing values for pre-apprenticeship education and the timing of the final examination and include corresponding imputation indicators.

¹⁸The AI-literacy items ask apprentices to what extent they agree that they (i) understand how ChatGPT generates answers, (ii) can recognize errors in ChatGPT’s answers, (iii) can ask ChatGPT appropriate and effective questions, (iv) can use ChatGPT to support their decision-making process, (v) can use ChatGPT to work more efficiently and productively, and (vi) can use ChatGPT to achieve their learning and educational goals. To construct the index, we standardize each item and average the standardized responses. We then split apprentices at the sample median into lower- and higher-AI-literacy groups. Treatment assignment has no detectable effect on either the composite index or the median-split indicator. We therefore use AI literacy as a moderator, while interpreting the resulting heterogeneity with some caution because the measure was collected after treatment exposure.

¹⁹Our analysis closely follows our pre-analysis plan, with minor deviations documented in Appendix E.

5.1. Productivity Effects

Figure 1: Average Treatment Effects on Productivity (Share of Points Achieved), by Task



Notes: Outcome is the share of points achieved on each task. Tasks are ordered by required skill level: Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). Bars show control and treatment group means; error bars are 95% confidence intervals. Treatment effects are from OLS regressions with controls; heteroscedasticity-robust standard errors. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Share of Points Achieved. Figure 1 reports ITT effects of AI access on the share of points achieved on each task (see Appendix Table A.4 for regression results). AI access substantially increases productivity across all post-Baseline tasks. In the full sample, it raises productivity by 44 percentage points on the Within task and by 26 and 31 percentage points on the Beyond 1 and Beyond 2 tasks, respectively. The sizable gains on tasks beyond apprentices' formal training provide direct evidence of AI-enabled task expansion.

Among Application Development apprentices, AI increases productivity across all three tasks, but the gains differ substantially by unaided performance. The largest effect is observed on the Beyond 2 machine-learning task, where AI access raises productivity by 54 percentage points. Without AI, apprentices achieve only 4% of the available points on this task. AI also substantially improves performance on the Within task, increasing the share of points achieved by 51 percentage points from a control-group mean of 20%. On the Beyond 1 task, control-group performance is somewhat higher at 28%, and AI raises performance by 25 percentage points. Thus, the largest productivity gain occurs on the task with the lowest unaided performance.

Among System Integration apprentices, AI also increases performance across all three tasks. The largest effect is observed on the Within task, where AI access raises perfor-

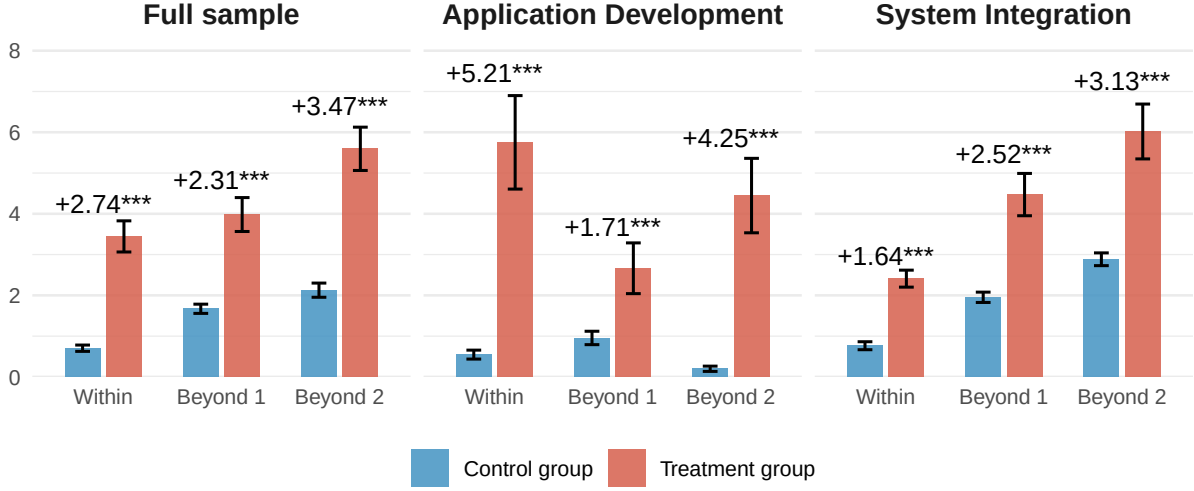
mance by 40 percentage points from a control-group mean of 23%. This task requires IPv4 subnet planning and is part of apprentices' formal training. One possible explanation for the particularly large gain is that apprentices already have the knowledge needed to evaluate and effectively use the AI output. At the same time, AI also substantially improves performance on both tasks beyond the apprenticeship curriculum, increasing the share of points achieved by 27 percentage points on Beyond 1 and 20 percentage points on Beyond 2.

Extensive Margin. The low level of control-group performance suggests that part of the treatment effect may operate on the extensive margin, with AI encouraging apprentices to engage with tasks they might otherwise not attempt. To examine this, Appendix Table A.5 and Figure A.6 report treatment effects on whether apprentices submit a substantive answer.²⁰ AI increases the probability of submitting a substantive answer by 11 percentage points on the Within task and by 6 percentage points on Beyond 2 in the full sample, while the effect on Beyond 1 is small and statistically insignificant. These extensive-margin effects are thus strongest on tasks where apprentices are less likely to engage without AI.

Appendix Table A.6 examines performance among apprentices who submit a substantive answer. Within this selected sample, apprentices with AI access achieve higher scores across all post-treatment tasks, both in the full sample and within each specialization. In the full sample, these differences range from 27 to 43 percentage points. Because AI access itself affects the probability of submitting an answer, however, these conditional comparisons are descriptive and cannot be interpreted as causal treatment effects.

²⁰We define 'no answer submitted' as no answer containing content relevant to the task. This also accounts for cases in which apprentices entered a placeholder (e.g., ".") rather than leaving the response field blank.

Figure 2: Average Treatment Effects on Productivity per Minute, by Task



Notes: Outcome is the number of points achieved per minute spent on the task. Tasks are ordered by required skill level: Within (within curriculum), Beyond 1 and Beyond 2 (beyond curriculum). Bars show control and treatment group means; error bars are 95% confidence intervals. Treatment effects are from OLS regressions with controls; heteroscedasticity-robust standard errors. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Points per Minute. While the previous analysis focused on task performance, Figure 2 and Appendix Table A.7 additionally account for the time apprentices spend on each task by examining points achieved per minute.²¹ This measure captures whether apprentices achieve higher scores relative to the time spent on the task. The results reinforce the pattern observed for task performance. Treatment effects on the post-treatment tasks Within, Beyond 1 and Beyond 2 are large and statistically significant. Apprentices in the treatment group achieve on average between 2.3 and 3.5 additional points per minute, with the largest gain observed on Beyond 2. The pattern differs somewhat across the two specialization tracks. AI access generates particularly large gains in points per minute on the Within task in the Application Development track, where apprentices analyze a long code snippet, while the effect is smallest on the Within task in the System Integration track, where apprentices enter calculated results into tables, limiting the scope for time savings from AI. Effects on both Beyond tasks in both tracks are large and highly signif-

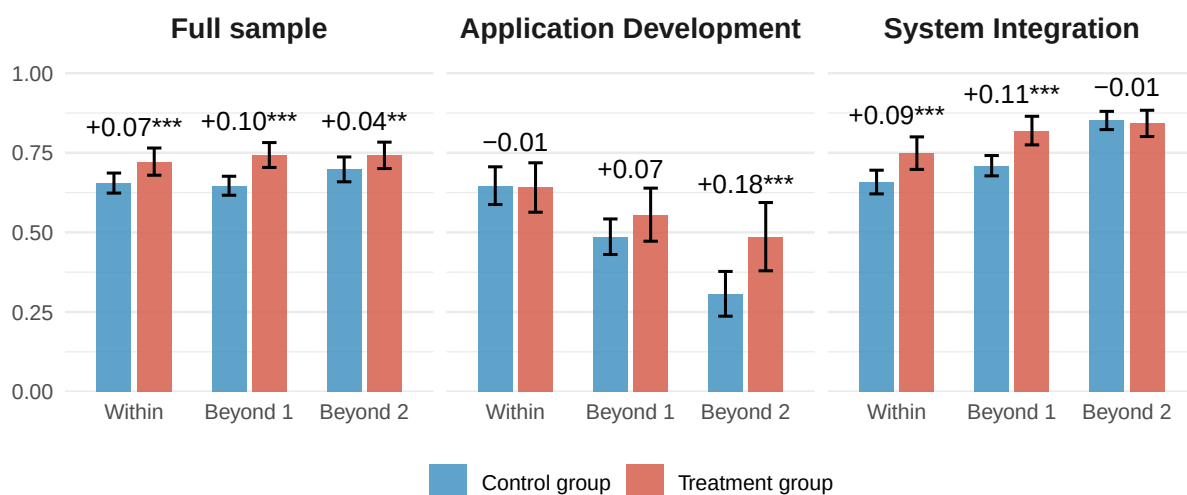
²¹The variables for points per minute exhibit a right-skewed distribution, as depicted in Appendix Figure A.2. To ensure that the results are not driven by this skewness, we estimate the same models using a winsorized version of the variable. The results remain robust under this specification and are provided in Appendix Table A.8. Appendix Figures A.1 to A.5 show that the other outcome variables are approximately normally distributed.

icant, ranging from 1.7 to 4.3 additional points per minute for Application Development and from 2.5 to 3.1 additional points per minute for System Integration.

Taken together, the evidence suggests that AI access—in addition to enabling task expansion—substantially increases the quality of apprentices’ solutions relative to the time spent on the task.

5.2. Effects on Comprehension

Figure 3: Average Treatment Effects on Comprehension (Share of Questions Correct), by Task



Notes: Outcome is the share of comprehension questions answered correctly after each task. Tasks are ordered by required skill level: Within (within curriculum), Beyond 1 and Beyond 2 (beyond curriculum). Bars show control and treatment group means; error bars are 95% confidence intervals. Treatment effects are from OLS regressions with controls; heteroscedasticity-robust standard errors. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The productivity gains documented above do not necessarily imply a corresponding increase in apprentices’ understanding of the material. To examine this, Figure 3 (see Table A.9 for the regression table) reports treatment effects on comprehension, measured as the share of comprehension questions answered correctly after each task.

In the full sample, the treatment-control difference for the pre-treatment Baseline task is close to zero and statistically insignificant, consistent with random assignment. By contrast, AI access increases comprehension by 6.7 percentage points on the Within task and 9.7 percentage points on Beyond 1. The effect on Beyond 2 is smaller (4.4 percentage points) but remains statistically significant.

The results across specializations also vary with task difficulty. Among Application Development apprentices, comprehension gains are concentrated on Beyond 2 (17.9 percentage points), the task with by far the lowest control-group performance (30.7%). Among System Integration apprentices, treatment effects are concentrated on the Within and Beyond 1 tasks (9.1 and 11.0 percentage points, respectively), while effects on Beyond 2 are close to zero. Thus, within each specialization, comprehension gains tend to be largest on the tasks where control-group performance is lowest.

The results are robust to excluding apprentices who may have used AI during the comprehension blocks (Appendix Table A.10).²² Furthermore, the results remain unchanged when excluding apprentices who did not attempt the task. As AI increases task engagement on the extensive margin, one concern is that the positive effects on comprehension merely reflect greater engagement with the task rather than greater understanding of its content. Appendix Table A.11 shows that our results are robust to excluding these observations.

5.3. *Effects on Task Enjoyment and Frustration*

Apart from productivity and comprehension, subjective task experience is economically relevant because it may affect effort provision, persistence on difficult tasks, and the willingness to engage in similar tasks in the future. Appendix Tables A.12 and A.13 report ITT effects on two subjective task-experience outcomes. The clearest pattern emerges on the task in each specialization where performance without AI is lowest: AI access reduces frustration and raises enjoyment on these tasks.

For Application Development apprentices, treatment substantially reduces frustration on both the Within and Beyond 2 tasks, by 0.51 SD and 0.49 SD, respectively. On the Beyond 2 machine-learning task, treatment also raises enjoyment by 0.47 SD. All three effects are significant at the 1% level. For System Integration apprentices, treatment reduces frustration by 0.57 SD on the Within task, the largest frustration-reducing effect across all tasks, and raises enjoyment by 0.31 SD. On the two Beyond tasks, effects on enjoyment are small and statistically insignificant. Frustration also declines on both Beyond tasks, although the effects are considerably smaller and statistically significant only for Beyond 2 (−0.16 SD).

²²For practical reasons related to the school setting, treatment-group apprentices retained internet access during the comprehension blocks. To discourage AI use, apprentices were instructed not to use any aids, asked to confirm compliance after completing the questions, and all prompt submissions were tracked. Overall, 12.68% of treated apprentices either self-reported using AI (6.05%) or were observed submitting a prompt (9.34%) during at least one comprehension block.

Read together with the productivity and comprehension results, these findings suggest that AI can provide particularly large benefits on tasks that apprentices struggle to perform without AI. On these tasks, AI not only raises productivity without reducing comprehension, but can also make the work less frustrating and, in some cases, more enjoyable.

5.4. Heterogeneous Treatment Effects

The average treatment effects may mask substantial heterogeneity in who benefits from access to AI. If effective use of AI requires users to formulate appropriate prompts and evaluate generated outputs, gains may be larger among apprentices with higher AI literacy. At the same time, AI may complement broader forms of human capital, implying larger returns for apprentices with stronger educational backgrounds. We therefore examine treatment effect heterogeneity along these two dimensions.

Table 3: Heterogeneous Treatment Effects on Productivity (Share of Points Achieved), by AI Literacy

	Share of points achieved							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment (Low AI literacy)	-0.003 (0.025)	-0.008 (0.024)	0.368*** (0.032)	0.375*** (0.032)	0.238*** (0.031)	0.254*** (0.029)	0.253*** (0.034)	0.275*** (0.028)
Treatment (High AI literacy)	0.028 (0.024)	0.022 (0.023)	0.515*** (0.029)	0.517*** (0.029)	0.294*** (0.030)	0.290*** (0.029)	0.351*** (0.035)	0.344*** (0.029)
Difference	0.031 (0.034)	0.030 (0.033)	0.147*** (0.043)	0.142*** (0.043)	0.055 (0.043)	0.036 (0.041)	0.098** (0.048)	0.069* (0.041)
Controls		✓		✓		✓		✓
Control mean (Low AI literacy)	0.418	0.418	0.223	0.223	0.395	0.395	0.436	0.436
Control mean (High AI literacy)	0.393	0.393	0.213	0.213	0.394	0.394	0.417	0.417
Observations	653	653	653	653	653	653	653	653

Notes: Outcome is the share of points achieved on each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Treatment effects are from OLS regressions of the outcome on the treatment indicator, an indicator for above-median AI literacy, and their interaction; the interaction term identifies the differential treatment effect for apprentices with above-median AI literacy. Heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome within the corresponding AI-literacy subgroup. The full sample pools the Application Development and System Integration sub-samples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

AI Literacy. Table 3 reports treatment effects separately for apprentices with below- and above-median AI literacy. AI access increases performance significantly for both groups on all post-Baseline tasks. However, the gains are larger for apprentices with above-median AI literacy. This difference is most pronounced on the Within task, where the

treatment effect is 52 percentage points for apprentices with high AI literacy compared with 38 percentage points for those with low AI literacy, a differential of 14 percentage points significant at the 1% level. On Beyond 2, the corresponding treatment effects are 34 and 28 percentage points, with a smaller differential of 7 percentage points that is statistically significant at the 10% level. The differential on Beyond 1 is smaller and not statistically significant.

Appendix Table A.14 provides a robustness check using the individual items of the AI literacy index. The table shows that the interaction effect observed for the AI-literacy index is concentrated in the following three items: "I can use ChatGPT to support my decision-making process.", "I can use ChatGPT to work more efficiently and productively." and "I can use ChatGPT to achieve my learning and education goals.", as well as "I can formulate appropriate and effective questions for ChatGPT" for the Within task only. This suggests that the ability to use the tool effectively, rather than understanding how it works, is more strongly associated with larger treatment effects. Notably, self-reporting above-median ability to identify errors in ChatGPT responses is associated with a negative interaction effect for the Beyond 1 task.

Heterogeneity analyses by AI literacy for points achieved per minute and comprehension are reported in Appendix Tables A.15 and A.16. The results for points achieved per minute broadly support the heterogeneity observed for points achieved, while we find no consistent or significant pattern for comprehension.

Pre-Apprenticeship Education. Appendix Table A.17 reports heterogeneous treatment effects on productivity for apprentices with lower and higher pre-apprenticeship education.²³

Treatment effects are positive and significant for both education groups across all post-Baseline tasks. For the Within and Beyond 2 tasks, the effects do not differ significantly by education. On Beyond 1, however, apprentices with lower pre-apprenticeship education benefit more from AI: the treatment effect is 32 percentage points for apprentices without a university-of-applied-sciences entrance qualification, compared with 24 percentage points for those with this qualification, a difference of 8 percentage points that is statistically significant at the 10% level. The robustness check using the general university entrance qualification as the cutoff shows a similar pattern, although the difference on Beyond 1 is no longer statistically significant.

²³Appendix Table A.18 provides a robustness check, with a split based on whether apprentices hold a general university entrance qualification. Results remain robust.

Appendix Tables A.19 and A.20 report heterogeneity in points per minute and comprehension, respectively. We find no significant differences by pre-apprenticeship education for either outcome.

Overall, the results suggest that larger AI gains are more consistently associated with apprentices' self-reported ability to use the tool effectively than with conventional educational attainment. This finding is in line with the descriptive evidence from Zisler and Backes-Gellner (2026), who use survey data on entry-level IT workers with migration backgrounds and find that AI-attributed shifts toward a more complex task portfolio are associated with active AI use rather than with educational background.

5.5. *Interaction with the AI tool*

To better understand the treatment effects documented above, this section descriptively examines how apprentices use the AI tool when access is provided. We first document treatment take-up and then examine how usage patterns are associated with productivity and comprehension. More detailed analyses can be found in Appendix B.

Who is Using AI? We define never-takers as people in the treatment group who visited ChatGPT for less than fifteen seconds and created no prompts, which is the case for 15 apprentices.²⁴ This high treatment take-up is consistent with survey evidence on apprentices' pre-existing AI experience (see Appendix Table B.1). AI use is nearly universal: only 5% report never using such tools, while the large majority use them at vocational school (84%), at their training firm (77%), and privately (77%). AI users also report higher prior AI use and self-reported AI literacy than never-takers (Appendix Table B.2). Additional comparisons of AI users and never-takers in productivity, comprehension, task experience, and effort are reported in Appendix Tables B.3 and B.4. Given that only 15 apprentices are classified as never-takers, these comparisons are descriptive and should be interpreted with caution.

Usage Behavior: How do Apprentices Use AI? In addition to whether apprentices use AI at all, we examine how they interact with the tool once access is provided. We characterize prompting intensity, prompt types, and the incorporation of AI-generated content into submitted answers. Prompt-level interaction data are available for 65% of the treatment group; Appendix Table B.5 shows that data availability is not systematically related to education, prior AI use, or AI use during the experiment.

²⁴This excludes $N = 11$ observations for which AI interaction cannot be observed.

Among AI users, apprentices submit 6.3 prompts on average across the three post-treatment tasks. Around half of all prompts copy task material into ChatGPT, while roughly one quarter combine copied task material with further instructions. Requests for explanations and prompts that transform task material are considerably less frequent. We also observe that apprentices directly incorporate AI-generated content into their submitted answers. Overall, these patterns show substantial variation in how apprentices interact with AI. Detailed task- and specialization-level descriptive statistics are reported in Appendix Table B.6.

Usage Behavior and Productivity. We next examine whether differences in AI usage are associated with productivity. Table 4 relates prompting intensity, prompt types, and the incorporation of AI-generated content to productivity across the three post-treatment tasks. These associations are descriptive and should not be interpreted as causal.

The strongest and most consistent associations concern copying behavior. Prompts that copy task text into ChatGPT are positively associated with productivity across all three tasks. Direct incorporation of ChatGPT output into submitted answers is also strongly associated with higher productivity: in the multivariate specifications, exact copying is associated with increases in the share of points achieved of 18 percentage points on Within, 20 percentage points on Beyond 1, and 13 percentage points on Beyond 2. More intensive prompting is also positively associated with productivity, although this association is statistically significant only for Beyond 1 once other dimensions of AI usage are taken into account. Other prompt types show less consistent associations with productivity.

Usage patterns associated with higher productivity are generally not associated with lower comprehension. Appendix Table B.7 reports the detailed results; neither prompting intensity nor copying AI output into the final answer is associated with lower comprehension across the three tasks.

Table 4: Association Between AI Usage and Productivity (Among AI Users), by Task

	Within		Beyond 1		Beyond 2	
	Bivar.	Multivar.	Bivar.	Multivar.	Bivar.	Multivar.
	(1)	(2)	(3)	(4)	(5)	(6)
Prompt input						
No. of prompts	0.017 (0.013)	0.005 (0.012)	0.086*** (0.010)	0.034** (0.016)	0.064*** (0.012)	0.019 (0.014)
Copy task text	0.152*** (0.042)	0.182*** (0.047)	0.233*** (0.041)	0.116*** (0.044)	0.236*** (0.034)	0.142*** (0.049)
Instruction + copy task	0.104** (0.045)	0.163*** (0.046)	0.122*** (0.043)	0.053 (0.039)	0.051 (0.044)	0.020 (0.040)
Transform response	0.034 (0.073)	-0.040 (0.084)	0.223*** (0.037)	0.005 (0.044)	0.154*** (0.041)	0.015 (0.051)
Explain	-0.116* (0.064)	-0.126* (0.066)	-0.041 (0.047)	0.018 (0.042)	-0.031 (0.063)	-0.020 (0.050)
Usage of AI output						
Answer exact copy	0.261*** (0.078)	0.177** (0.079)	0.320*** (0.033)	0.204*** (0.038)	0.226*** (0.032)	0.127*** (0.037)
Observations	225		225		225	

Notes: Outcome is the share of points achieved (Productivity). Column pairs correspond to three tasks ordered by required skill level: Within (within curriculum), Beyond 1 and Beyond 2 (beyond curriculum). Odd-numbered columns (Bivar.) report coefficients from separate OLS regressions, each including one AI usage measure at a time. Even-numbered columns (Multivar.) report coefficients from a single OLS regression including all AI usage measures jointly. All regressions control for age, gender, education, whether the final exam took place during the field period, and session fixed effects. The usage measures are: number of prompts submitted to ChatGPT during the task (count variable); and four indicator variables equal to one if at least one prompt was classified as copy, explain, instructed copy, or transform, respectively; and an indicator for whether the apprentice’s answer was copied from ChatGPT (LCS score = 1). Heteroscedasticity-robust standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.6. Robustness Checks

We conduct a broad set of robustness checks to assess the validity of our findings (see Appendix D). Throughout the paper, we have already shown that our results are robust to alternative definitions of the outcome variables (points per minute, see Section 5.1) and heterogeneity measures (pre-apprenticeship education and sub-items of AI literacy, see Section 5.4), to excluding treatment-group apprentices who visited other websites during the experiment (see Section 3.3), and, for the comprehension outcomes, to excluding apprentices who may have used AI during the comprehension blocks (see Section 5.2). In the appendix, we further examine the role of multiple hypothesis testing (Appendix Table D.1), experimenter-demand effects (Appendix Table D.2), participant motivation, the inclusion of apprentices from related specializations, classroom characteristics (Appendix

Tables D.3 to D.8), and changes in ChatGPT’s default model over the data-collection period (Appendix Table D.9). Across all specifications, our main conclusions remain unchanged.

6. Conclusion

Generative AI is rapidly entering the workplaces in which middle-skilled, early-career workers acquire the skills that will shape their careers. Using a preregistered field experiment with 673 final-year IT apprentices at twelve German vocational schools, we provide causal evidence that AI access raises productivity by 26 to 44 percentage points across tasks within and beyond apprentices’ formal training. Productivity increases by 26 to 31 percentage points on tasks deliberately designed beyond the apprenticeship curriculum, providing direct evidence of AI-enabled task expansion. At the same time, the largest treatment effect arises on the within-curriculum task, showing that AI can also substantially amplify performance within workers’ existing domain. AI further increases the probability that apprentices engage with tasks they might otherwise not attempt, particularly on the task in each specialization where unaided performance is lowest. Importantly, these productivity gains do not come at the expense of understanding: AI access tends to improve subsequent task comprehension, with no evidence of reductions in understanding. AI also improves task experience, reducing frustration and increasing enjoyment, with effects tending to be larger on tasks where performance without AI is lower.

In terms of policy implications, the more consequential results concern heterogeneity in apprentices’ ability to use AI effectively. Directly incorporating model output into submitted answers is associated with higher productivity, without being systematically associated with lower comprehension. At the same time, apprentices with higher self-reported AI literacy generally benefit more from AI, while treatment effects do not vary systematically with pre-apprenticeship education. The heterogeneity by AI literacy is concentrated in measures capturing apprentices’ ability to use ChatGPT effectively, rather than their understanding of how the technology works. These are learnable skills, but only 16% of apprentices report learning them at vocational school. Vocational education and training therefore faces a practical mismatch: skills associated with larger gains from AI are not yet systematically taught, while standardized examinations still tend to measure the kinds of outputs that AI can readily supply.

Like any field experiment, our study is bounded in scope and three features of the design are important for interpreting the estimates. First, the treatment compares AI with

internet access to a setting with no internet access. The results should therefore be interpreted as the effect of AI access relative to a no-internet condition and may overstate the effect relative to an internet-only comparison, since workers would typically have access to online resources at work even without access to AI tools. Second, our comprehension measures capture short-run unaided comprehension rather than long-term learning. Apprentices answered multiple-choice comprehension questions immediately after each task. Thus, the estimates show what they retained at that moment, not whether they retained the material over time. To assess longer-run effects, future work would need to include reassessments, ideally after days or weeks, and measure whether apprentices can recall, apply, and transfer the content without immediate support from AI. Third, our evidence comes from two IT apprenticeships within the German dual system, which we view as a feature of our study: IT apprentices are among the middle-skilled workers most directly exposed to AI, the two occupations are large and nationally standardized, and the institutional setting holds task content, assessment, and work environment unusually constant — conditions that are difficult to replicate in less regulated occupational contexts. Establishing whether the same patterns extend to other middle-skilled occupations, and to apprenticeship systems outside Germany, is a natural and important next step.

A further consideration is the rapid pace of technological change. Our estimates are necessarily tied to the capabilities of the AI models available in the experiment and may change as models improve.²⁵ At the same time, the broader mechanism we study is not specific to a particular model or interface. As AI systems become more capable and increasingly agentic, the set of tasks workers can delegate to or perform with AI is likely to expand further. Our results give a strong starting point: in a highly AI-exposed early-career segment, access to the tool expands the range of tasks that middle-skilled workers can perform without weakening their understanding of the underlying material.

²⁵Relatedly, trust in AI and patterns of AI adoption may themselves evolve as the technology develops. An important direction for future research is therefore to study whether AI-induced productivity gains vary with workers’ trust in AI and their propensity to adopt and use the technology, both of which have been shown to vary substantially across countries (Hazra et al., 2026; Bick et al., 2026a).

References

- Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2023). When Should You Adjust Standard Errors for Clustering? *The Quarterly Journal of Economics* 138(1), 1–35.
- Autor, D. (2024). Applying AI to Rebuild Middle Class Jobs. Working Paper 32140, National Bureau of Economic Research.
- Autor, D., T. Rodchenko, J. Martin, Z. Iscenko, S. Strand, D. Pearl, and M. Ferere (2026). Does AI Assistance Enhance or Erode Expertise? Evidence from a Three-Month Field Experiment in Patent Drafting. Working Paper 35720, National Bureau of Economic Research.
- Awuah, K., U. Krenk, and D. Yanagizawa-Drott (2025). Augment or Automate? An Early Field Experiment with Generative AI in Hiring. SSRN Scholarly Paper 5351335.
- Bastani, H., O. Bastani, A. Sungu, H. Ge, Kabakcı, and R. Mariman (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences* 122(26), e2422633122.
- BIBB (2026). Datenreport zum Berufsbildungsbericht 2026. Technical report, Bundesinstitut fuer Berufsbildung (BIBB).
- Bick, A., A. Blandin, D. Deming, N. Fuchs-Schündeln, and J. Jessen (2026a). Mind the Gap: AI Adoption in Europe and the U.S. Working Paper w34995, National Bureau of Economic Research.
- Bick, A., A. Blandin, and D. J. Deming (2026b). The Rapid Adoption of Generative AI. *Management Science*.
- Brynjolfsson, E., B. Chandar, and R. Chen (2026). Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence. *Stanford Digital Economy Lab Working Paper*.
- Brynjolfsson, E., D. Li, and L. Raymond (2025). Generative AI at Work. *The Quarterly Journal of Economics* 140(2), 889–942.
- Cruces, G., D. Fernandez Meijide, S. Galiani, R. H. Galvez, and M. Lombardi (2026). Does Generative AI Narrow Education-Based Productivity Gaps? Evidence from a Randomized Experiment. Working Paper 34851, National Bureau of Economic Research.
- Cui, K. Z., M. Demirer, S. Jaffe, L. Musolff, S. Peng, and T. Salz (2026). The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers. *Management Science*.
- Dell’Acqua, F., E. McFowland, E. Mollick, H. Lifshitz, K. C. Kellogg, S. Rajendran, L. Kraye, F. Candelon, and K. R. Lakhani (2026a). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. *Organization Science* 37(2), 403–423.
- Dell’Acqua, F., C. Ayoubi, H. Lifshitz, R. Sadun, E. Mollick, L. Mollick, Y. Han, J. Goldman, H. Nair, S. Taub, and K. R. Lakhani (2026b). The Cybernetic Teammate: A Field Experiment on Generative AI and Teamwork. *Organization Science* 37(4), 1217–1242.

- Gambacorta, L., H. Qiu, S. Shan, and D. M. Rees (2026). Generative AI and Labour Productivity: A Quasi Experiment on Coding. *Journal of Financial Stability* 84, 101543.
- Haslberger, M., J. Gingrich, and J. Bhatia (2025). No Great Equalizer: Experimental Evidence on Productivity Effects of Generative AI Use in the UK Labor Market. SSRN Scholarly Paper 4594466.
- Hazra, S., Y. Park, and M. Serra-Garcia (2026). Uneven Trust in AI: Experimental and Global Survey Evidence.
- Humlum, A. and E. Vestergaard (2025). The Unequal Adoption of ChatGPT Exacerbates Existing Inequalities Among Workers. *Proceedings of the National Academy of Sciences* 122(1), e2414972121.
- IAB (2017). Berufe im Spiegel der Statistik. Online Source, Institute for Employment Research.
- Jabarian, B. and L. Henkel (2026). Voice AI in Firms: A Natural Field Experiment on Automated Job Interviews. CESifo Working Paper No. 12984.
- Kanazawa, K., D. Kawaguchi, H. Shigeoka, and Y. Watanabe (2026). AI, Skill, and Productivity: The Case of Taxi Drivers. *Management Science* 72(2), 1376–1388.
- Lambert, P. J. and Y. Schindler (2026). The Broken Ladder: AI, Remote Work, and Early-Career Hiring. SSRN Scholarly Paper 6787638.
- Langer, C. and S. Wiederhold (2023). The Value of Early-Career Skills. *CESifo Working Paper No. 10288*.
- Lichtinger, G. and S. M. Hosseini Maasoum (2025). Generative AI as Seniority-Biased Technological Change: Evidence from U.S. Resume and Job Posting Data. SSRN Scholarly Paper 5425555.
- List, J. A., A. M. Shaikh, and Y. Xu (2019). Multiple Hypothesis Testing in Experimental Economics. *Experimental Economics* 22(4), 773–793.
- Mehmood, S., C. Goessmann, and E. Ash (2026). Courts of Tomorrow: Evidence from a Nationwide Rollout of Generative AI. SSRN Scholarly Paper 7119447.
- Noy, S. and W. Zhang (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science* 381(6654), 187–192.
- Otis, N. G., R. Clarke, S. Delecourt, D. Holtz, and R. Koning (2026). The Uneven Impact of Generative Artificial Intelligence on Entrepreneurial Performance: Evidence from a Field Experiment in Kenya. *Management Science*.
- Paradis, E., K. Grey, Q. Madison, D. Nam, A. Macvean, V. Meimand, N. Zhang, B. Ferrari-Church, and S. Chandra (2024). How Much Does AI Impact Development Speed? An Enterprise-Based Randomized Controlled Trial. arXiv:2410.12944 [cs.SE].
- Peng, S., E. Kalliamvakou, P. Cihon, and M. Demirer (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv:2302.06590.

- Schoenegger, P., I. Tuminauskaite, P. S. Park, R. V. S. Bastos, and P. E. Tetlock (2024). Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy. *Science Advances* 10(45), eadp1528.
- Shen, J. H. and A. Tamkin (2026). How AI Impacts Skill Formation. arXiv:2601.20245 [cs.CY].
- Tambe, P. B. (2026). Reskilling the Workforce for AI: Domain Expertise and Algorithmic Literacy. *Management Science* 72(1), 515–537.
- Yeverechyahu, D., R. Mayya, and G. Oestreicher-Singer (2026). The Impact of Large Language Models on Open-Source Innovation: Evidence from GitHub Copilot. SSRN Scholarly Paper 4684662.
- Zisler, C. and U. Backes-Gellner (2026). Generative AI and Task Recomposition in Entry-Level IT Work: A Worker-Side Task-Level Measure. *Economics of Education Working Paper Series*. Number: 0258.
- Zubaer, A. A., M. Granitzer, S. Geschwind, J. Graf Lambsdorff, and D. Voss (2025). GPT-4 Shows Comparable Performance to Human Examiners in Ranking Open-Text Answers. *Scientific Reports* 15(1), 35045.

Online Appendix

Appendix A. Additional Figures and Tables

Table A.1: Descriptive Statistics

	(1)	(2)	(3)	(4)	(5)
	Mean	SD	Min	Max	N
Experimental variables					
Treatment	0.52	0.50	0.00	1.00	673
Task order (within, then beyond)	0.53	0.50	0.00	1.00	673
Descriptives					
Age	23.47	4.30	18.00	45.33	673
Female	0.09	0.28	0.00	1.00	673
University entrance qualification	0.59	0.49	0.00	1.00	673
Apprenticeship					
Apprenticeship length in years	2.80	0.33	2.00	3.50	632
Specialization					
<i>Application Development</i>	0.30	0.46	0.00	1.00	673
<i>System Integration</i>	0.70	0.46	0.00	1.00	673
Average grade (interim exam)	2.80	0.86	1.00	4.40	579
Writes final exam during field phase	0.14	0.35	0.00	1.00	673
Outcomes in baseline tasks					
Share of received points	0.41	0.22	0.00	1.00	673
Points achieved per minute	1.89	1.15	0.00	8.00	673
Comprehension: Share correct	0.87	0.22	0.00	1.00	673
Frustrating task experience	0.01	1.00	-1.20	2.39	673
Fun task experience	-0.05	0.97	-1.91	1.80	673
Observations	673				

Notes: This table shows descriptive statistics for our experimental sample.

Table A.2: Task Scores and Pre-Apprenticeship Education

	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average grade (interim exam, reverse-coded)	0.068*** (0.016)	0.054*** (0.019)	0.079*** (0.016)	0.071*** (0.017)	0.049*** (0.017)	0.058*** (0.018)	0.011 (0.023)	0.047*** (0.016)
Application development		0.103*** (0.034)		-0.022 (0.035)		-0.187*** (0.036)		-0.527*** (0.030)
Age		-0.007** (0.003)		0.002 (0.003)		-0.004 (0.003)		-0.001 (0.003)
Female		0.030 (0.050)		-0.053* (0.030)		-0.085 (0.052)		-0.036 (0.041)
University entrance qualification		0.094*** (0.026)		0.006 (0.026)		0.036 (0.031)		0.040 (0.027)
Task order (within, then beyond)		0.042* (0.022)		0.025 (0.024)		-0.050* (0.026)		-0.075*** (0.023)
Writes final exam during field phase		0.013 (0.033)		0.128*** (0.046)		0.071* (0.039)		0.020 (0.034)
Mean of dep. var.	0.404	0.404	0.219	0.219	0.394	0.394	0.423	0.423
N	326	326	326	326	326	326	326	326

Notes: OLS estimates. Dependent variable: share of points achieved. Even columns include controls for age, gender, task order, session fixed effects, and educational background. Missing dummies for schooling, exam date, and interim exam grade are included but not reported. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3: Overview of Average Treatment Effects, When Graded With Different AI Models

Model	Full Sample			Application Development			System Integration		
	Within	Beyond 1	Beyond 2	Within	Beyond 1	Beyond 2	Within	Beyond 1	Beyond 2
GPT-5.4 mini	0.443*** (0.021)	0.265*** (0.020)	0.306*** (0.020)	0.508*** (0.044)	0.249*** (0.040)	0.544*** (0.039)	0.402*** (0.025)	0.269*** (0.024)	0.203*** (0.022)
GPT-5.4 nano	0.448*** (0.021)	0.282*** (0.020)	0.333*** (0.019)	0.525*** (0.042)	0.261*** (0.037)	0.537*** (0.038)	0.402*** (0.025)	0.287*** (0.024)	0.245*** (0.021)
Gemini 3.5 flash	0.452*** (0.022)	0.290*** (0.021)	0.332*** (0.021)	0.540*** (0.045)	0.274*** (0.040)	0.582*** (0.041)	0.402*** (0.025)	0.296*** (0.025)	0.225*** (0.023)
Gemini 3.5 flash lite	0.454*** (0.021)	0.292*** (0.021)	0.341*** (0.021)	0.548*** (0.044)	0.254*** (0.039)	0.587*** (0.041)	0.402*** (0.025)	0.307*** (0.025)	0.235*** (0.023)
Aggregated (mean)	0.448*** (0.021)	0.282*** (0.020)	0.326*** (0.020)	0.525*** (0.042)	0.266*** (0.038)	0.551*** (0.039)	0.402*** (0.025)	0.288*** (0.024)	0.228*** (0.022)
Aggregated (median)	0.453*** (0.021)	0.287*** (0.020)	0.332*** (0.020)	0.543*** (0.044)	0.272*** (0.039)	0.560*** (0.039)	0.402*** (0.025)	0.291*** (0.024)	0.234*** (0.022)

Notes: Each cell reports the average treatment effect for share of points achieved on the corresponding task, when graded with the corresponding model, separately for the full sample, Application Development subsample, and System Integration subsample. AI grading models include ChatGPT-5.4 mini, ChatGPT-5.4 nano, Gemini 3.5 flash, Gemini 3.5 flash lite, as well as aggregated versions of all models, using the mean and the median grading score across models, respectively. Heteroscedasticity-robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

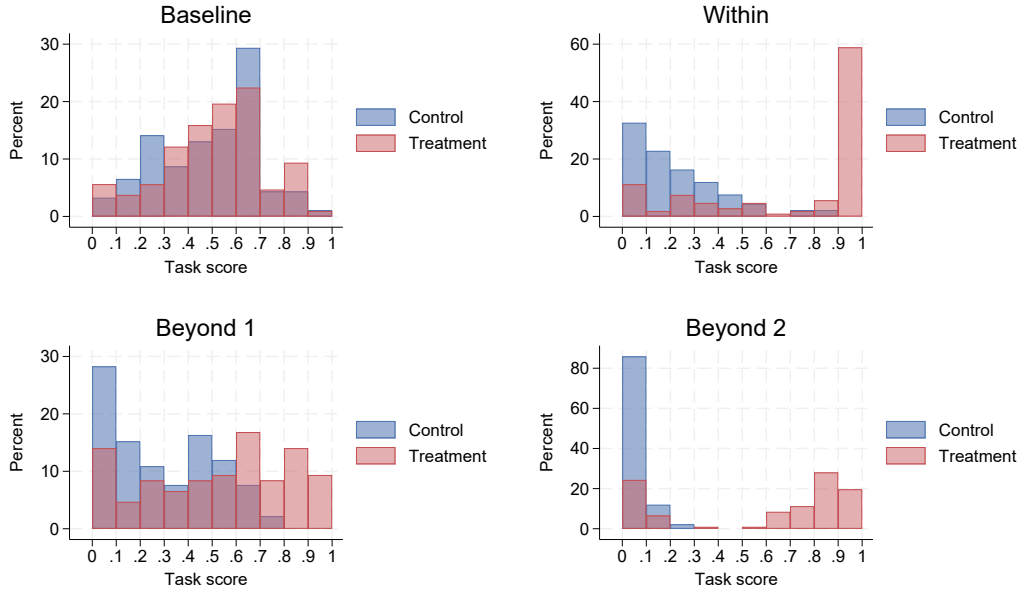
Table A.4: Average Treatment Effects on Productivity (Share of Points Achieved), by Task

	Share of points achieved							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Full sample								
Treatment	0.016 (0.017)	0.009 (0.016)	0.442*** (0.021)	0.443*** (0.021)	0.259*** (0.022)	0.265*** (0.020)	0.298*** (0.024)	0.306*** (0.020)
Controls		✓		✓		✓		✓
Control mean	0.404	0.404	0.219	0.219	0.394	0.394	0.423	0.423
Observations	673	673	673	673	673	673	673	673
Application Development								
Treatment	0.024 (0.031)	0.021 (0.032)	0.528*** (0.041)	0.508*** (0.044)	0.244*** (0.038)	0.249*** (0.040)	0.552*** (0.037)	0.544*** (0.039)
Controls		✓		✓		✓		✓
Control mean	0.473	0.473	0.204	0.204	0.275	0.275	0.036	0.036
Observations	199	199	199	199	199	199	199	199
System Integration								
Treatment	0.009 (0.019)	0.009 (0.019)	0.404*** (0.024)	0.402*** (0.025)	0.271*** (0.025)	0.269*** (0.024)	0.206*** (0.022)	0.203*** (0.022)
Controls		✓		✓		✓		✓
Control mean	0.376	0.376	0.225	0.225	0.440	0.440	0.575	0.575
Observations	474	474	474	474	474	474	474	474

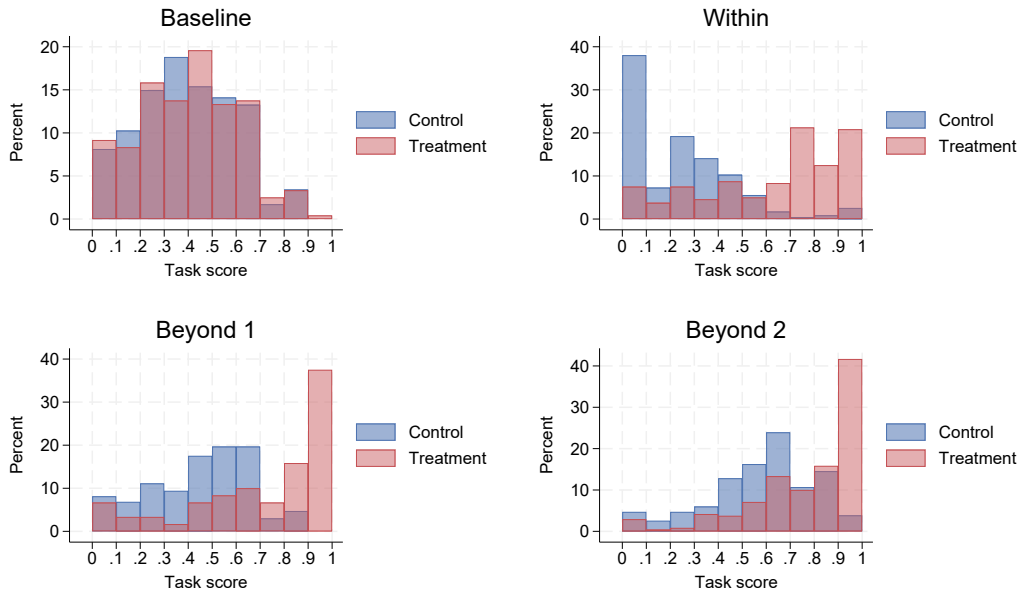
Notes: Outcome is the share of points achieved on each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Treatment coefficients are from OLS regressions of the outcome on the treatment indicator; heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure A.1: Histogram of Share of Points Achieved, by Task

(a) Application Development



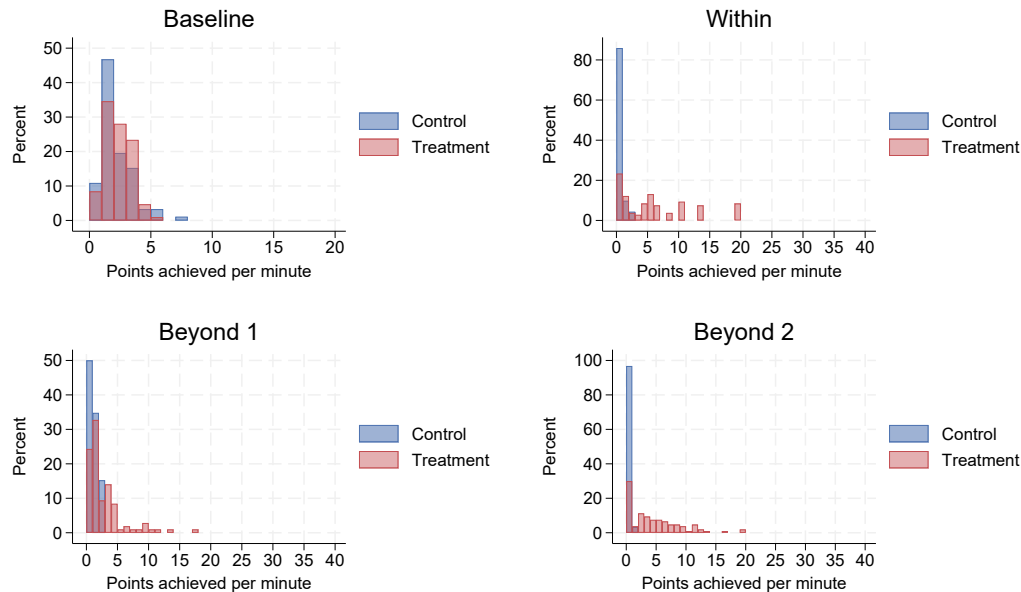
(b) System Integration



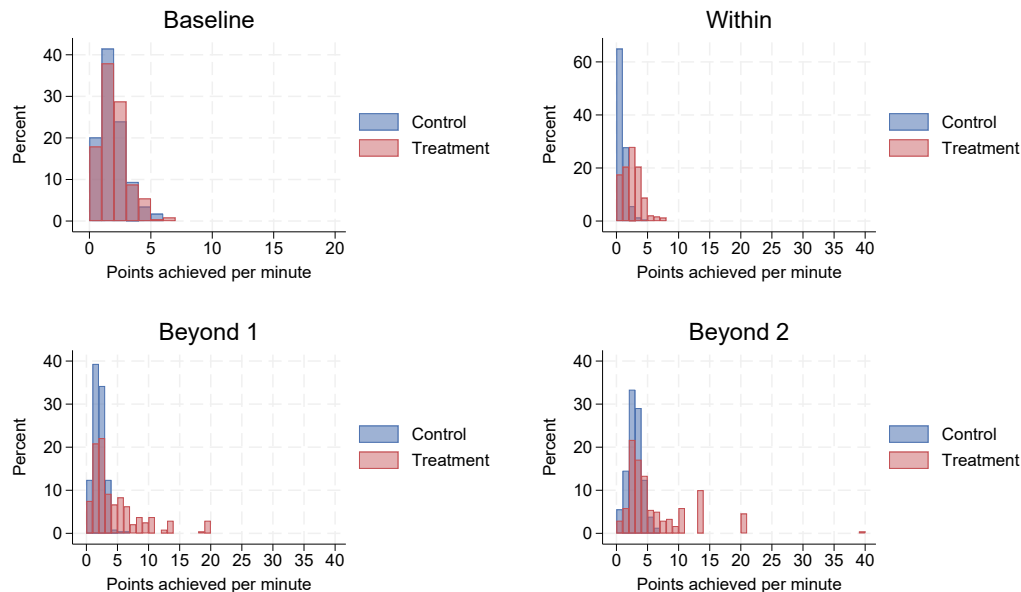
Notes: Histograms show the distribution of the share of maximum points achieved, separately by specialization.

Figure A.2: Histogram of Points Achieved per Minute, by Task

(a) Application Development



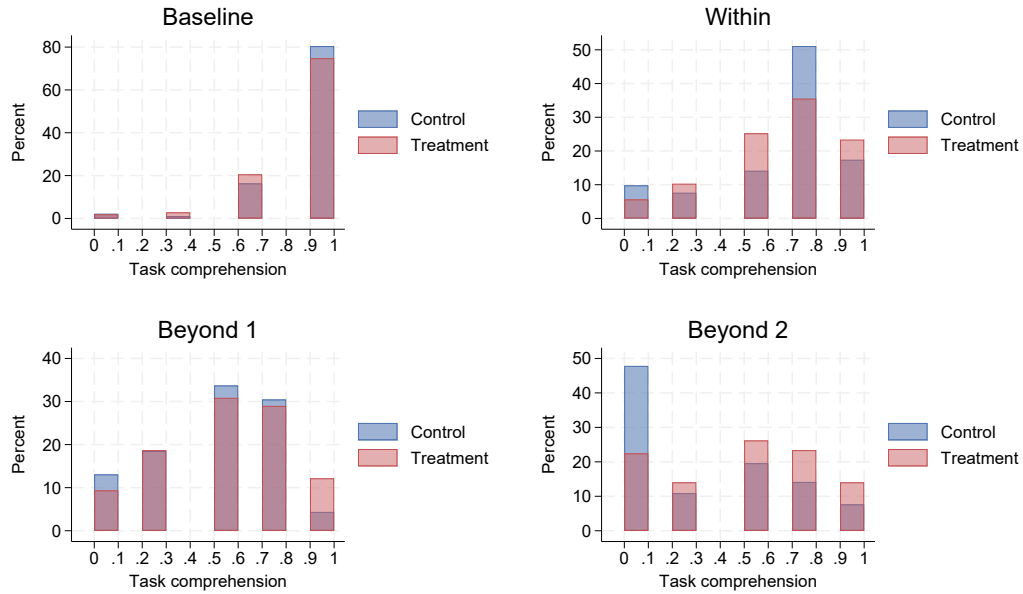
(b) System Integration



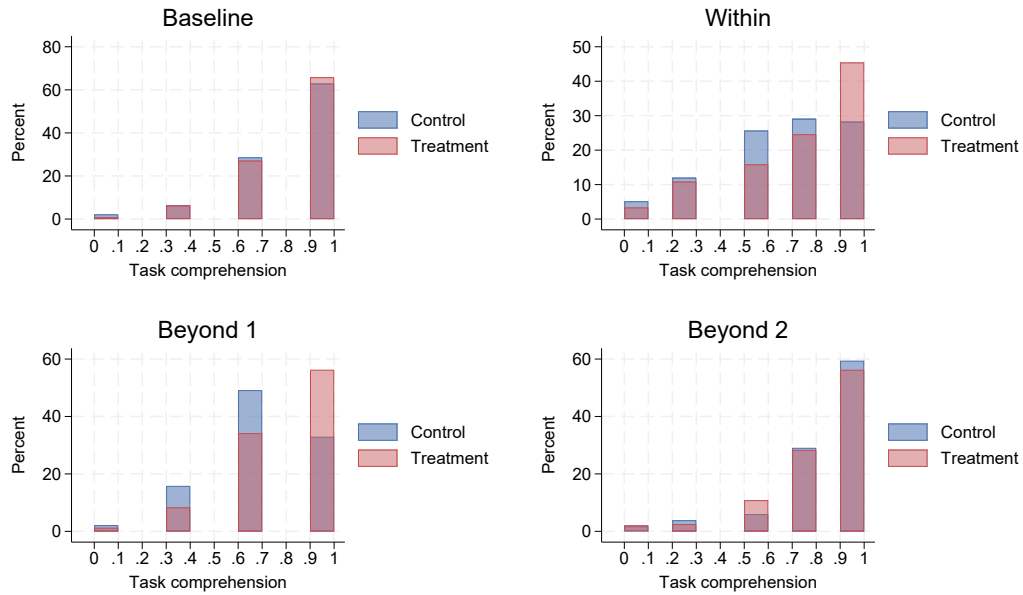
Notes: Histograms show the distribution of points achieved per minute, separately by specialization.

Figure A.3: Histogram of Comprehension, by Task

(a) Application Development



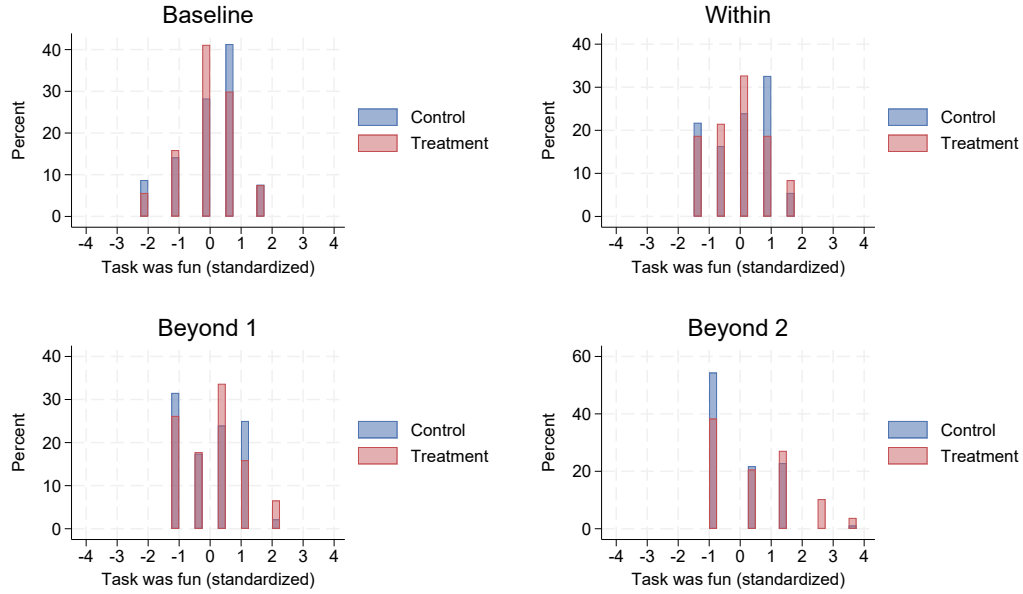
(b) System Integration



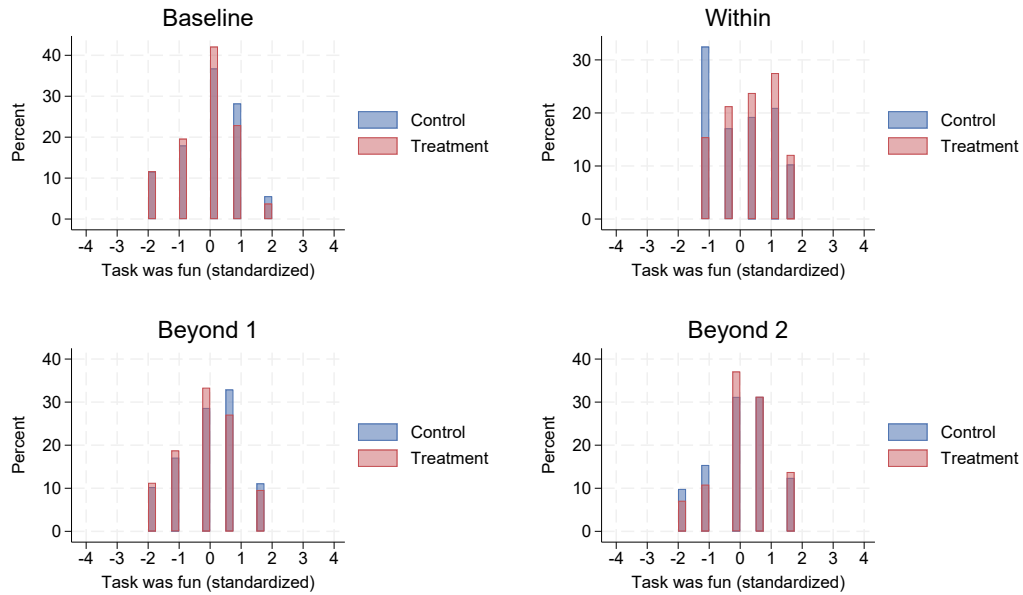
Notes: Histograms show the distribution of comprehension, separately by specialization.

Figure A.4: Histogram of Task Experience (Fun), by Task

(a) Application Development



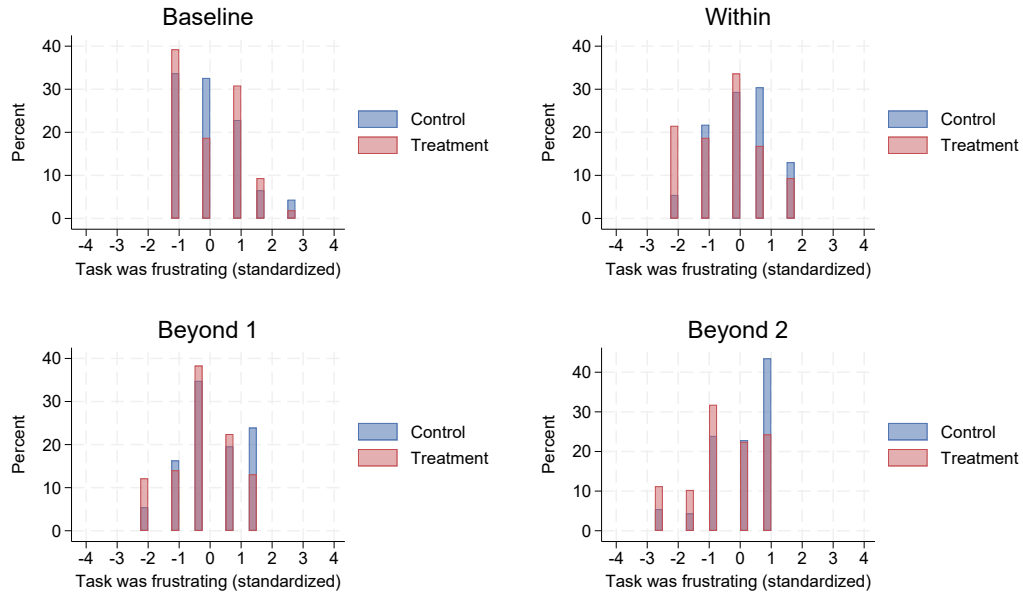
(b) System Integration



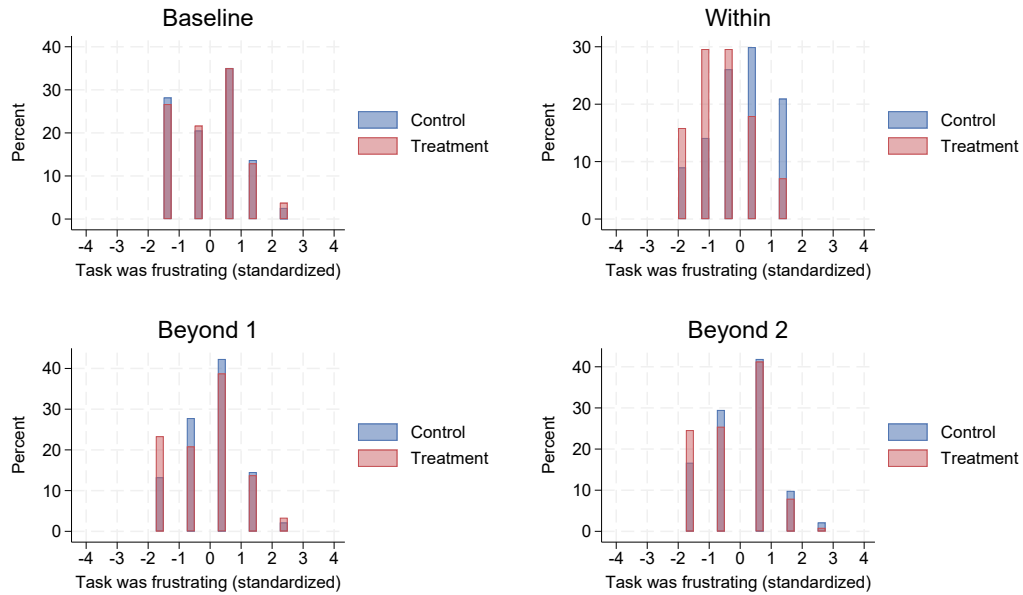
Notes: Histograms show the distribution of task experience (fun, standardized at the control group mean), separately by specialization.

Figure A.5: Histogram of Task Experience (Frustrating), by Task

(a) Application Development



(b) System Integration



Notes: Histograms show the distribution of task experience (frustrating, standardized at the control group mean), separately by specialization.

Table A.5: Average Treatment Effects on Extensive Margin: Gives an Answer, by Task

	Answer given (=1)							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Full sample								
Treatment	-0.019 (0.014)	-0.020 (0.014)	0.107*** (0.022)	0.113*** (0.022)	0.009 (0.018)	0.013 (0.018)	0.058*** (0.020)	0.060*** (0.020)
Controls		✓		✓		✓		✓
Control mean	0.975	0.975	0.853	0.853	0.939	0.939	0.896	0.896
Observations	673	673	673	673	673	673	673	673
Application Development								
Treatment	-0.017 (0.019)	-0.011 (0.019)	0.107** (0.045)	0.109** (0.053)	0.046 (0.044)	0.058 (0.046)	0.169*** (0.055)	0.163*** (0.059)
Controls		✓		✓		✓		✓
Control mean	0.989	0.989	0.837	0.837	0.870	0.870	0.728	0.728
Observations	199	199	199	199	199	199	199	199
System Integration								
Treatment	-0.020 (0.018)	-0.020 (0.018)	0.108*** (0.026)	0.110*** (0.025)	-0.003 (0.017)	-0.002 (0.017)	0.018 (0.016)	0.019 (0.015)
Controls		✓		✓		✓		✓
Control mean	0.970	0.970	0.859	0.859	0.966	0.966	0.962	0.962
Observations	474	474	474	474	474	474	474	474

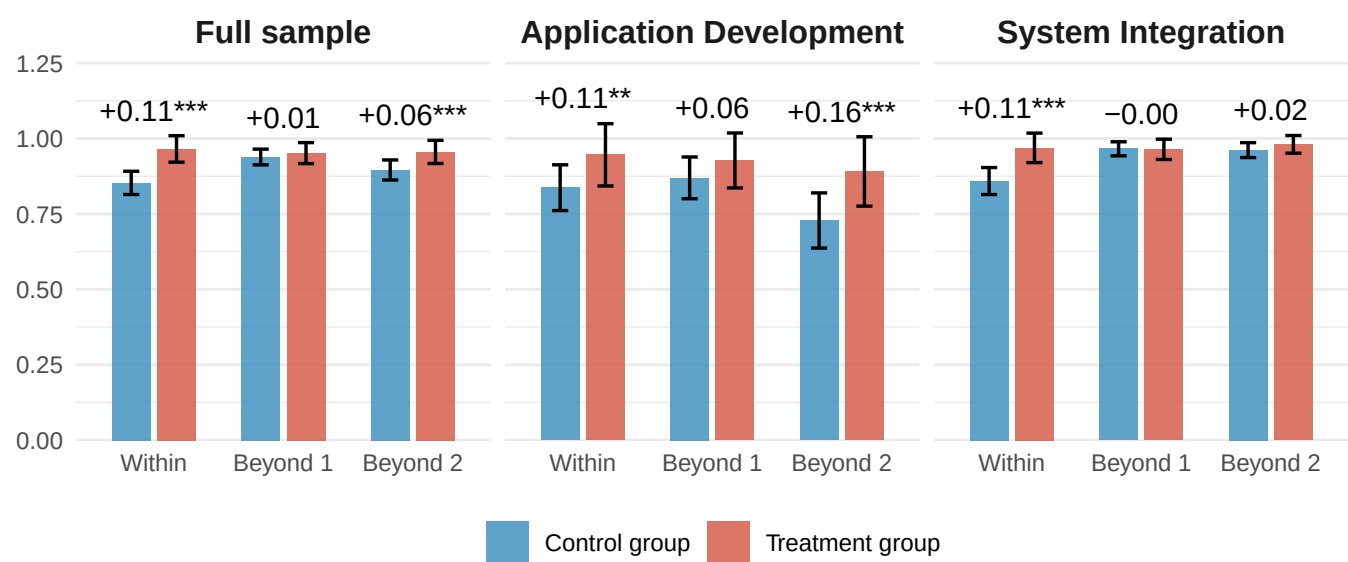
Notes: Outcome is an indicator equal to one if the participant gave an answer on the respective task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Treatment coefficients are obtained from OLS regressions of the outcome on the treatment indicator; heteroscedasticity-robust standard errors are reported in parentheses. Control means refer to the average outcome in the control group. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.6: Average Treatment Effects on Intensive Margin: Share of Points Achieved (if Answer Given), by Task

	Share of points achieved							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Full sample								
Treatment	0.025 (0.016)	0.019 (0.016)	0.432*** (0.021)	0.432*** (0.021)	0.268*** (0.020)	0.274*** (0.019)	0.284*** (0.023)	0.302*** (0.019)
Controls		✓		✓		✓		✓
Control mean	0.414	0.414	0.257	0.257	0.420	0.420	0.472	0.472
Observations	650	650	611	611	635	635	623	623
Application Development								
Treatment	0.034 (0.030)	0.030 (0.031)	0.532*** (0.040)	0.504*** (0.045)	0.250*** (0.037)	0.251*** (0.039)	0.606*** (0.036)	0.602*** (0.040)
Controls		✓		✓		✓		✓
Control mean	0.478	0.478	0.244	0.244	0.316	0.316	0.049	0.049
Observations	195	195	178	178	178	178	163	163
System Integration								
Treatment	0.017 (0.019)	0.018 (0.019)	0.389*** (0.024)	0.390*** (0.025)	0.283*** (0.023)	0.277*** (0.023)	0.199*** (0.020)	0.196*** (0.020)
Controls		✓		✓		✓		✓
Control mean	0.388	0.388	0.262	0.262	0.456	0.456	0.598	0.598
Observations	455	455	433	433	457	457	460	460

Notes: Outcome is the share of points achieved on each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Treatment coefficients are from OLS regressions of the outcome on the treatment indicator; heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure A.6: Average Treatment Effects on the Extensive Margin of Productivity (Answer Given), by Task



Notes: Outcome is an indicator equal to one if the participant gave an answer on the respective task. Tasks are ordered by required skill level: Within (within curriculum), Beyond 1 and Beyond 2 (beyond curriculum). Bars show control and treatment group means; error bars are 95% confidence intervals. Treatment effects are from OLS regressions with controls; heteroscedasticity-robust standard errors. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.7: Average Treatment Effects on Productivity per Minute, by Task

	Points per minute							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Full sample								
Treatment	0.089 (0.089)	0.043 (0.086)	2.824*** (0.207)	2.742*** (0.195)	2.308*** (0.221)	2.311*** (0.212)	3.439*** (0.283)	3.467*** (0.271)
Controls		✓		✓		✓		✓
Control mean	1.848	1.848	0.704	0.704	1.671	1.671	2.127	2.127
Observations	673	673	673	673	673	673	673	673
Application Development								
Treatment	0.096 (0.174)	0.135 (0.176)	5.395*** (0.561)	5.206*** (0.585)	1.742*** (0.303)	1.709*** (0.317)	4.284*** (0.421)	4.248*** (0.466)
Controls		✓		✓		✓		✓
Control mean	2.096	2.096	0.547	0.547	0.955	0.955	0.199	0.199
Observations	199	199	199	199	199	199	199	199
System Integration								
Treatment	0.073 (0.102)	0.043 (0.101)	1.686*** (0.109)	1.645*** (0.106)	2.599*** (0.279)	2.519*** (0.265)	3.164*** (0.346)	3.134*** (0.343)
Controls		✓		✓		✓		✓
Control mean	1.750	1.750	0.765	0.765	1.952	1.952	2.885	2.885
Observations	474	474	474	474	474	474	474	474

Notes: Outcome is productivity per minute, measured as points achieved per minute spent on each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1 /Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Treatment coefficients are obtained from OLS regressions of the outcome on the treatment indicator; heteroscedasticity-robust standard errors are reported in parentheses. Control means refer to the average outcome in the control group. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.8: Average Treatment Effects on Productivity per Minute (Winsorized), by Task

	Points per minute							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Full sample								
Treatment	0.090 (0.087)	0.042 (0.085)	2.824*** (0.207)	2.742*** (0.195)	2.301*** (0.219)	2.305*** (0.211)	3.381*** (0.268)	3.425*** (0.263)
Controls		✓		✓		✓		✓
Control mean	1.839	1.839	0.704	0.704	1.671	1.671	2.127	2.127
Observations	673	673	673	673	673	673	673	673
Application Development								
Treatment	0.118 (0.164)	0.135 (0.174)	5.395*** (0.561)	5.206*** (0.585)	1.742*** (0.303)	1.709*** (0.317)	4.284*** (0.421)	4.248*** (0.466)
Controls		✓		✓		✓		✓
Control mean	2.067	2.067	0.547	0.547	0.955	0.955	0.199	0.199
Observations	199	199	199	199	199	199	199	199
System Integration								
Treatment	0.066 (0.101)	0.037 (0.100)	1.686*** (0.109)	1.645*** (0.106)	2.588*** (0.277)	2.511*** (0.263)	3.080*** (0.321)	3.064*** (0.326)
Controls		✓		✓		✓		✓
Control mean	1.750	1.750	0.765	0.765	1.952	1.952	2.885	2.885
Observations	474	474	474	474	474	474	474	474

Notes: Outcome is productivity per minute, measured as points achieved per minute spent on each task, winsorized at the 1st and 99th percentiles. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Treatment coefficients are obtained from OLS regressions of the outcome on the treatment indicator; heteroscedasticity-robust standard errors are reported in parentheses. Control means refer to the average outcome in the control group. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.9: Average Treatment Effects on Comprehension (Share of Comprehension Questions Answered Correctly), by Task

	Comprehension: Share correct							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Full sample								
Treatment	0.008 (0.017)	0.007 (0.017)	0.061*** (0.022)	0.067*** (0.022)	0.086*** (0.021)	0.097*** (0.020)	0.028 (0.026)	0.044** (0.021)
Controls		✓		✓		✓		✓
Control mean	0.862	0.862	0.655	0.655	0.646	0.646	0.698	0.698
Observations	673	673	673	673	673	673	673	673
Application Development								
Treatment	-0.023 (0.029)	-0.014 (0.029)	0.005 (0.041)	-0.006 (0.040)	0.053 (0.040)	0.069 (0.043)	0.174*** (0.049)	0.179*** (0.055)
Controls		✓		✓		✓		✓
Control mean	0.917	0.917	0.647	0.647	0.486	0.486	0.307	0.307
Observations	199	199	199	199	199	199	199	199
System Integration								
Treatment	0.019 (0.021)	0.023 (0.021)	0.087*** (0.027)	0.091*** (0.026)	0.109*** (0.022)	0.110*** (0.023)	-0.016 (0.021)	-0.009 (0.021)
Controls		✓		✓		✓		✓
Control mean	0.840	0.840	0.658	0.658	0.709	0.709	0.851	0.851
Observations	474	474	474	474	474	474	474	474

Notes: Outcome is the share of comprehension questions answered correctly for each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Treatment coefficients are from OLS regressions of the outcome on the treatment indicator; heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.10: Average Treatment Effects on Comprehension, Excluding Treatment-Group Apprentices Who Accessed AI During the Comprehension Blocks

	Comprehension: Share correct							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Full sample								
Treatment	0.015 (0.017)	0.014 (0.017)	0.069*** (0.023)	0.070*** (0.022)	0.080*** (0.022)	0.087*** (0.020)	0.027 (0.027)	0.033 (0.021)
Controls		✓		✓		✓		✓
Control mean	0.862	0.862	0.655	0.655	0.646	0.646	0.698	0.698
Observations	629	629	629	629	629	629	629	629
Application Development								
Treatment	-0.011 (0.029)	-0.008 (0.027)	0.012 (0.042)	0.011 (0.041)	0.036 (0.040)	0.053 (0.043)	0.130** (0.050)	0.133** (0.057)
Controls		✓		✓		✓		✓
Control mean	0.917	0.917	0.647	0.647	0.486	0.486	0.307	0.307
Observations	180	180	180	180	180	180	180	180
System Integration								
Treatment	0.025 (0.021)	0.027 (0.022)	0.092*** (0.027)	0.089*** (0.027)	0.100*** (0.023)	0.101*** (0.023)	-0.010 (0.021)	-0.001 (0.021)
Controls		✓		✓		✓		✓
Control mean	0.840	0.840	0.658	0.658	0.709	0.709	0.851	0.851
Observations	449	449	449	449	449	449	449	449

Notes: Robustness check for the comprehension results in Table A.9. The sample excludes treatment-group apprentices who either self-reported having used AI during a comprehension block or were observed submitting a prompt to the AI tracker during a comprehension block. Outcome, specifications, and column structure are identical to Table A.9. Treatment coefficients are from OLS regressions of the outcome on the treatment indicator; heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.11: Average Treatment Effects on Comprehension, Excluding Apprentices Who Have Not Given an Answer in the Specific Task, by Task

	Comprehension: Share correct							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Full sample								
Treatment	0.017 (0.017)	0.015 (0.017)	0.041* (0.023)	0.055** (0.022)	0.090*** (0.021)	0.102*** (0.020)	0.001 (0.025)	0.031 (0.021)
Controls		✓		✓		✓		✓
Control mean	0.861	0.861	0.686	0.686	0.659	0.659	0.745	0.745
Observations	650	650	611	611	635	635	623	623
Application Development								
Treatment	-0.012 (0.028)	-0.006 (0.028)	-0.001 (0.041)	-0.012 (0.042)	0.054 (0.040)	0.074* (0.041)	0.165*** (0.055)	0.173*** (0.063)
Controls		✓		✓		✓		✓
Control mean	0.916	0.916	0.669	0.669	0.509	0.509	0.343	0.343
Observations	195	195	178	178	178	178	163	163
System Integration								
Treatment	0.027 (0.021)	0.028 (0.021)	0.060** (0.027)	0.072*** (0.026)	0.116*** (0.022)	0.116*** (0.023)	-0.022 (0.020)	-0.014 (0.020)
Controls		✓		✓		✓		✓
Control mean	0.838	0.838	0.693	0.693	0.712	0.712	0.864	0.864
Observations	455	455	433	433	457	457	460	460

Notes: Outcome is the share of comprehension questions answered correctly for each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Treatment coefficients are from OLS regressions of the outcome on the treatment indicator; heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.12: Average Treatment Effects on Task Enjoyment (“Working on the task was fun”), by Task

	Working on the task was fun							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Full sample								
Treatment	-0.083 (0.075)	-0.079 (0.075)	0.194*** (0.074)	0.209*** (0.075)	-0.061 (0.076)	-0.044 (0.073)	0.157** (0.076)	0.172*** (0.066)
Controls		✓		✓		✓		✓
Control mean	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003
Observations	673	673	673	673	673	673	673	673
Application Development								
Treatment	-0.067 (0.136)	-0.103 (0.138)	-0.056 (0.140)	-0.087 (0.150)	0.081 (0.141)	0.099 (0.155)	0.546*** (0.164)	0.470*** (0.172)
Controls		✓		✓		✓		✓
Control mean	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000	0.000
Observations	199	199	199	199	199	199	199	199
System Integration								
Treatment	-0.101 (0.090)	-0.069 (0.091)	0.289*** (0.088)	0.312*** (0.088)	-0.109 (0.091)	-0.121 (0.088)	0.112 (0.089)	0.073 (0.087)
Controls		✓		✓		✓		✓
Control mean	-0.004	-0.004	-0.004	-0.004	-0.003	-0.003	-0.003	-0.003
Observations	474	474	474	474	474	474	474	474

Notes: Outcome is apprentices’ agreement with the statement “Working on the task was fun” (Likert scale), standardized by the control-group mean and standard deviation within each sample. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Treatment coefficients are from OLS regressions of the outcome on the treatment indicator; heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.13: Average Treatment Effects on Frustration (“The task was frustrating”), by Task

	The task was frustrating							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Full sample								
Treatment	0.018 (0.078)	-0.013 (0.077)	-0.536*** (0.077)	-0.550*** (0.076)	-0.141* (0.079)	-0.151** (0.075)	-0.206*** (0.076)	-0.218*** (0.066)
Controls		✓		✓		✓		✓
Control mean	0.004	0.004	0.001	0.001	0.005	0.005	0.005	0.005
Observations	673	673	673	673	673	673	673	673
Application Development								
Treatment	0.006 (0.143)	-0.001 (0.145)	-0.454*** (0.151)	-0.505*** (0.160)	-0.254* (0.142)	-0.290* (0.147)	-0.484*** (0.149)	-0.491*** (0.150)
Controls		✓		✓		✓		✓
Control mean	0.000	0.000	0.000	0.000	-0.000	-0.000	-0.000	-0.000
Observations	199	199	199	199	199	199	199	199
System Integration								
Treatment	0.032 (0.092)	-0.017 (0.094)	-0.567*** (0.089)	-0.568*** (0.090)	-0.121 (0.098)	-0.098 (0.098)	-0.170* (0.092)	-0.160* (0.094)
Controls		✓		✓		✓		✓
Control mean	0.005	0.005	0.001	0.001	0.007	0.007	0.007	0.007
Observations	474	474	474	474	474	474	474	474

Notes: Outcome is apprentices’ agreement with the statement “The task was frustrating” (Likert scale), standardized by the control-group mean and standard deviation within each sample. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Treatment coefficients are from OLS regressions of the outcome on the treatment indicator; heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.14: Heterogeneous Treatment Effects on Productivity (Share of Points Achieved), by AI Literacy

	Share of points achieved							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
I understand how ChatGPT generates answers.								
Treatment (Below Median)	0.012 (0.017)	0.009 (0.017)	0.439*** (0.022)	0.442*** (0.022)	0.263*** (0.022)	0.269*** (0.020)	0.300*** (0.024)	0.307*** (0.020)
Treatment (Above Median)	0.023 (0.024)	0.019 (0.025)	0.474*** (0.030)	0.473*** (0.029)	0.278*** (0.031)	0.284*** (0.029)	0.313*** (0.035)	0.310*** (0.029)
Difference	0.010 (0.018)	0.009 (0.018)	0.035 (0.022)	0.030 (0.022)	0.014 (0.023)	0.015 (0.021)	0.013 (0.023)	0.003 (0.021)
I can identify errors in ChatGPT's responses.								
Treatment (Below Median)	0.013 (0.017)	0.010 (0.017)	0.440*** (0.022)	0.443*** (0.022)	0.265*** (0.022)	0.273*** (0.020)	0.301*** (0.024)	0.309*** (0.020)
Treatment (Above Median)	0.005 (0.024)	0.003 (0.024)	0.459*** (0.031)	0.455*** (0.031)	0.205*** (0.030)	0.200*** (0.029)	0.316*** (0.036)	0.305*** (0.030)
Difference	-0.008 (0.018)	-0.006 (0.018)	0.020 (0.024)	0.012 (0.025)	-0.060*** (0.021)	-0.073*** (0.020)	0.014 (0.025)	-0.004 (0.021)
I can formulate appropriate and effective questions for ChatGPT.								
Treatment (Below Median)	0.013 (0.017)	0.010 (0.017)	0.439*** (0.022)	0.443*** (0.021)	0.264*** (0.022)	0.271*** (0.020)	0.301*** (0.024)	0.308*** (0.020)
Treatment (Above Median)	-0.005 (0.024)	-0.001 (0.024)	0.500*** (0.030)	0.498*** (0.030)	0.291*** (0.030)	0.279*** (0.029)	0.350*** (0.035)	0.335*** (0.030)
Difference	-0.018 (0.017)	-0.011 (0.017)	0.061*** (0.023)	0.055** (0.023)	0.027 (0.022)	0.008 (0.020)	0.050* (0.025)	0.026 (0.022)
I can use ChatGPT to support my decision-making process.								
Treatment (Below Median)	0.013 (0.017)	0.010 (0.017)	0.436*** (0.022)	0.439*** (0.022)	0.262*** (0.021)	0.267*** (0.020)	0.298*** (0.024)	0.304*** (0.020)
Treatment (Above Median)	0.013 (0.024)	0.014 (0.023)	0.508*** (0.029)	0.509*** (0.030)	0.314*** (0.030)	0.307*** (0.029)	0.368*** (0.035)	0.361*** (0.029)
Difference	0.000 (0.017)	0.004 (0.017)	0.072*** (0.023)	0.070*** (0.023)	0.051** (0.022)	0.039* (0.022)	0.071*** (0.025)	0.057*** (0.022)
I can use ChatGPT to work more efficiently and productively.								
Treatment (Below Median)	0.013 (0.017)	0.010 (0.017)	0.440*** (0.022)	0.443*** (0.022)	0.265*** (0.021)	0.271*** (0.020)	0.301*** (0.024)	0.309*** (0.020)
Treatment (Above Median)	0.024 (0.024)	0.023 (0.024)	0.506*** (0.030)	0.510*** (0.030)	0.338*** (0.030)	0.332*** (0.029)	0.360*** (0.034)	0.342*** (0.029)
Difference	0.011 (0.017)	0.013 (0.017)	0.066*** (0.024)	0.067*** (0.024)	0.072*** (0.023)	0.060*** (0.022)	0.059** (0.025)	0.034 (0.021)
I can use ChatGPT to achieve my learning and educational goals.								
Treatment (Below Median)	0.012 (0.017)	0.010 (0.017)	0.440*** (0.022)	0.442*** (0.022)	0.266*** (0.022)	0.271*** (0.020)	0.302*** (0.024)	0.309*** (0.020)
Treatment (Above Median)	0.006 (0.024)	0.005 (0.024)	0.504*** (0.030)	0.505*** (0.029)	0.314*** (0.030)	0.311*** (0.029)	0.344*** (0.034)	0.340*** (0.028)
Difference	-0.007 (0.017)	-0.005 (0.018)	0.064*** (0.023)	0.062*** (0.023)	0.048** (0.023)	0.039* (0.021)	0.042* (0.025)	0.032 (0.020)

Notes: Outcome is the share of points achieved on each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). Each row reports heterogeneity by a different AI literacy item (self-assessed on a 5-point scale, median-split into below and above median). Treatment effects are from OLS regressions of the outcome on the treatment indicator, the binary moderator, and their interaction; the interaction term (“Difference”) identifies the differential treatment effect for apprentices above the median. Heteroscedasticity-robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.15: Heterogeneous Treatment Effects on Points per Minute, by AI Literacy

	Points per minute							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment (Low AI literacy)	-0.100 (0.125)	-0.149 (0.121)	2.268*** (0.269)	2.120*** (0.265)	1.889*** (0.277)	1.909*** (0.267)	2.797*** (0.378)	2.913*** (0.382)
Treatment (High AI literacy)	0.306** (0.130)	0.263** (0.128)	3.436*** (0.328)	3.454*** (0.318)	2.937*** (0.355)	2.891*** (0.344)	4.299*** (0.437)	4.206*** (0.408)
Difference	0.406** (0.181)	0.412** (0.176)	1.168*** (0.424)	1.335*** (0.435)	1.048** (0.451)	0.982** (0.437)	1.502*** (0.578)	1.293** (0.563)
Controls		✓		✓		✓		✓
Control mean (Low AI literacy)	1.921	1.921	0.710	0.710	1.615	1.615	2.213	2.213
Control mean (High AI literacy)	1.771	1.771	0.704	0.704	1.743	1.743	2.077	2.077
Observations	653	653	653	653	653	653	653	653

Notes: Outcome is points achieved per minute on each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Treatment effects are from OLS regressions of the outcome on the treatment indicator, an indicator for above-median AI literacy, and their interaction; the interaction term identifies the differential treatment effect for apprentices with above-median AI literacy. Heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome within the corresponding AI-literacy subgroup. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.16: Heterogeneous Treatment Effects on Comprehension, by AI Literacy

	Comprehension: Share correct							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment (Low AI literacy)	-0.005 (0.023)	-0.008 (0.023)	0.051 (0.033)	0.074** (0.031)	0.057* (0.030)	0.072** (0.028)	-0.003 (0.037)	0.044 (0.030)
Treatment (High AI literacy)	0.015 (0.026)	0.016 (0.027)	0.069** (0.032)	0.064** (0.032)	0.111*** (0.032)	0.111*** (0.029)	0.064* (0.037)	0.045 (0.029)
Difference	0.020 (0.035)	0.024 (0.036)	0.018 (0.046)	-0.010 (0.044)	0.054 (0.044)	0.039 (0.040)	0.067 (0.053)	0.001 (0.041)
Controls		✓		✓		✓		✓
Control mean (Low AI literacy)	0.873	0.873	0.655	0.655	0.671	0.671	0.718	0.718
Control mean (High AI literacy)	0.850	0.850	0.652	0.652	0.626	0.626	0.684	0.684
Observations	653	653	653	653	653	653	653	653

Notes: Outcome is the share of comprehension questions answered correctly for each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Treatment effects are from OLS regressions of the outcome on the treatment indicator, an indicator for above-median AI literacy, and their interaction; the interaction term identifies the differential treatment effect for apprentices with above-median AI literacy. Heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome within the corresponding AI-literacy subgroup. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.17: Heterogeneous Treatment Effects on Productivity (Share of Points Achieved), by Education

	Share of points achieved							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment (No university entrance qualification)	0.011 (0.027)	0.003 (0.027)	0.457*** (0.037)	0.463*** (0.038)	0.320*** (0.037)	0.322*** (0.036)	0.340*** (0.040)	0.343*** (0.036)
Treatment (University entrance qualification)	-0.002 (0.022)	-0.005 (0.021)	0.434*** (0.028)	0.429*** (0.028)	0.235*** (0.028)	0.243*** (0.027)	0.280*** (0.032)	0.288*** (0.026)
Difference	-0.013 (0.035)	-0.008 (0.034)	-0.023 (0.047)	-0.034 (0.048)	-0.085* (0.047)	-0.079* (0.045)	-0.060 (0.051)	-0.055 (0.045)
Controls		✓		✓		✓		✓
Control mean (No university entrance qualification)	0.341	0.341	0.188	0.188	0.363	0.363	0.406	0.406
Control mean (University entrance qualification)	0.451	0.451	0.233	0.233	0.414	0.414	0.448	0.448
Observations	616	616	616	616	616	616	616	616

Notes: Outcome is the share of points achieved on each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Education control variables are not included this time. Treatment effects are from OLS regressions of the outcome on the treatment indicator, an indicator for holding a university-of-applied-sciences entrance qualification or a higher degree, and their interaction; the interaction term identifies the differential treatment effect for apprentices with higher pre-apprenticeship education. Heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome within the corresponding educational subgroup. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.18: Robustness: Heterogeneous Treatment Effects on Productivity (Share of Points Achieved), by Education (Dummy = 1 if General University Entrance Qualification or Higher)

	Share of points achieved							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment (No university entrance qualification)	0.041 (0.025)	0.032 (0.025)	0.453*** (0.033)	0.459*** (0.033)	0.292*** (0.033)	0.297*** (0.032)	0.325*** (0.036)	0.332*** (0.032)
Treatment (University entrance qualification)	-0.002 (0.022)	-0.004 (0.021)	0.434*** (0.028)	0.431*** (0.028)	0.235*** (0.028)	0.242*** (0.027)	0.280*** (0.032)	0.288*** (0.026)
Difference	-0.043 (0.033)	-0.037 (0.033)	-0.019 (0.043)	-0.028 (0.044)	-0.057 (0.044)	-0.055 (0.042)	-0.045 (0.048)	-0.044 (0.042)
Controls		✓		✓		✓		✓
Control mean (No university entrance qualification)	0.337	0.337	0.198	0.198	0.365	0.365	0.387	0.387
Control mean (University entrance qualification)	0.451	0.451	0.233	0.233	0.414	0.414	0.448	0.448
Observations	673	673	673	673	673	673	673	673

Notes: Outcome is the share of points achieved on each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Education control variables are not included this time. Treatment effects are from OLS regressions of the outcome on the treatment indicator, an indicator for holding a general university entrance qualification or a higher degree, and their interaction; the interaction term identifies the differential treatment effect for apprentices with higher pre-apprenticeship education. Heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome within the corresponding educational subgroup. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.19: Heterogeneous Treatment Effects on Points per Minute, by Education.

	Points per minute							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment (No university entrance qualification)	0.116 (0.153)	0.015 (0.154)	2.933*** (0.373)	2.815*** (0.366)	2.750*** (0.371)	2.656*** (0.369)	4.161*** (0.510)	4.127*** (0.525)
Treatment (University entrance qualification)	0.037 (0.116)	0.005 (0.116)	2.839*** (0.275)	2.734*** (0.261)	2.325*** (0.306)	2.377*** (0.305)	3.295*** (0.376)	3.367*** (0.376)
Difference	-0.078 (0.192)	-0.010 (0.195)	-0.094 (0.463)	-0.081 (0.470)	-0.426 (0.481)	-0.279 (0.490)	-0.866 (0.634)	-0.761 (0.673)
Controls		✓		✓		✓		✓
Control mean (No university entrance qualification)	1.626	1.626	0.651	0.651	1.666	1.666	2.249	2.249
Control mean (University entrance qualification)	2.021	2.021	0.731	0.731	1.703	1.703	2.146	2.146
Observations	616	616	616	616	616	616	616	616

Notes: Outcome is the points achieved per minute on each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Education control variables are not included this time. Treatment effects are from OLS regressions of the outcome on the treatment indicator, an indicator for holding a university-of-applied-sciences entrance qualification or a higher degree, and their interaction; the interaction term identifies the differential treatment effect for apprentices with higher pre-apprenticeship education. Heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome within the corresponding educational subgroup. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.20: Heterogeneous Treatment Effects on Comprehension, by Education

	Comprehension: Share correct							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment (No university entrance qualification)	0.034 (0.033)	0.039 (0.033)	0.069 (0.042)	0.111*** (0.042)	0.093** (0.037)	0.103*** (0.035)	0.005 (0.043)	0.014 (0.037)
Treatment (University entrance qualification)	-0.016 (0.021)	-0.019 (0.021)	0.063** (0.028)	0.060** (0.028)	0.074*** (0.028)	0.084*** (0.027)	0.036 (0.035)	0.055** (0.027)
Difference	-0.049 (0.039)	-0.058 (0.040)	-0.006 (0.051)	-0.051 (0.051)	-0.019 (0.046)	-0.019 (0.044)	0.032 (0.055)	0.041 (0.046)
Controls		✓		✓		✓		✓
Control mean (No university entrance qualification)	0.804	0.804	0.601	0.601	0.649	0.649	0.716	0.716
Control mean (University entrance qualification)	0.898	0.898	0.668	0.668	0.654	0.654	0.697	0.697
Observations	616	616	616	616	616	616	616	616

Notes: Outcome is the share of comprehension questions answered correctly for each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Education control variables are not included this time. Treatment effects are from OLS regressions of the outcome on the treatment indicator, an indicator for holding a university-of-applied-sciences entrance qualification or a higher degree, and their interaction; the interaction term identifies the differential treatment effect for apprentices with higher pre-apprenticeship education. Heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome within the corresponding educational subgroup. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Appendix B. AI Usage and Treatment Take-up

Appendix B.1. Comparison of AI Users and Never-takers

Self-reported AI usage in our sample is typically frequent, with about 91% of school users reporting use at least weekly. Across contexts, AI is used primarily for research and comprehension followed by writing and coding tasks. ChatGPT is by far the dominant tool, with two-thirds using the free version and 32% using ChatGPT Plus.

Table B.2 compares never-takers with AI users, i.e. treatment-group apprentices who actively used the tool. We find no significant differences in age, gender, or educational background between the two groups. However, AI users report significantly higher prior AI usage and AI literacy than never-takers, suggesting that treatment take-up is related to pre-existing familiarity and engagement with AI tools.

Table B.3 compares never-takers and AI users along our main outcome measures. Across all tasks, AI users exhibit substantially higher productivity levels than never-takers, both in terms of points achieved and points achieved per minute. Differences in comprehension are more nuanced. For the Within tasks, AI users are slightly less likely to answer comprehension questions correctly, whereas for the Beyond tasks, they perform slightly better than never-takers. Thus, never-takers perform relatively well on comprehension of material within the apprenticeship curriculum, but not on tasks beyond it. Given the small number of never-takers, these differences should be interpreted cautiously.

Table B.4 reports differences in task perception. Here, in contrast, we find little evidence of differences between the two groups in overall task enjoyment or frustration. This suggests that differences in productivity and comprehension between never-takers and AI users are not accompanied by differences in task experience. Notably, never-takers report exerting higher effort, yet this does not appear to translate into a worse task experience. Given the small number of never-takers, however, these comparisons should be interpreted with caution.

Appendix B.2. Prompting Behavior

Table B.6 describes prompting behavior among apprentices who used the provided AI tool at least once during the experiment.

On average, users submitted 6.3 prompts across the three tasks, with prompting activity relatively evenly distributed across task blocks. System Integration apprentices submitted more prompts overall than Application Development apprentices, primarily driven by higher prompting activity in the Within and Beyond 1 tasks. Differences in the number of prompts across tasks may partly reflect differences in the task structure, i.e., the number of subtasks. In the Application Development specialization, the Within task contained one subtask, while the Beyond 1 and Beyond 2 tasks each contained two subtasks. In the System Integration specialization, the Within task consisted of two

subtasks with tables as response input, while the Beyond 1 and Beyond 2 tasks each contained three subtasks.

The share of AI users submitting no prompt for the Within task is 0.17 in the full sample, 0.33 in the Application Development specialization and 0.11 in the System Integration specialization. For task Beyond 1, it is 0.20 in the full sample, 0.24 in the Application Development specialization and 0.19 in the System Integration specialization. For task Beyond 2, it is 0.23 in the full sample, 0.16 in the Application Development specialization and 0.25 in the System Integration specialization. This shows that a non-negligible share of AI users did not submit any prompt in a given task block, despite using the tool elsewhere in the experiment. Within each specialization, the share is smallest for the task where performance without AI is lowest.

The second panel shows substantial heterogeneity in the types of prompts submitted. Prompts were individually reviewed and categorized manually, following the criteria described in the table notes for Table B.6. Around half of all prompts are classified as copy prompts, indicating that apprentices frequently pasted parts of the task description or materials directly into AI. Instructed copy prompts account for roughly one quarter of prompts, while explain and transform prompts account for about 10% of prompts. Overall, the distribution of prompt types is similar across Application Development and System Integration, although Application Development apprentices rely somewhat more on explain prompts.

The third panel reports the share of apprentices who used an exact copy of ChatGPT-generated content as their submitted answers, conditional on querying ChatGPT in the respective task block, across task blocks (Within, Beyond 1, Beyond 2) and specializations (Application Development, System Integration).

Overall, these results indicate that apprentices use AI in diverse ways and with varying intensity, both across and within task types.

Appendix B.3. Prompting Behavior and Performance

The previous section documents substantial heterogeneity in how participants interact with AI. We next examine whether these differences in usage behavior are associated with differences in performance outcomes. In particular, we study whether prompting intensity, prompt types, and the incorporation of AI-generated content correlate with productivity and comprehension across tasks.

Table 4 examines how AI usage behavior is associated with productivity across the three post-treatment tasks. Several patterns emerge. First, the number of prompts submitted is positively associated with productivity, on the Beyond 1 and Beyond 2 tasks.

Second, prompt types are differentially associated with productivity. Prompts that copy task text into ChatGPT are strongly positively associated with productivity across all three tasks. Instructed copy prompts are also positively associated with productivity on the Within task, while transform prompts show no significant association once other usage measures are taken into account. Explain prompts are negatively associated with productivity on the Within task, although the association is

only marginally significant. Overall, these patterns suggest that different prompting strategies are associated with substantially different productivity levels.

Third, direct incorporation of AI-generated content is strongly associated with higher productivity. Apprentices whose final answers contain exact copies of ChatGPT output achieve substantially higher productivity scores across all three tasks. In the multivariate specifications, exact copying is associated with increases in the share of points achieved of 18 percentage points on Within, 20 percentage points on Beyond 1, and 13 percentage points on Beyond 2.

At the same time, usage patterns associated with higher productivity are generally not associated with lower comprehension. Table B.7 shows that most AI-usage measures are not significantly related to comprehension. In particular, neither prompting intensity nor copying AI output into the final answer is associated with lower comprehension across the three tasks. We find some associations for specific prompting strategies: transform prompts are positively associated with comprehension on the Within task, while instructed copy prompts are negatively associated with comprehension on Beyond 2. Overall, these results show that usage patterns linked to higher productivity are generally not accompanied by lower task understanding.

Table B.1: Descriptive Evidence on AI Usage

	(1) Mean	(2) SD
Context of AI usage		
At vocational school	0.84	0.37
At training firm	0.77	0.42
Privately	0.77	0.42
No AI usage	0.05	0.22
Frequent AI usage (at least weekly)		
At vocational school	0.91	0.29
At training firm	0.89	0.32
Tasks using AI at vocational school		
Writing tasks	0.69	0.46
Research & comprehension	0.92	0.27
Coding tasks	0.69	0.46
Creative tasks	0.34	0.47
Tasks using AI at training firm		
Writing tasks	0.63	0.48
Research & comprehension	0.88	0.32
Coding tasks	0.65	0.48
Creative tasks	0.29	0.45
AI tools used		
ChatGPT (free)	0.66	0.47
ChatGPT (Plus)	0.32	0.47
Google Gemini (free)	0.22	0.42
Perplexity	0.20	0.40
Observations	659	

Notes: All variables are indicator variables (0/1) unless noted otherwise; means correspond to the share of respondents answering yes. Categories under *Context of AI usage* are not mutually exclusive (multiple answers possible). *Frequent AI usage* variables are recoded from a 4-category frequency question (daily; less than daily but at least weekly; less than weekly but at least monthly; less than monthly) into an indicator equal to one if AI tools are used at least weekly, and zero otherwise. Questions on frequency and tasks of AI usage at vocational school were only asked to respondents reporting AI usage at vocational school ($N = 551$); analogously for the training firm ($N = 505$). Questions on AI tools used were only asked to respondents reporting any AI usage ($N = 626$); we report the 4 most frequently used tools.

Table B.2: Who is Using AI?

	Treatment group			
	(1)	(2)	(3)	(4)
	Control	Never-taker	AI users	Diff: (3)-(2)
Demographics				
Age	23.67	23.67	23.29	-0.37
Female	0.07	0.13	0.11	-0.03
Application development	0.28	0.47	0.30	-0.16
University entrance qualification	0.59	0.53	0.59	0.06
Average grade (interim exam)	2.82	3.02	2.77	-0.25
AI				
AI usage in school	0.83	0.47	0.86	0.40***
No AI usage	0.05	0.27	0.04	-0.23***
AI literacy	-0.02	-0.43	0.04	0.47***
Observations	326	15	321	336

Notes: Sample restricted to the treatment group. Never takers are treated apprentices who used the provided AI tool for less than fifteen seconds and made no prompts; AI users are treated apprentices who used it at least fifteen seconds. Column (1) reports means for the full control group for reference. Columns (2) and (3) report means for never takers and AI users among treated apprentices, respectively. Column (4) reports the difference (3)–(2) from a two-sample *t*-test. Demographic characteristics include age, gender, occupational track, high pre-apprenticeship education (split at entrance qualification to a university of applied sciences), and interim exam grade. AI-related characteristics include prior AI usage in school, no prior AI usage, and a measure of AI literacy. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.3: Differences in Outcomes Between Never-Takers and AI Users

	(1) Control	Treatment group		
		(2) Never-taker	(3) AI users	(4) Diff: (3)-(2)
Share of points achieved				
Within	0.22	0.43	0.67	0.24***
Beyond 1	0.39	0.24	0.67	0.43***
Beyond 2	0.42	0.37	0.73	0.36***
Points achieved per minute				
Within	0.70	1.70	3.59	1.89*
Beyond 1	1.67	1.51	4.00	2.49**
Beyond 2	2.13	2.31	5.55	3.24***
Comprehension				
Within	0.65	0.88	0.71	-0.17**
Beyond 1	0.65	0.54	0.74	0.19***
Beyond 2	0.70	0.58	0.74	0.15*
Observations	326	15	321	336

Notes: Sample restricted to the treatment group. Never-takers are treated apprentices who used the provided AI tool for less than fifteen seconds and made no prompts; AI users are treated apprentices who used it at least fifteen seconds. Excludes 11 observations for which AI interaction cannot be observed. Column (1) reports means for the full control group for reference. Columns (2) and (3) report means for never-takers and AI users among treated apprentices, respectively. Column (4) reports the difference (3)–(2) from a two-sample t -test. Outcomes cover task productivity (share of points achieved, points per minute) and comprehension (share of questions answered correctly), each measured separately for the Within, Beyond 1, and Beyond 2 tasks. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.4: Differences in Outcomes Between Never-Takers and AI Users; Secondary Outcomes

	(1) Control	Treatment group		
		(2) Never-taker	(3) AI users	(4) Diff: (3)-(2)
Task experience: Fun				
Within	0.33	0.40	0.36	-0.04
Beyond 1	0.39	0.33	0.32	-0.01
Beyond 2	0.32	0.20	0.36	0.16
Task experience: Frustrating				
Within	0.49	0.20	0.25	0.05
Beyond 1	0.24	0.40	0.22	-0.18
Beyond 2	0.27	0.13	0.20	0.07
Task experience: Boring				
Within	0.13	0.13	0.21	0.07
Beyond 1	0.14	0.13	0.23	0.09
Beyond 2	0.15	0.13	0.19	0.05
Task experience: Difficulties				
Within	0.64	0.33	0.31	-0.02
Beyond 1	0.49	0.53	0.33	-0.20
Beyond 2	0.47	0.47	0.26	-0.20*
Task effort				
Within	64.57	66.00	48.41	-17.59**
Beyond 1	60.40	66.00	47.29	-18.71**
Beyond 2	54.63	46.67	46.98	0.31
Belief task correctness				
Within	30.74	50.67	61.59	10.92
Beyond 1	45.18	46.00	58.13	12.13
Beyond 2	40.89	37.33	59.25	21.92***
Observations	326	15	321	336

Notes: Sample restricted to the treatment group. Never takers are treated apprentices who used the provided AI tool for less than fifteen seconds and made no prompts; AI users are treated apprentices who used it at least fifteen seconds. Excludes 11 observations for which AI interaction cannot be observed. Column (1) reports means for the full control group for reference. Columns (2) and (3) report means for never takers and AI users among treated apprentices, respectively. Column (4) reports the difference (3)–(2) from a two-sample t -test. Outcomes cover task experience (fun, boring, frustrating, difficulties), self-reported effort, and belief about task correctness, each measured separately for the Within, Beyond 1, and Beyond 2 tasks. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.5: Descriptive Comparison of Treatment-Group Apprentices by AI Interaction Data Availability

	(1) Never-taker	AI interaction data		
		(2) Not available	(3) Available	(4) Diff: (2)-(3)
Demographics				
Age	23.67	23.02	23.37	-0.35
Female	0.13	0.06	0.12	-0.07*
University entrance qualification	0.53	0.61	0.59	0.02
Average grade (interim exam)	3.02	2.69	2.81	-0.12
AI usage before experiment				
AI usage in school	0.47	0.83	0.88	-0.05
No AI usage	0.27	0.04	0.05	-0.01
AI literacy	-0.43	0.01	0.06	-0.04
ChatGPT usage during experiment				
Number of visits	0.07	25.82	23.50	2.32
Time spent on website (in seconds)	0.27	428.94	418.53	10.41
Share of points achieved				
Within	0.43	0.70	0.66	0.04
Beyond 1	0.24	0.67	0.67	-0.01
Beyond 2	0.37	0.71	0.75	-0.04
Points achieved per minute				
Within	1.70	4.20	3.34	0.86*
Beyond 1	1.51	4.15	4.06	0.08
Beyond 2	2.31	5.70	5.72	-0.02
Comprehension				
Within	0.88	0.67	0.73	-0.06*
Beyond 1	0.54	0.72	0.75	-0.03
Beyond 2	0.58	0.68	0.76	-0.08**
Observations	15	105	227	332

Notes: Sample restricted to the treatment group. Column (1) shows means for *never-takers*, defined as participants who did not use the AI tool during the experiment. Columns (2) and (3) show means for participants for whom AI interaction data are unavailable or available, respectively; never-takers are excluded from both columns. Column (4) reports the difference in means between columns (2) and (3), estimated via a two-sample *t*-test; stars indicate statistical significance. *AI Literacy* is an index of self-assessed AI knowledge and skills. *AI Usage in School* and *No AI Usage* are indicators for whether apprentices use generative AI tools in a school context or do not use them at all. *ChatGPT visits* and *ChatGPT seconds* are based on browser history and measure the number of visits to ChatGPT and total time spent on the platform, respectively. Productivity outcomes are reported for the three treatment tasks (Within, Beyond 1, Beyond 2). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.6: Prompting Behavior Among AI Users

	(1) All	(2) Application Development	(3) System Integration
Prompt input			
<i>Average number of prompts</i>			
Total	6.28	4.62	6.86
Within task	2.06	0.88	2.47
Beyond 1 task	2.27	1.66	2.49
Beyond 2 task	1.95	2.09	1.90
<i>Share of prompts by type</i>			
Copy task text	0.50	0.49	0.50
Instructed copy	0.24	0.25	0.24
Transform response	0.10	0.10	0.10
Explain	0.12	0.15	0.11
Other	0.04	0.02	0.05
Usage of AI output			
<i>Share of students copying AI response</i>			
Within task	0.46	0.46	.
Beyond 1 task	0.43	0.25	0.49
Beyond 2 task	0.52	0.57	0.50
Observations	225	58	167

Notes: Sample restricted to treated apprentices who used the provided AI tool at least once during the three tasks. Columns report means separately for all users, the Application Development subsample, and the System Integration subsample. The first panel reports the average number of prompts submitted in total and separately for the Within, Beyond 1, and Beyond 2 tasks. The second panel reports the share of prompts by classification type, where each prompt was classified as copy (prompts copy of task text), instructed copy (prompts copy of task text with further instruction), transform (prompts to alter the text, e.g., shorten or simplify), explain (prompts to understand aspects of the task), or other; shares sum to one within each apprentice. The third panel reports the share of apprentices whose submitted answer is an exact copy of the ChatGPT response, conditional on having queried ChatGPT for the respective task block. For this measure, the Within task block is only available for Application Development apprentices. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.7: Association Between AI Usage and Comprehension (Among AI Users), by Task

	Within		Beyond 1		Beyond 2	
	Bivar.	Multivar.	Bivar.	Multivar.	Bivar.	Multivar.
	(1)	(2)	(3)	(4)	(5)	(6)
Prompt input						
No. of prompts	0.004 (0.013)	-0.000 (0.014)	-0.008 (0.010)	-0.019 (0.019)	-0.012 (0.011)	0.008 (0.019)
Copy task text	-0.012 (0.041)	-0.006 (0.050)	-0.007 (0.037)	0.013 (0.049)	-0.032 (0.036)	-0.026 (0.057)
Instruction + copy task	0.018 (0.041)	0.010 (0.048)	0.042 (0.038)	0.058 (0.044)	-0.099** (0.047)	-0.095* (0.056)
Transform response	0.103* (0.062)	0.120* (0.070)	-0.024 (0.039)	-0.017 (0.055)	-0.047 (0.052)	-0.023 (0.060)
Explain	0.010 (0.065)	-0.010 (0.066)	-0.070 (0.056)	-0.038 (0.063)	0.071 (0.052)	0.046 (0.061)
Usage of AI output						
Answer exact copy	-0.090 (0.071)	-0.113 (0.073)	0.036 (0.040)	0.058 (0.048)	-0.080** (0.039)	-0.061 (0.046)
Observations	225		225		225	

Notes: Outcome is share of comprehension questions answered correctly (Comprehension). Column pairs correspond to three tasks ordered by required skill level: Within (within curriculum), Beyond 1 and Beyond 2 (beyond curriculum). Odd-numbered columns (Bivar.) report coefficients from separate OLS regressions, each including one AI usage measure at a time. Even-numbered columns (Multivar.) report coefficients from a single OLS regression including all AI usage measures jointly. All regressions control for age, gender, education, whether the final exam took place during the field period, and session fixed effects. The usage measures are: number of OpenAI prompts sent during the task (count variable); and four indicator variables equal to one if at least one prompt was classified as copy, explain, instructed copy, or transform, respectively; and an indicator for whether the apprentice's answer was copied from ChatGPT (LCS score = 1). Heteroscedasticity-robust standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Appendix C. Experimental Details

Appendix C.1. School Recruitment

Vocational schooling in Germany is regulated at the state level; consequently, field experiments require research approval from each state’s education ministry. Between May and August 2025, we submitted parallel applications to all 16 state ministries and informed the Standing Conference of the Ministers of Education and Cultural Affairs (*Kultusministerkonferenz*, KMK). Each application package included the state-specific approval form, the apprentice survey, the task pool, the project timeline, cover letters for principals and teachers, the apprentice information flyer, the institute sponsor statement, the analysis plan, as well as the data-protection concept and consent forms.

Nine ministries approved the study: Baden-Württemberg, Berlin, Hesse, Lower Saxony, Mecklenburg-Vorpommern, North Rhine-Westphalia, Rhineland-Palatinate, Saxony-Anhalt, and Thuringia.

Across the nine approved states, we identified 163 vocational schools that met the eligibility criteria, namely offering final-year classes in the *Fachinformatiker:in* apprenticeship program. All eligible schools were contacted via a standardized information email. Schools expressing interest subsequently received details on the experiment including information material for teachers and apprentices.

In total, 25 schools agreed to participate in the study. We ultimately declined 12 of these schools: 11 because the available class size fell below the minimum threshold of 30 participants, rendering participation too costly, and one because the school could only host a session after the end of the data-collection period. Data collection was conducted in 12 schools during the study period.²⁶

Table C.1 reports the number of eligible, participating, and visited schools by federal state.

²⁶In addition, one school participated during the pilot phase but was not included in the final sample. Two schools were visited multiple times.

Table C.1: School Recruitment Outcomes by Federal State

Federal state	Eligible schools	Schools agreeing to participate	Schools visited
Baden-Württemberg	33	1	1
Berlin	2	1	0
Hesse	21	2	1
Lower Saxony	38	8	3
Mecklenburg-Vorpommern	4	0	0
North Rhine-Westphalia	41	11	5
Rhineland-Palatinate	19	1	1
Saxony-Anhalt	2	1	1
Thuringia	3	0	0
Total	163	25	12

Appendix C.2. Final Sample and Exclusion Criteria

A total of 840 apprentices participated in the experiment. Our final estimation sample consists of 673 participants. We exclude 137 individuals who did not complete the full task sequence due to internet-connectivity interruptions and an additional 19 individuals affected by other technical problems. We further exclude 10 apprentices enrolled in the *Kaufleute für Digitalisierungsmanagement* program, an IT-adjacent commercial occupation that was part of the original study design but ultimately yielded too few observations for meaningful statistical analysis. Finally, we exclude one participant in the control group who was identified as an always-taker.

Importantly, treatment assignment does not predict exclusion from the final estimation sample, suggesting that attrition and technical failures are unrelated to experimental treatment status.

Appendix C.3. Grouping the Occupational Specializations

In addition to the two larger specialization tracks *Application Development* and *System Integration*, the two smaller tracks *Data and Process Analysis* and *Digital Networking* were part of the experiment. Due to their limited sample sizes, which was anticipated prior to data collection, they were grouped with the larger ones. Thus, 5 apprentices in *Data and Process Analysis* were grouped with *Application Development*, and 15 apprentices in *Digital Networking* were grouped with *System Integration*. This reflects common practice in vocational education, where these specializations are often taught together. Robustness checks excluding these individuals are provided in Section 5.6.

Appendix C.4. Experimental Set-up and Randomization Procedure

This subsection documents the physical implementation of the experimental sessions described in Section 3.2. Figure C.1 shows representative treatment and control classrooms. Figure C.2 illustrates the on-site randomisation procedure used to assign apprentices to experimental conditions.

Each school provided two adjacent classrooms, each equipped with up to 30 laptops numbered consecutively, and apprentices were received in the hallway before the session started. Upon entering, every apprentice drew a physical card showing a PC number. Apprentices with cards 1–30 went to the treatment classroom, apprentices with cards 31–60 went to the control classroom, and each apprentice sat at the PC matching the number on their card.

Figure C.1: Representative Classroom Set-Up



Figure C.2: On-Site Randomization Procedure



Appendix D. Robustness Checks

We conduct several sets of robustness checks. We first assess the sensitivity of our results to multiple hypothesis testing. We then re-estimate our main specifications on a series of subsamples, reported as rows in Tables D.3–D.8: rows are grouped by specialization (Application Development, System Integration); within each specialization, “Main analysis sample” reports the treatment effect estimated on the full estimation sample, corresponding to the same specification underlying the main results reported in the text, and the remaining rows correspond, in the order in which they appear in the tables, to the checks described below. Finally, we assess robustness to variation in the ChatGPT model version served during data collection.

Multiple hypothesis testing. Because treatment effects vary across tasks depending on task type and difficulty level, we report treatment effects separately for each task rather than aggregating them by difficulty level or field of specialization. Across our five main outcomes, this results in a total of 30 estimated treatment effects, raising concerns about multiple hypothesis testing. To address this issue, we apply randomization inference in combination with three correction procedures: List-Shaikh-Xu, Bonferroni, and Romano-Wolf (List et al., 2019). All productivity and comprehension effects that are statistically significant in the main analysis remain significant under each correction procedure when conservatively correcting for all 30 hypotheses. For task enjoyment, one of the two significant effects remains robust to the corrections. For frustration, three out of the five significant effects remain robust (Appendix Table D.1).

Experimenter demand effects. A second concern for the interpretation of our treatment effects is experimenter demand. To mitigate this concern, sessions were framed as assessments of occupation-relevant task performance rather than as studies of AI, neither the research question nor the treatment–control contrast was disclosed to participants.

To assess potential demand effects ex post, respondents were asked at the end of the survey what they believed the study aimed to investigate. Appendix Table D.2 reports the distribution of responses by treatment status. Only a small share of respondents correctly identified any of the three research questions. As expected, treatment-group respondents were more likely to mention AI in general (17% vs. 0%), whereas control-group respondents more often referred to unrelated topics such as learning or ability (84% vs. 70%). Importantly, even among respondents mentioning AI, only a minority articulated the specific hypotheses under investigation. In the treatment group (vs. control), 2% (0%) mentioned AI productivity effects, 10% (2%) referred to a productivity–comprehension trade-off, and 0% (0%) mentioned our interest in heterogeneous productivity effects by task difficulty. Overall, these patterns suggest that respondents were largely unaware of the study’s objectives.

We further re-estimate our main specifications excluding apprentices who mention one of the study hypotheses. Results remain unchanged and are reported in Appendix Tables D.3–D.8.

Motivation and lack of incentives. A third concern is that apprentices may not have been sufficiently motivated because the study did not include financial incentives. As German

vocational-school regulations prevented us from implementing monetary performance incentives (see Section 3.3), we use an end-of-session self-report measure as an ex-post proxy for effort. Specifically, apprentices rated their agreement with the statement: “Overall, I took the study seriously and put effort into the tasks” on a Likert scale. Only 12% of apprentices (fully) disagreed with this statement (15% in the Application Development subsample and 11% in the System Integration subsample). Moreover, average agreement does not differ significantly between treatment and control groups.

To assess whether low motivation affects our findings, we re-estimate our main specifications restricting the sample to apprentices who report having taken the study seriously (88% of the full sample). Coefficient magnitudes and significance patterns remain nearly identical across outcomes and complexity levels, suggesting that our results are unlikely to be driven by differences in motivation or effort provision.

Classroom rules for leaving. One potential concern is that incentives may have differed across classrooms. Depending on the timing of the experiment during the school day and school-specific rules, some apprentices were allowed to leave once they had completed the experiment, whereas others remained in class. Being allowed to leave may have increased the incentive to finish quickly and reduced effort. To address this concern, we re-estimate our main specifications excluding apprentices who were allowed to leave after the experiment. The results remain unchanged.

Excluding other fields of specialization. Another potential concern is that apprentices from related but distinct specializations may perform differently because the tasks were designed for the two larger specializations, Application Development and System Integration. Specifically, apprentices in Data and Process Analysis were assigned the Application Development tasks, while apprentices in Digital Networking were assigned the System Integration tasks. However, this is likely a minor concern, as these specializations share the same curriculum during the first years of training and are typically taught in the same classes due to the small number of apprentices in these specializations. In line with that, only 20 apprentices in our sample belong to these additional specializations. Re-estimating our main specifications after excluding these apprentices leaves the main treatment effects virtually unchanged.

ChatGPT versions. Lastly, ChatGPT’s default model changed during our data-collection window (October 2025–January 2026): GPT-4o, GPT-5.1, and GPT-5.2 each served as the default model during part of the field phase.²⁷ This raises the concern that treatment effects differ systematically across model versions.

We address this concern in two ways. First, we test all three models on the six occupation-specific tasks (Within, Beyond 1, Beyond 2 for both Application Development and System Integration) by

²⁷Treated apprentices received access to a ChatGPT Pro account rather than a pinned API endpoint, implying that the served model corresponded to OpenAI’s current default at the time of the session. Because the interface may also route prompts to different models depending on task type when no model is pinned, within-session model variation is unobservable.

independently submitting each task prompt and grading responses using the same rubrics as for apprentices. Appendix Table D.9 shows that all three models achieve near-maximum scores on all tasks except System Integration Within task.

Second, because these calibration exercises rely on single queries and apprentices' interactions with the models may generate different outputs, we test whether average task scores and ITT estimates vary across model windows. In the treatment group, 32% used GPT-4o, 58% GPT-5.1, and 10% GPT-5.2 as the default model. Neither average scores nor treatment effects differ systematically across model windows.

Finally, because the default model was constant within sessions and treatment was randomized within session, session fixed effects absorb model variation by construction.

Table D.1: Corrections for Multiple Hypothesis Testing, Average Treatment Effects

	(1) Coefficient	(2) Rand. Inference	(3) List-Shaikh-Xu	(4) Bonferroni	(5) Romano-Wolf
Treatment effects on productivity (System Integration)					
Within	0.402***	<0.001	<0.001	0.010	0.006
Beyond 1	0.269***	<0.001	<0.001	0.010	0.009
Beyond 2	0.203***	<0.001	<0.001	0.010	0.008
Treatment effects on productivity (Application Development)					
Within	0.508***	<0.001	<0.001	0.010	0.009
Beyond 1	0.249***	<0.001	<0.001	0.010	0.006
Beyond 2	0.544***	<0.001	<0.001	0.010	0.005
Treatment effects on productivity per minute (System Integration)					
Within	1.645***	<0.001	<0.001	0.010	0.007
Beyond 1	2.519***	<0.001	<0.001	0.010	0.010
Beyond 2	3.134***	<0.001	<0.001	0.010	0.006
Treatment effects on productivity per minute (Application Development)					
Within	5.206***	<0.001	<0.001	0.010	0.007
Beyond 1	1.709***	<0.001	<0.001	0.010	0.005
Beyond 2	4.248***	<0.001	<0.001	0.010	0.008
Treatment effects on comprehension (System Integration)					
Within	0.091***	<0.001	<0.001	0.010	0.007
Beyond 1	0.110***	<0.001	<0.001	0.010	0.008
Beyond 2	-0.009	0.662	0.892	1.000	1.000
Treatment effects on comprehension (Application Development)					
Within	-0.006	0.884	0.884	1.000	0.884
Beyond 1	0.069	0.105	0.571	1.000	0.837
Beyond 2	0.179***	0.002	0.017	0.050	0.022
Treatment effects on task enjoyment (System Integration)					
Within	0.312***	<0.001	<0.001	0.010	0.009
Beyond 1	-0.121	0.158	0.680	1.000	1.000
Beyond 2	0.073	0.401	0.923	1.000	1.000
Treatment effects on task enjoyment (Application Development)					
Within	-0.087	0.554	0.907	1.000	1.000
Beyond 1	0.099	0.515	0.942	1.000	1.000
Beyond 2	0.470***	0.011	0.107	0.330	0.121
Treatment effects on frustration (System Integration)					
Within	-0.568***	<0.001	<0.001	0.010	0.010
Beyond 1	-0.098	0.326	0.904	1.000	1.000
Beyond 2	-0.160*	0.087	0.534	1.000	0.780
Treatment effects on frustration (Application Development)					
Within	-0.505***	<0.001	0.009	0.030	0.014
Beyond 1	-0.290***	0.049	0.381	1.000	0.493
Beyond 2	-0.491***	0.002	0.019	0.060	0.024

Notes: This table reports treatment effects estimated via multiple hypothesis testing regression with 3,000 bootstrap replications. P-values < 0.10 are highlighted in bold. The sample is split into two subgroups: Application Development and System Integration. For comparison, Column (1) reports OLS coefficients along with significance stars based on robust standard errors (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$), as displayed in the main tables. The p-values in Column (2) are derived using randomization inference (RI) based on 3,000 permutations. Column (3) reports p-values adjusted for multiple hypothesis testing following List, Shaikh, and Xu (2019). Column (4) reports Bonferroni-corrected p-values. Column (5) reports Romano-Wolf stepdown p-values. All regressions include the full set of controls and use robust standard errors.

Table D.2: Experimenter Demand: Respondents’ Beliefs About the Study Purpose

	Treatment (1)	Control (2)
Mentions generative AI (any context)	0.17	0.00
States a different (non-AI) study purpose	0.70	0.84
No, unclear, or meaningless answer	0.10	0.14
Mentions a specific research question		
Productivity effects	0.02	0.00
Productivity–comprehension trade-off	0.10	0.02
Heterogeneous effects by task difficulty	0.00	0.00
Observations	297	284

Notes: This table reports the share of respondents whose answer to the open-ended question “What do you think we are trying to find out with this study?” falls into each category, separately for the treatment and control group. Responses were classified by GPT-4o.

Table D.3: Share of Points Achieved: Treatment Effects (Robustness)

	Share of points achieved			
	Baseline	Within	Beyond 1	Beyond 2
	(1)	(2)	(3)	(4)
Application Development				
Main analysis sample	0.021 (0.032)	0.508*** (0.044)	0.249*** (0.040)	0.544*** (0.039)
Excluding unmotivated students	0.008 (0.037)	0.505*** (0.047)	0.228*** (0.046)	0.539*** (0.043)
Excluding students aware of the research questions	-0.007 (0.036)	0.511*** (0.048)	0.223*** (0.047)	0.521*** (0.045)
Excluding other specializations	0.024 (0.033)	0.510*** (0.044)	0.255*** (0.041)	0.551*** (0.040)
Excluding students who use other internet sources	0.007 (0.034)	0.507*** (0.047)	0.238*** (0.042)	0.548*** (0.041)
Excluding students free to leave after the experiment	-0.025 (0.034)	0.516*** (0.051)	0.226*** (0.047)	0.494*** (0.048)
System Integration				
Main analysis sample	0.009 (0.019)	0.402*** (0.025)	0.269*** (0.024)	0.203*** (0.022)
Excluding unmotivated students	0.013 (0.021)	0.405*** (0.026)	0.264*** (0.025)	0.186*** (0.023)
Excluding students aware of the research questions	-0.000 (0.022)	0.384*** (0.028)	0.251*** (0.028)	0.187*** (0.024)
Excluding other specializations	0.004 (0.020)	0.402*** (0.025)	0.264*** (0.024)	0.196*** (0.022)
Excluding students who use other internet sources	0.022 (0.020)	0.406*** (0.026)	0.276*** (0.025)	0.203*** (0.022)
Excluding students free to leave after the experiment	0.003 (0.023)	0.410*** (0.029)	0.276*** (0.029)	0.213*** (0.025)

Notes: Outcome is the share of points achieved on each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). Treatment coefficients are from OLS regressions of the outcome on the treatment indicator, including the full set of controls; heteroscedasticity-robust standard errors in parentheses. Rows are grouped by specialization (Application Development, System Integration); within each specialization: *Main analysis sample* (full sample); *Excluding unmotivated apprentices* ($N = 589$; AD $N = 168$, SI $N = 421$); *Excluding apprentices aware of the research questions* ($N = 614$; AD $N = 172$, SI $N = 442$); *Excluding other specializations* ($N = 653$; AD $N = 194$, SI $N = 459$); *Excluding apprentices who use other internet sources* ($N = 619$; AD $N = 178$, SI $N = 441$); *Excluding apprentices free to leave after the experiment* ($N = 510$; AD $N = 156$, SI $N = 354$). See Appendix D for subsample definitions. The full regression table for the main analysis sample is reported in Appendix Table A.4. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table D.4: Points per Minute: Treatment Effects (Robustness)

	Points per minute			
	Baseline	Within	Beyond 1	Beyond 2
	(1)	(2)	(3)	(4)
Application Development				
Main analysis sample	0.135 (0.176)	5.206*** (0.585)	1.709*** (0.317)	4.248*** (0.466)
Excluding unmotivated students	-0.016 (0.197)	4.616*** (0.557)	1.349*** (0.291)	3.822*** (0.409)
Excluding students aware of the research questions	-0.080 (0.209)	5.127*** (0.680)	1.599*** (0.400)	3.847*** (0.511)
Excluding other specializations	0.157 (0.178)	5.279*** (0.601)	1.752*** (0.325)	4.307*** (0.475)
Excluding students who use other internet sources	0.063 (0.183)	5.128*** (0.640)	1.454*** (0.308)	4.054*** (0.463)
Excluding students free to leave after the experiment	-0.190 (0.202)	4.680*** (0.601)	1.531*** (0.375)	3.743*** (0.543)
System Integration				
Main analysis sample	0.043 (0.101)	1.645*** (0.106)	2.519*** (0.265)	3.134*** (0.343)
Excluding unmotivated students	0.033 (0.107)	1.648*** (0.108)	2.391*** (0.275)	3.012*** (0.367)
Excluding students aware of the research questions	-0.028 (0.110)	1.566*** (0.120)	2.359*** (0.288)	2.867*** (0.365)
Excluding other specializations	0.045 (0.101)	1.665*** (0.108)	2.564*** (0.270)	3.158*** (0.351)
Excluding students who use other internet sources	0.085 (0.103)	1.638*** (0.113)	2.582*** (0.283)	3.183*** (0.369)
Excluding students free to leave after the experiment	-0.076 (0.114)	1.537*** (0.120)	2.273*** (0.284)	2.936*** (0.368)

Notes: Outcome is points achieved per minute spent on each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). Treatment coefficients are from OLS regressions of the outcome on the treatment indicator, including the full set of controls; heteroscedasticity-robust standard errors in parentheses. Rows are grouped by specialization (Application Development, System Integration); within each specialization: *Main analysis sample* (full sample); *Excluding unmotivated apprentices* ($N = 589$; AD $N = 168$, SI $N = 421$); *Excluding apprentices aware of the research questions* ($N = 614$; AD $N = 172$, SI $N = 442$); *Excluding other specializations* ($N = 653$; AD $N = 194$, SI $N = 459$); *Excluding apprentices who use other internet sources* ($N = 619$; AD $N = 178$, SI $N = 441$); *Excluding apprentices free to leave after the experiment* ($N = 510$; AD $N = 156$, SI $N = 354$). See Appendix D for subsample definitions. The full regression table for the main analysis sample is reported in Appendix Table A.7. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table D.5: Answer Given: Treatment Effects (Robustness)

	Answer given (=1)			
	Baseline	Within	Beyond 1	Beyond 2
	(1)	(2)	(3)	(4)
Application Development				
Main analysis sample	-0.011 (0.019)	0.109** (0.053)	0.058 (0.046)	0.163*** (0.059)
Excluding unmotivated students	-0.043* (0.024)	0.125** (0.054)	0.028 (0.055)	0.173*** (0.062)
Excluding students aware of the research questions	-0.015 (0.022)	0.121** (0.055)	0.063 (0.055)	0.195*** (0.065)
Excluding other specializations	-0.011 (0.020)	0.106** (0.053)	0.052 (0.047)	0.160*** (0.059)
Excluding students who use other internet sources	-0.015 (0.022)	0.138*** (0.052)	0.060 (0.047)	0.198*** (0.058)
Excluding students free to leave after the experiment	–	0.057 (0.062)	0.060 (0.056)	0.152** (0.071)
System Integration				
Main analysis sample	-0.020 (0.018)	0.110*** (0.025)	-0.002 (0.017)	0.019 (0.015)
Excluding unmotivated students	-0.016 (0.018)	0.106*** (0.025)	-0.005 (0.014)	0.010 (0.014)
Excluding students aware of the research questions	-0.008 (0.019)	0.121*** (0.029)	-0.025 (0.018)	0.013 (0.015)
Excluding other specializations	-0.026 (0.017)	0.101*** (0.025)	-0.001 (0.018)	0.020 (0.015)
Excluding students who use other internet sources	-0.009 (0.017)	0.111*** (0.026)	-0.007 (0.018)	0.017 (0.016)
Excluding students free to leave after the experiment	-0.024 (0.018)	0.146*** (0.029)	0.013 (0.021)	0.023 (0.017)

Notes: Outcome is an indicator equal to one if the participant provided an answer to the task (i.e., did not skip it). Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). Treatment coefficients are from OLS regressions of the outcome on the treatment indicator, including the full set of controls; heteroscedasticity-robust standard errors in parentheses. Rows are grouped by specialization (Application Development, System Integration); within each specialization: *Main analysis sample* (full sample); *Excluding unmotivated apprentices* ($N = 589$; AD $N = 168$, SI $N = 421$); *Excluding apprentices aware of the research questions* ($N = 614$; AD $N = 172$, SI $N = 442$); *Excluding other specializations* ($N = 653$; AD $N = 194$, SI $N = 459$); *Excluding apprentices who use other internet sources* ($N = 619$; AD $N = 178$, SI $N = 441$); *Excluding apprentices free to leave after the experiment* ($N = 510$; AD $N = 156$, SI $N = 354$). See Appendix D for subsample definitions. The full regression table for the main analysis sample is reported in Appendix Table A.5. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table D.6: Comprehension: Treatment Effects (Robustness)

	Comprehension: Share correct			
	Baseline	Within	Beyond 1	Beyond 2
	(1)	(2)	(3)	(4)
Application Development				
Main analysis sample	-0.014 (0.029)	-0.006 (0.040)	0.069 (0.043)	0.179*** (0.055)
Excluding unmotivated students	-0.033 (0.033)	-0.028 (0.039)	0.055 (0.047)	0.144** (0.062)
Excluding students aware of the research questions	-0.064* (0.033)	-0.006 (0.046)	0.058 (0.051)	0.201*** (0.062)
Excluding other specializations	-0.014 (0.030)	-0.011 (0.040)	0.065 (0.043)	0.180*** (0.055)
Excluding students who use other internet sources	-0.023 (0.032)	0.003 (0.042)	0.098** (0.044)	0.179*** (0.058)
Excluding students free to leave after the experiment	-0.035 (0.030)	-0.002 (0.044)	0.109** (0.047)	0.209*** (0.060)
System Integration				
Main analysis sample	0.023 (0.021)	0.091*** (0.026)	0.110*** (0.023)	-0.009 (0.021)
Excluding unmotivated students	0.016 (0.022)	0.069** (0.027)	0.105*** (0.025)	-0.021 (0.021)
Excluding students aware of the research questions	0.023 (0.023)	0.069** (0.029)	0.096*** (0.025)	-0.017 (0.023)
Excluding other specializations	0.018 (0.021)	0.088*** (0.026)	0.111*** (0.023)	-0.007 (0.021)
Excluding students who use other internet sources	0.015 (0.022)	0.078*** (0.027)	0.103*** (0.024)	-0.015 (0.022)
Excluding students free to leave after the experiment	0.024 (0.024)	0.099*** (0.031)	0.129*** (0.026)	-0.016 (0.024)

Notes: Outcome is the share of comprehension questions answered correctly for each task. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). Treatment coefficients are from OLS regressions of the outcome on the treatment indicator, including the full set of controls; heteroscedasticity-robust standard errors in parentheses. Rows are grouped by specialization (Application Development, System Integration); within each specialization: *Main analysis sample* (full sample); *Excluding unmotivated apprentices* ($N = 589$; AD $N = 168$, SI $N = 421$); *Excluding apprentices aware of the research questions* ($N = 614$; AD $N = 172$, SI $N = 442$); *Excluding other specializations* ($N = 653$; AD $N = 194$, SI $N = 459$); *Excluding apprentices who use other internet sources* ($N = 619$; AD $N = 178$, SI $N = 441$); *Excluding apprentices free to leave after the experiment* ($N = 510$; AD $N = 156$, SI $N = 354$). See Appendix D for subsample definitions. The full regression table for the main analysis sample is reported in Appendix Table A.9. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table D.7: Task Experience – Fun: Treatment Effects (Robustness)

	Working on the task was fun			
	Baseline	Within	Beyond 1	Beyond 2
	(1)	(2)	(3)	(4)
Application Development				
Main analysis sample	-0.103 (0.138)	-0.087 (0.150)	0.099 (0.155)	0.470*** (0.172)
Excluding unmotivated students	0.019 (0.136)	-0.038 (0.163)	0.176 (0.165)	0.605*** (0.188)
Excluding students aware of the research questions	-0.180 (0.159)	-0.093 (0.176)	0.101 (0.184)	0.490** (0.201)
Excluding other specializations	-0.112 (0.138)	-0.103 (0.152)	0.085 (0.156)	0.441** (0.172)
Excluding students who use other internet sources	-0.115 (0.147)	-0.062 (0.158)	0.217 (0.161)	0.507*** (0.183)
Excluding students free to leave after the experiment	-0.208 (0.154)	-0.140 (0.173)	0.061 (0.180)	0.536*** (0.184)
System Integration				
Main analysis sample	-0.069 (0.091)	0.312*** (0.088)	-0.121 (0.088)	0.073 (0.087)
Excluding unmotivated students	-0.010 (0.094)	0.323*** (0.095)	-0.128 (0.092)	0.012 (0.091)
Excluding students aware of the research questions	-0.069 (0.104)	0.340*** (0.099)	-0.094 (0.095)	0.132 (0.098)
Excluding other specializations	-0.083 (0.092)	0.305*** (0.089)	-0.139 (0.089)	0.059 (0.089)
Excluding students who use other internet sources	-0.058 (0.093)	0.302*** (0.090)	-0.098 (0.092)	0.122 (0.090)
Excluding students free to leave after the experiment	-0.092 (0.104)	0.410*** (0.099)	-0.091 (0.103)	0.048 (0.101)

Notes: Outcome is apprentices' agreement that the task was fun. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). Treatment coefficients are from OLS regressions of the outcome on the treatment indicator, including the full set of controls; heteroscedasticity-robust standard errors in parentheses. Rows are grouped by specialization (Application Development, System Integration); within each specialization: *Main analysis sample* (full sample); *Excluding unmotivated apprentices* ($N = 589$; AD $N = 168$, SI $N = 421$); *Excluding apprentices aware of the research questions* ($N = 614$; AD $N = 172$, SI $N = 442$); *Excluding other specializations* ($N = 653$; AD $N = 194$, SI $N = 459$); *Excluding apprentices who use other internet sources* ($N = 619$; AD $N = 178$, SI $N = 441$); *Excluding apprentices free to leave after the experiment* ($N = 510$; AD $N = 156$, SI $N = 354$). See Appendix D for subsample definitions. The full regression table for the main analysis sample is reported in Appendix Table A.12. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table D.8: Task Experience – Frustrating: Treatment Effects (Robustness)

	The task was frustrating			
	Baseline	Within	Beyond 1	Beyond 2
	(1)	(2)	(3)	(4)
Application Development				
Main analysis sample	-0.001 (0.145)	-0.505*** (0.160)	-0.290* (0.147)	-0.491*** (0.150)
Excluding unmotivated students	0.064 (0.158)	-0.484*** (0.173)	-0.300* (0.166)	-0.545*** (0.170)
Excluding students aware of the research questions	0.136 (0.171)	-0.455** (0.183)	-0.283 (0.172)	-0.458*** (0.174)
Excluding other specializations	0.021 (0.144)	-0.520*** (0.161)	-0.270* (0.150)	-0.468*** (0.152)
Excluding students who use other internet sources	-0.046 (0.155)	-0.517*** (0.172)	-0.233 (0.156)	-0.567*** (0.156)
Excluding students free to leave after the experiment	0.046 (0.163)	-0.459** (0.184)	-0.281 (0.172)	-0.502*** (0.172)
System Integration				
Main analysis sample	-0.017 (0.094)	-0.568*** (0.090)	-0.098 (0.098)	-0.160* (0.094)
Excluding unmotivated students	-0.049 (0.098)	-0.587*** (0.096)	-0.119 (0.103)	-0.141 (0.102)
Excluding students aware of the research questions	0.055 (0.103)	-0.561*** (0.101)	-0.044 (0.109)	-0.141 (0.105)
Excluding other specializations	-0.005 (0.094)	-0.550*** (0.091)	-0.089 (0.099)	-0.147 (0.095)
Excluding students who use other internet sources	-0.022 (0.098)	-0.587*** (0.094)	-0.102 (0.102)	-0.182* (0.097)
Excluding students free to leave after the experiment	-0.074 (0.105)	-0.721*** (0.102)	-0.076 (0.111)	-0.175 (0.107)

Notes: Outcome is apprentices' agreement that the task was frustrating. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). Treatment coefficients are from OLS regressions of the outcome on the treatment indicator, including the full set of controls; heteroscedasticity-robust standard errors in parentheses. Rows are grouped by specialization (Application Development, System Integration); within each specialization: *Main analysis sample* (full sample); *Excluding unmotivated apprentices* ($N = 589$; AD $N = 168$, SI $N = 421$); *Excluding apprentices aware of the research questions* ($N = 614$; AD $N = 172$, SI $N = 442$); *Excluding other specializations* ($N = 653$; AD $N = 194$, SI $N = 459$); *Excluding apprentices who use other internet sources* ($N = 619$; AD $N = 178$, SI $N = 441$); *Excluding apprentices free to leave after the experiment* ($N = 510$; AD $N = 156$, SI $N = 354$). See Appendix D for subsample definitions. The full regression table for the main analysis sample is reported in Appendix Table A.13. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

To assess what the public ChatGPT interface can in principle deliver on the occupation-specific tasks, we graded each of the three default models active during our data-collection window—GPT-4o, GPT-5.1, and GPT-5.2—on the Within, Beyond 1, and Beyond 2 tasks. We pasted the task prompts verbatim into the public ChatGPT interface and graded the resulting answers using the same rubric as for the apprentices, separately for the Application Development and System Integration versions. Table D.9 reports the resulting raw point scores. The pattern is twofold: on the Application Development Within and Beyond 1 tasks, as well as the System Integration Beyond 1 and Beyond 2 tasks, all three models reach (or essentially reach) the maximum number of points, so the ceiling provided by the AI tool is the same regardless of which default model the apprentice encountered. The model differences are concentrated on the System Integration Within task, where the achievable points vary the most across models, while the three models perform similarly on the Application Development Beyond 2 task, by not reaching the maximum achievable points.

Table D.9: ChatGPT Performance on the Tasks

Model	Application Development			System Integration		
	Within	Beyond 1	Beyond 2	Within	Beyond 1	Beyond 2
GPT-4o	40	40	36	34	39	40
GPT-5.1	40	40	36	39	40	40
GPT-5.2	40	40	35	31	40	40

Notes: Each cell reports the raw points achieved by the corresponding model on the corresponding task, graded with GPT-5.4 mini, using the same rubric used for the apprentices. The maximum score in each task is 40 points. Task contents are summarized in Table 2, and the full task materials and rubrics are reproduced in Appendix F.

Appendix E. Deviations from the Pre-Analysis Plan

Relative to the pre-analysis plan, we made a small number of deviations motivated by implementation constraints and data considerations. None of these changes affect the randomized assignment of treatment.

Sample. The pre-analysis plan included apprentices in management assistance in digitalization management. Due to lower-than-expected recruitment, only 10 apprentices from this occupation participated. Because this occupation completed a different set of experimental tasks and treatment effects are estimated at the task level, we restrict the analysis to IT specialists.

Primary outcomes. First, instead of reporting task completion time as a separate primary outcome, we report productivity measured as points earned per minute. This measure better captures productivity because tasks differed in the number of subtasks and substantial treatment effects on task completion make completion time difficult to interpret in isolation. Second, we deviate from the preregistered positive task experience index. The pre-analysis plan proposed an index combining enjoyment, frustration, and boredom. Because treatment effects on boredom were opposite in sign to those on enjoyment and frustration, we instead analyze enjoyment and frustration separately and report boredom as a secondary outcome in Appendix Table E.1. In the multiple hypothesis testing we control for the fact that we now have two outcomes for task experience instead of one index.

Outcome aggregation. Instead of aggregating outcomes across Within and Beyond tasks within occupation, we present results separately for each experimental task and specialization. During the analysis, it became apparent that treatment effects differed across tasks, making task-specific estimates more informative than the preregistered aggregation. Here again, we take this into account in the multiple hypothesis testing.

Treatment effect estimation. Although the pre-analysis plan proposed reporting both intention-to-treat (ITT) and treatment-on-the-treated (TOT) estimates, we report only ITT effects. Compliance with treatment assignment was almost perfect, implying that TOT estimates would be nearly identical to the ITT estimates. Moreover, defining compliance consistently at the task level is not feasible for the full sample.

Table E.1: Average Treatment Effects on Task Boredom (“I found the task boring”), by Task

	I found the task boring							
	Baseline		Within		Beyond 1		Beyond 2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Full sample								
Treatment	0.124*	0.074	0.286***	0.271***	0.211***	0.182**	0.101	0.099
	(0.074)	(0.074)	(0.076)	(0.077)	(0.080)	(0.081)	(0.079)	(0.079)
Controls		✓		✓		✓		✓
Control mean	0.003	0.003	0.001	0.001	0.002	0.002	0.002	0.002
Observations	673	673	673	673	673	673	673	673
Application Development								
Treatment	0.197	0.131	0.362**	0.409***	0.253*	0.265	0.336**	0.380**
	(0.139)	(0.141)	(0.147)	(0.148)	(0.145)	(0.164)	(0.145)	(0.152)
Controls		✓		✓		✓		✓
Control mean	0.000	0.000	0.000	0.000	-0.000	-0.000	0.000	0.000
Observations	199	199	199	199	199	199	199	199
System Integration								
Treatment	0.095	0.055	0.254***	0.237***	0.184*	0.161*	-0.017	-0.003
	(0.087)	(0.088)	(0.089)	(0.091)	(0.095)	(0.097)	(0.093)	(0.093)
Controls		✓		✓		✓		✓
Control mean	0.004	0.004	0.001	0.001	0.002	0.002	0.003	0.003
Observations	474	474	474	474	474	474	474	474

Notes: Outcome is apprentices’ agreement with the statement “I found the task boring” (Likert scale), standardized by the control-group mean and standard deviation within each sample. Columns correspond to the four tasks ordered by required skill level: Baseline (pre-treatment), Within (within curriculum), and Beyond 1/Beyond 2 (beyond curriculum). For each task, two specifications are reported: without and with controls. Treatment coefficients are from OLS regressions of the outcome on the treatment indicator; heteroscedasticity-robust standard errors in parentheses. Control means refer to the control-group average of the outcome. The full sample pools the Application Development and System Integration subsamples. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Appendix F. Task Materials

This appendix provides English translations of all task materials administered during the experiment. The original materials were in German; the translations below preserve the wording and structure of the original as closely as possible.

Appendix F.1. Baseline Task (Common to Application Development and System Integration Subsamples)

Task 1 (Baseline): ER Model. Look at the ER model and answer the following questions. [An ER diagram with the entity types *Company*, *Building*, *Room*, and *Equipment*, together with their cardinalities *C* and *D*, is shown to participants. The diagram is omitted from this translation.]

- (a) What kind of relationship exists between the entity types *Room* and *Building*? Explain the relationship between the entities. Interpret the cardinalities in both reading directions and indicate any dependencies, if present.
- (b) State the cardinalities used for *Room* (C) and *Equipment* (D).
- (c) According to the ER model, can a company exist without any buildings? Justify your decision.

Appendix F.2. Task Materials: Application Development Subsample

Task 2 (Within): Hotel Occupancy. As part of a statistical analysis, a function is to determine the occupancy of the rooms in a hotel. The following classes are available:

Occupancy

- `guestId` : Integer
- `dateFrom` : Date // arrival day
- `dateTo` : Date // departure day, does not count as an occupancy day
- `floorId` : Integer

An *Occupancy* object stores, for a guest's stay, the time period during which the guest was on a given floor. A public get-method exists for each attribute.

Floor

- `floorId` : Integer
 - `numberOfRooms` : Integer
-

A *Floor* object stores, for each floor, the number of rooms available. The number of rooms is not changed over the observed period. A public get-method exists for each attribute. A one-dimensional array of type *Occupancy* records on which floor a guest occupied a room from *dateFrom* until *dateTo*.

Index	guestId	dateFrom	dateTo	floorId
0	10342	01.02.25	08.02.25	3
1	10563	03.02.25	13.02.25	5

Each row of the table represents the data of one *Occupancy* object.

Note. Date objects can be compared using the operators $<$, $\leq$, $>$, $\geq$. The expression `date = date + 1` yields the date immediately following `date` (as a Date object).

Write a function in pseudocode that determines on how many days, within the given period, the specified floor was more than 80% occupied. The array `occupancy`, the start and end date of the period, and the floor are passed as parameters.

Task 3 (Beyond 1): Resilience Concept for a Payment Service. You work at an e-commerce company. The Payment Service of your application communicates with an external payment service provider and is increasingly showing stability issues. Timeouts and temporary outages of the external service lead to abandoned orders and customer complaints. An Architecture Review (ID: #7892) found that the current code does not implement appropriate error handling. To meet the agreed Service Level Agreement (SLA) of 99.9% availability, the service should be made more robust.

Review the following Java code of the Payment Service.

```
package com.shop.payment;

import java.net.http.HttpClient;
import java.net.http.HttpRequest;
import java.net.http.HttpResponse;
import java.net.URI;
import java.time.Duration;

public class PaymentGateway {

    private enum CircuitState { CLOSED, OPEN, HALF_OPEN }
    private CircuitState state = CircuitState.CLOSED;
    private int failureCount = 0;
    private long lastFailureTime = 0;
    private final HttpClient httpClient;

    public PaymentGateway() {
        this.httpClient = HttpClient.newBuilder()
            .connectTimeout(Duration.ofSeconds(5))
            .build();
    }
}
```

```

public PaymentResponse processPayment(double amount, String cardToken)
    throws PaymentException {
    // TODO: Circuit Breaker logic incomplete
    if (state == CircuitState.OPEN) {
        if (System.currentTimeMillis() - lastFailureTime < 30000) {
            throw new PaymentException("Circuit breaker open");
        } else {
            state = CircuitState.HALF_OPEN;
        }
    }
    try {
        // TODO: No retry logic for transient failures
        String requestBody = String.format(
            "{ \"amount\": %.2f, \"token\": \"%s\" }", amount, cardToken);
        HttpRequest request = HttpRequest.newBuilder()
            .uri(URI.create("https://api.payment-provider.com/charge"))
            .header("Content-Type", "application/json")
            .timeout(Duration.ofSeconds(5))
            .POST(HttpRequest.BodyPublishers.ofString(requestBody))
            .build();
        HttpResponse<String> response = httpClient.send(request,
            HttpResponse.BodyHandlers.ofString());
        if (response.statusCode() == 200) {
            // TODO: State reset on success missing
            return new PaymentResponse(true, response.body());
        } else {
            // TODO: No distinction between permanent and transient errors
            handleFailure();
            throw new PaymentException("Payment failed: "
                + response.statusCode());
        }
    } catch (Exception e) {
        handleFailure();
        throw new PaymentException("Payment service unavailable: "
            + e.getMessage());
    }
}

```

```

    }
}

private void handleFailure() {
    failureCount++;
    lastFailureTime = System.currentTimeMillis();
    // TODO: Threshold management for state transitions
    if (failureCount >= 3) {
        state = CircuitState.OPEN;
    }
}

}

class PaymentResponse {
    private final boolean success;
    private final String data;
    public PaymentResponse(boolean success, String data) {
        this.success = success;
        this.data = data;
    }
    public boolean isSuccess() { return success; }
    public String getData()    { return data; }
}

class PaymentException extends Exception {
    public PaymentException(String message) { super(message); }
}

```

Sub-questions:

- (a) Describe three concrete weaknesses in the current code and explain why they are problematic for system stability.
- (b) For two of the identified weaknesses, describe a solution approach that improves the stability and reliability of the service. Explain:
 - what concretely needs to be changed,

- what benefit the change brings to the system,
- and which parameter values you would propose (with justification).

Task 4 (Beyond 2): Classification Model for Customer Data. You are asked to implement a model that predicts, based on customer data, whether a customer purchases a product or not. Logistic regression with gradient descent is a machine-learning method for such questions. The model learns from example data by adjusting its internal parameters (weights) step by step, much like turning dials.

Given:

- data points $i = 1, 2, \dots, n$, each with $j = 1, 2, \dots, d$ features x_{ij} ,
- labels $y_1, y_2, \dots, y_n$ taking the value either 0 or 1,
- a learning rate $\eta > 0$.

- What is logistic regression, and why is logistic regression well suited for a binary classification task such as “customer buys” / “customer does not buy”?
- Formulate pseudocode for training the logistic regression:
 - Initialize the parameters/weights.
 - Implement the sigmoid function, which maps real numbers to probabilities that $y = 1$.
 - Define the objective function, often called the binary cross-entropy.
 - Update the weights via gradient descent over all data. Iterate once through all data points.
 - Return the final parameters.

Appendix F.3. Task Materials: System Integration Subsample

Task 2 (Within): IPv4 Subnet Planning. The operator of a hotel plans to build a comprehensive WLAN network to be used by different user groups. The data traffic of the user groups is to be transported in logically separated form. You are asked to contribute to the design of the WLAN concept.

The operator has identified the following user groups:

- Guests (approx. 2,500)
- Employees (approx. 150)
- IT operations (approx. 10)
- External service providers (approx. 70)

A separate IPv4 subnetwork is to be defined for each of the four user groups.

(a) Complete the information on the size of the required networks.

Reference: Overview of powers of two.

2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}	2^{11}	2^{12}	2^{13}	2^{14}
2	4	8	16	32	64	128	256	512	1024	2048	4096	8192	16384

	(Minimum) network size	Subnet mask (decimal)	Subnet mask (CIDR)
Guests			
Employees			
IT operations			
External service providers			

(b) Based on the network sizes, design an appropriate IPv4 addressing scheme. The IT director has defined several constraints that must be observed:

- The IP addressing scheme shall use the private IPv4 network 172.22.0.0/16.
- The subnets shall be no larger than necessary.
- The IP ranges of the subnets shall be directly adjacent.

Carry out the IPv4 subnet planning and complete the table below.

Subnet	Network address	First host address	Last host address	Broadcast address	Subnet mask (CIDR)

Participants may use the space below for notes or calculation steps. The final answer should be entered in the table.

Task 3 (Beyond 1): Zero-Trust Security Concept. Schneider Handels GmbH currently operates a classical network-based security architecture with a firewall at the network perimeter and VPN access for mobile employees. Due to the rise of home-office workplaces and several security incidents in recent months, management has decided to modernize the security concept.

You have been commissioned to design a concept for introducing basic Zero-Trust principles.

Current IT infrastructure:

- 3 sites, each with a local network
- Central file servers and application servers at the main site

- VPN access for external employees
- Classical firewall structure (perimeter protection only)
- Heterogeneous end devices (Windows PCs, notebooks, some personal mobile devices)

Your tasks:

- (a) Name three weaknesses of the current security architecture in view of the described situation, and explain why they are problematic.
- (b) Develop three priority measures for introducing basic Zero-Trust principles that are implementable within the next 12 months. For each measure, describe:
 - what is to be implemented,
 - which Zero-Trust principle is thereby realized,
 - which technical components/concepts are required.
- (c) Decide, taking into account security benefit, implementation effort, and dependencies, which measure should be implemented first. Briefly justify your decision.

Task 4 (Beyond 2): Hybrid Cloud Solution. You are involved in modernizing the IT infrastructure of a company. Management wants to make the IT infrastructure future-proof and is evaluating a hybrid cloud solution. The goal is a 25% cost reduction while simultaneously increasing availability to at least 99.5%.

Current situation:

- 12 physical servers (8 years old, no warranty remaining)
- 45 virtual machines (VMs) on a virtualization platform
- 3 TB productive database (customer data, order management)
- Monthly operating costs: €8,500

- Current availability: 98.5%
- Compliance requirement: customer data must be stored in Germany

Sub-questions:

- (a) Identify three central problems of the current infrastructure and briefly explain why these problems are relevant for the company.
- (b) Decide, for the following three categories, what remains on-premises and what moves to the cloud. Justify your decision.
 - Databases, customer data
 - Standard applications and office services
 - Development and test environments
- (c) Name one risk of the hybrid cloud solution and propose a countermeasure.

Appendix G. Scoring Rubric

Appendix G.1. Scoring Rubric: Baseline Task (Both Subsamples)

Table G.1: Scoring Rubric for the Baseline Task (Both Subsamples)

Task	Criterion	Pts.	Max.
Task 1 (Baseline): ER Model			20
1a)	Relationship type: correctly identifies the 1:n or n:1 relationship	2	10
	Building $\rightarrow$ room: correctly explains that a building can contain multiple rooms	2	
	Room $\rightarrow$ building: correctly explains that a room belongs to exactly one building	2	
	Dependency: correctly describes the existence dependency according to the ER model	2	
	Technical understanding: coherent presentation of the relationship with appropriate terminology and level of detail	2	
1b)	Correctly identifies the cardinality of Room (C); Description of relationship as $m : n$ and $n : m$ equally correct	2	4
	Correctly identifies the cardinality of Equipment (D); Description of relationship as $m : n$ and $n : m$ equally correct	2	
1c)	Correctness of answer: correctly answers “Yes”	2	6
	Justification: refers to the (absent) existence dependency as justification	2	
	Technical understanding: coherently links cardinality (n includes 0) and existence dependency	2	

Appendix G.2. Scoring Rubric: Application Development Subsample

Table G.2: Scoring Rubric for the Application Development Subsample

Task	Criterion	Pts.	Max.
Task 2 (Within): Hotel Occupancy			40
	Function signature: declares all parameters (occupancy array, start date, end date, floor or floor ID) and a meaningful return type (number of days)	8	40
	Return value: returns the number of days with more than 80 % occupancy (4) using a correct counter variable (4)	8	
	Date-range iteration: loop that considers every day from start date to end date (<code>day = day + 1</code>)	4	
	Occupancy date logic: an occupancy counts for a day if <code>datumVon ≤ day < datumBis</code>	4	
	Floor filter: only occupancies of the given floor are counted (comparison of <code>stockwerkId</code>)	4	
	Threshold calculation: daily occupancy = occupied rooms / number of rooms on the floor; the day counts if occupancy exceeds 80 %	4	
	Technical understanding: correct order of steps, executability in principle, readability and efficiency of the pseudocode	8	
Task 3 (Beyond 1): Resilience Concept for Payment Service			40
3a)	Weaknesses 1–3, identification: correctly names or locates three distinct, applicable weaknesses in the code	3×2	20
	Weaknesses 1–3, reasoning: plausibly explains why each weakness is problematic for system stability	3×2	
	Core problem recognition: recognizes that the service does not handle temporary failures robustly and that the Circuit Breaker logic is incomplete	4	
	Technical understanding: uses appropriate terminology (e.g., retry, backoff, Circuit Breaker, CLOSED/OPEN/HALF-OPEN, permanent/temporary, idempotency, threshold)	4	
3b)	Solutions 1–2, change: describes a concrete technical change for each of two different weaknesses	2×4	20

Continued on next page.

(Table G.2 continued.)

Task	Criterion	Pts.	Max.
	Solutions 1–2, benefit: explains the benefit of each change for system stability and reliability	2×2	
	Solutions 1–2, parameters: proposes parameter values for each solution	2×2	
	Technical understanding: technically sound answer using appropriate terminology	4	
Task 4 (Beyond 2): Classification Model			40
4a)	Method description: describes logistic regression as a statistical/machine-learning method that models the probability of $y = 1$ from a weighted linear combination of the features via the sigmoid function	5	10
	Suitability for binary classification: the sigmoid maps any real value to $(0, 1)$, interpretable as a probability; a threshold (e.g., 0.5) separates “buys” from “does not buy”	5	
4b)	Initialization: weights w (and bias b) sensibly initialized, e.g., $w = 0$ (or small random values), $b = 0$	4	30
	Sigmoid function: $\sigma(z) = 1/(1+e^{-z})$ with $z = w \cdot x + b$ correctly stated	4	
	Cost function: binary cross-entropy $J = -\frac{1}{n} \sum [y \log \hat{y} + (1 - y) \log(1 - \hat{y})]$ with $\hat{y} = \sigma(w \cdot x + b)$	4	
	Gradient-descent update: $w_j \leftarrow w_j - \eta(\hat{y} - y)x_j$ (or averaged over all n) and $b \leftarrow b - \eta(\hat{y} - y)$; one pass over all data points	4	
	Parameter return: returns the final parameters (w, b) at the end of training	4	
	Technical understanding: completeness and logic of the training procedure (initialization $\rightarrow$ prediction $\rightarrow$ error $\rightarrow$ update $\rightarrow$ return), correctness and consistency of the formulas, level of detail, structure, and readability of the pseudocode	10	

Appendix G.3. Scoring Rubric: System Integration Subsample

Table G.3: Scoring Rubric for the System Integration Subsample

Task	Criterion	Pts.	Max.
Task 2 (Within): IPv4 Subnet Planning			40
2a)	Network sizes: correct minimum network size (power of two) for each of the four user groups	4×2	16
	Subnet masks: correct subnet mask in decimal notation for each subnet	4×1	
	CIDR notation: correct subnet mask in CIDR notation for each subnet	4×1	
2b)	Subnet assignment: correct assignment of the address ranges to the four subnets (0.5 per subnet)	2	24
	Network addresses: correct network address for each subnet	4×2	
	First host addresses: correct first usable host address for each subnet	4×1	
	Last host addresses: correct last usable host address for each subnet	4×1	
	Broadcast addresses: correct broadcast address for each subnet	4×1	
	CIDR notation in the addressing scheme: correct prefix for each subnet (0.5 per subnet)	2	
Task 3 (Beyond 1): Zero-Trust Security Concept			40
3a)	Weaknesses 1–3, identification: names three distinct weaknesses of the current security architecture that apply to the described situation	3×2	12
	Weaknesses 1–3, reasoning: plausibly explains why each weakness is problematic in the described situation	3×2	
3b)	Measures 1–3, action: names three suitable measures for introducing Zero-Trust principles	3×2	20
	Measures 1–3, principle: correctly assigns a Zero-Trust principle to each measure	3×2	
	Measures 1–3, components: names concrete technical components or concepts matching each measure	3×2	
	Feasibility: all measures can realistically be implemented within 12 months	2	

Continued on next page.

(Table G.3 continued.)

Task	Criterion	Pts.	Max.
3c)	Selection: selects a suitable measure to be implemented first	2	8
	Reasoning: addresses security benefit, implementation effort, and dependencies	3×2	
Task 4 (Beyond 2): Hybrid Cloud Solution			40
4a)	Problems 1–3, identification: names three central problems of the current infrastructure	3×2	15
	Problems 1–3, reasoning: explains why each problem is business-relevant for the company	3×3	
4b)	Decisions: states an on-premises/cloud decision for each of the three categories (databases and customer data, standard applications and office services, development/test environments)	3×2	18
	Reasoning: plausibly justifies each decision, taking the requirements from the scenario into account	3×4	
4c)	Risk: names a relevant risk of the hybrid cloud solution	3	7
	Countermeasure: describes an appropriate countermeasure	4	

Appendix H. Comprehension Questions

After completing each task and the associated experience questions, participants answered a set of multiple-choice comprehension questions targeting the core concepts of that task. All questions were administered in German; the translations below are ours. Each question has exactly one correct answer (marked in bold). The response option “I don’t know” was always offered in order to discourage guessing. The Baseline block was answered by all participants; the Within/Beyond 1/Beyond 2 blocks were answered only by participants assigned to the corresponding occupational subsample.²⁸

Baseline Task — ER Model (all participants)

Q1. Which statement about existential dependency is correct?

- (a) **Rooms only exist if there is a corresponding building.**
- (b) Buildings only exist if there are rooms.
- (c) Rooms and buildings are independent of each other.
- (d) I don’t know.

Q2. What does the cardinality $n:1$ between Room and Building mean?

- (a) A building can contain several rooms, and a room can belong to several buildings.
- (b) **A building can contain one or more rooms, but a room always belongs to exactly one building.**
- (c) Neither rooms nor buildings have a fixed assignment.
- (d) I don’t know.

Q3. Which of the following statements correctly describes the difference between attributes and entities?

- (a) **Attributes describe properties of entities and are directly attached to them.**
- (b) Attributes are objects that have relationships with other objects.
- (c) Attributes and entities are identical and can be interchanged at will.
- (d) I don’t know.

²⁸One comprehension question had to be dropped from the analysis because, due to a survey-flow issue, it was not displayed to all participants.

Application Development Subsample — Within: Hotel Occupancy

Q1. Which approach best matches the required function?

- (a) Determine the maximum daily occupancy.
- (b) **For every day in the period, check all bookings for overlap and floor ID, count occupied rooms, and compute the ratio.**
- (c) Iterate through the bookings once and remember the longest booking.
- (d) I don't know.

Q2. How is the next day obtained correctly?

- (a) `date++` is mandatory.
- (b) **`date = date + 1.`**
- (c) `date = next(date).`
- (d) I don't know.

Q3. A guest checks out on 10 February (`datumBis = 10.02`). Does 10 February count as an occupied day?

- (a) **No, `datumBis` is exclusive.**
- (b) Yes, always.
- (c) Only if `start = 10.02`.
- (d) I don't know.

Q4. Which parameters does the function require at minimum, according to the task description?

- (a) **`(belegung[], start, ende, stockwerk).`**
- (b) `(start, ende, stockwerkId).`
- (c) `(belegung[], start, ende).`
- (d) I don't know.

Application Development Subsample — Beyond 1: Payment Service Resilience

Q1. Which main problem is the circuit-breaker pattern in the payment service meant to solve?

- (a) Memory leaks in the payment service.
- (b) **Timeouts of external payment providers.**

- (c) Overload of the internal database.
- (d) I don't know.

Q2. A 99.9% SLA means ...

- (a) at most approximately 9 hours of downtime per month.
- (b) **at most approximately 43 minutes of downtime per month.**
- (c) that the service must successfully process at least 999 out of 1,000 requests.
- (d) I don't know.

Q3. Which state transition is missing in the given code when a request succeeds in the `HALF_OPEN` state?

- (a) **From `HALF_OPEN` to `CLOSED` with a reset of `failureCount`.**
- (b) From `HALF_OPEN` to `OPEN` with an increased timeout.
- (c) From `HALF_OPEN` to `CLOSED` without changing `failureCount`.
- (d) I don't know.

Q4. Why is it important to distinguish between temporary and permanent errors?

- (a) To store different error messages in the database.
- (b) To compute performance metrics correctly.
- (c) **To retry on temporary errors but abort immediately on permanent errors.**
- (d) I don't know.

Application Development Subsample — Beyond 2: Classification Model

Q1. Which order corresponds to the required pseudocode structure?

- (a) Loss (BCE) → prediction (sigmoid) → initialization → update → return.
- (b) **Initialization → prediction (sigmoid) → loss (BCE) → update → return.**
- (c) Initialization → update → prediction (sigmoid) → loss (BCE) → return.
- (d) I don't know.

Q2. Why is logistic regression well suited for binary classification?

- (a) **Because it produces probabilities for two classes and learns a linear decision boundary.**

- (b) Because it requires no training data.
- (c) Because it always produces nonlinear boundaries.
- (d) I don't know.

Q3. What should be returned at the end of training?

- (a) The trajectory of the loss values.
- (b) The learning rate.
- (c) **The final model parameters (e.g., w and b).**
- (d) I don't know.

Q4. Which additional, sensible stopping criterion (in addition to a fixed number of epochs E) fits the pseudocode approach?

- (a) Stop as soon as accuracy equals 100%.
- (b) Stop when the gradients are exactly zero.
- (c) **Stop when the decrease in loss falls below a threshold ϵ .**
- (d) I don't know.

System Integration Subsample — Within: Subnetting

Q1. In which order do you plan the subnets if they must be directly adjacent and no larger than necessary?

- (a) Alphabetically by group.
- (b) **From the largest to the smallest network.**
- (c) From the smallest to the largest network.
- (d) I don't know.

Q2. The base network is 172.22.0.0/16. Which subnet is a correct first assignment for the guests?

- (a) **172.22.0.0/20.**
- (b) 172.22.8.0/20.
- (c) 172.22.15.0/20.
- (d) I don't know.

Q3. Which statement about the host rule is correct?

- (a) First host = network address, last host = broadcast.
- (b) **First host = network + 1, last host = broadcast - 1.**
- (c) Both can be chosen freely.
- (d) I don't know.

Q4. In the range 172.22.16.0/24 (employees), which is the first host address?

- (a) **172.22.16.1.**
- (b) 172.22.16.0.
- (c) 172.22.16.255.
- (d) I don't know.

System Integration Subsample — Beyond 1: Zero-Trust Security

Q1. Which principle is at the heart of the Zero-Trust model?

- (a) Access within the corporate network is granted automatically.
- (b) **By default, every access request is denied until identity and authorization have been verified.**
- (c) Once authenticated, access remains permanently valid.
- (d) I don't know.

Q2. Why is the classic VPN-based perimeter model problematic?

- (a) VPNs are not encrypted.
- (b) **After a single successful login, the entire internal network is treated as trusted.**
- (c) VPNs only work with a fixed IP address.
- (d) I don't know.

Q3. Which item is *not* an obvious vulnerability candidate in this situation?

- (a) Missing internal segmentation and lateral movement.
- (b) Private mobile devices (BYOD).
- (c) **Central file and application servers at the main site.**
- (d) I don't know.

Q4. What is a sensible first step towards implementing the Zero-Trust principle?

- (a) **Multi-factor authentication (MFA).**
- (b) Turning off the VPN.
- (c) Granting access based solely on the network location.
- (d) I don't know.

System Integration Subsample — Beyond 2: Hybrid Cloud Solution

Q1. Which compliance requirement is explicitly stated?

- (a) Customer data must be strongly encrypted.
- (b) **Customer data must be stored in Germany.**
- (c) Backups must be performed daily.
- (d) I don't know.

Q2. Which point most clearly signals the need for modernization?

- (a) **12 physical servers, 8 years old, no warranty.**
- (b) 45 VMs on a single virtualization platform.
- (c) Use of standard office software.
- (d) I don't know.

Q3. Which argument most strongly supports moving the development and test environments to the cloud in this scenario?

- (a) Development environments contain particularly sensitive data.
- (b) **Cloud environments offer scalability and can be provisioned and decommis-
sioned quickly on demand, which saves costs.**
- (c) On-premises servers are technically unsuitable for development environments.
- (d) I don't know.

Q4. Which of the following is a typical risk of a hybrid cloud solution?

- (a) Cloud services are fundamentally less secure than on-premises solutions.
- (b) Hybrid clouds are always more expensive than pure on-premises solutions.
- (c) **The internet connection becomes a single point of failure — if it goes down,
cloud resources are unreachable.**
- (d) I don't know.