Pilot GenAI

Research Report

Research Group Learning Technology & Analytics

2025

let's change
YOU. US. THE WORLD.

THE HAGUE
UNIVERSITY OF
APPLIED SCIENCES

Pilot GenAI

Research Report

Authors*

Manika Garg, Aslam Tanjung, Sunčica Bruck, Theo Bakker

Department

Research Group Learning Technology & Analytics

Date

7 October 2025

Type project

Research project

Version

1.0

*Manika Garg served as Project Leader (Research), designed the study, supervised the pilot, and led the writing of the report. Theo Bakker provided supervision, reviewed the report, and offered critical feedback. Aslam Tanjung acted as Data Engineer, carrying out data processing and analysis. Sunčica Bruck conducted the literature review and contributed to developing the Theory of Change. All authors approved the final version of the report.

Foreword

At the request of Facilitaire Zaken & Information Technology (FZ/IT), the research group Learning Technology & Analytics (LTA) at The Hague University of Applied Sciences (THUAS) designed and implemented the research into the Pilot GenAI. The study investigated the influence of generative AI on the perceived usability, quality, and productivity of education, research, and support staff.

This report presents the research questions, methods, findings, and reflections from the pilot. It combines literature review, theory of change, survey data, interviews, and an experimental study to provide evidence on where GenAI adds value, where limitations remain, and what conditions are needed for responsible scaling.

The report is intended to inform the AI Steering Group, the wider THUAS community and all the stakeholders in their decisions about the future use of GenAI within the knowledge institution.

Acknowledgment

We thank the AI Steering Group of THUAS for commissioning the research into the Pilot GenAI. Special recognition goes to Eva Willems, project lead, and the FZ/IT team for their support. We also acknowledge the AI Expert Team for developing the personas, the HCTL team for providing training to participants, and Npuls for making the EduGenAI platform available. We are grateful to all staff members who participated in the pilot.

Table of Contents

Executive Summary

The Pilot GenAI aimed to examine how generative AI influences the usability, quality, and productivity of routine professional tasks carried out by education, research, and support staff within THUAS. The pilot was designed to generate empirical evidence to inform a roadmap for the responsible and sustainable integration of GenAI across the institution.

The pilot's findings show that GenAI has the potential to support the professional tasks of THUAS staff members, provided GenAI is integrated carefully and structured.

Recommendations for GenAI scaling within THUAS

Based on the pilot's findings, a set of recommendations has been developed to guide the AI Steering Committee in setting priorities for the responsible and sustainable integration of GenAI. The recommendations are grounded in the full evidence base of the pilot, including literature, theory of change, survey data, interviews, and a micro-RCT, which is presented in the subsequent sections of this report.

Differences in digital literacy, task appropriateness for GenAI support, and staff attitudes across domains (education, research, and support) underline that scaling GenAI within THUAS cannot follow a one-size-fits-all model but should instead be grounded in differentiated support and careful governance.

1. **Strategy**

- **Phased scaling.** Begin in domains with higher readiness, such as education and research, while continuing to explore relevant use cases in support services. Scaling should proceed gradually, with continuous monitoring and feedback to adjust training, governance, and integration.
- **Ensure sustainability.** Address environmental concerns voiced by staff by encouraging conscious use and exploring energy-efficient infrastructure.
- **Ensure human oversight.** Keep reminding staff that AI-generated outputs require verification and that reliance on GenAI must not come at the expense of critical thinking.

2. **HR and Training**

- **Targeted training.** Digital literacy strongly shaped outcomes: experienced users achieved better results, while beginners felt uncertain. Training should therefore be tiered, with introductory modules for newcomers and advanced sessions for experienced users.
- **Embedded support.** Tutorials, prompt libraries, and contextual examples inside the platform could lower the learning curve and reduce reliance on trial and error.
- **Peer learning.** A helpdesk function and communities of practice where staff exchange experiences would further support adoption.

3. **Processes**

- **Culture of safe experimentation.** Staff need space to try GenAI, make mistakes, and ask questions without fear of judgment. Small pilots within faculties or teams, framed as opportunities for collective learning can help normalize responsible experimentation.
- **Share use cases.** Highlighting positive and practical examples across domains can reduce skepticism and show how GenAI creates value.

4. Procedures and IT

- **Strengthen trust and governance.** GDPR compliance and institutional oversight were key reasons staff trusted EduGenAI over other platforms. THUAS should clearly communicate what data is collected, how it is secured, and what constitutes ethical use regarding the GenAI platform in use.
- **Ethical oversight.** An advisory mechanism or ethical board could provide guidance on issues such as bias, environmental impact, and appropriate boundaries of GenAI use.
- **Workflow integration.** Some experienced users viewed EduGenAI as less convenient than external tools because it required switching systems. For scaling to succeed, GenAI should integrate smoothly with the tools already in daily use.

Summary

The pilot GenAI examined how generative AI can support the daily work of education, research, and support staff, focusing on three dimensions: usability, productivity, and quality. This report presents the results of research into the Pilot GenAI conducted at THUAS.

The study was designed based on the Dutch 3E Framework (developed by LTA) that combines three levels of evidence to evaluate education technology.

- Bronze-level evidence included a literature review and a theory of change to map expectations.
- Silver-level evidence was collected through surveys and semi-structured interviews, capturing staff perceptions before and after using GenAI.
- Gold-level evidence was generated through a micro-Randomized Controlled Trial, which tested causal effects on task performance.

In total, 205 staff registered for the pilot. Of these, 159 met the inclusion criteria, 106 completed both surveys, 67 participated in the RCT, and 10 were interviewed. Participants represented education, research, and support roles, with diverse ages, digital literacy, and prior AI experience.

Findings based on Dimension (Usability, Quality, and Productivity):

- Findings show that usability was rated positively. Ease of use remained high across domains (Pre = 4.77; Post = 4.65; ns). However, perceived usefulness declined after hands-on experience (Pre = 5.33; Post = 4.20; $p < .001$) as staff became more critical once they engaged with the tool in practice.
- Quality showed the strongest effect. Outputs produced with GenAI received higher ratings in GPT-based evaluations (M = 4.62 vs. 3.68; $p < .001$), while self-reported scores showed no significant difference (M = 4.07 vs. 3.82; $p = .120$). At the same time, participants stressed the need for human oversight because of concerns about accuracy and hallucinations.
- Productivity results were mixed. GenAI-supported tasks were completed faster in the RCT (M = 38.8 minutes vs. 45.9 minutes; $p = .015$). GPT-based productivity scores were significantly higher ($p < .001$) while self-reported productivity increases were less pronounced ($p = .008$), reflecting the extra effort required to learn the new technology, refine prompts, and review outputs.

Findings based on Domain (Education, Research, and Support): Differences were shaped by varying levels of digital literacy, prior AI experience, and the appropriateness of tasks for GenAI support.

- Teachers valued GenAI for structuring and drafting lesson materials, but stressed that pedagogical judgment was essential to adapt outputs to student needs.
- Researchers, with generally higher digital literacy, reported the strongest benefits, especially for writing tasks such as summarization and report drafting. However, they also raised concerns about hallucinations and fabricated references.
- Support staff had the most divided experiences. Some recognized time savings in tasks like meeting summaries, while others struggled to connect GenAI to their context-specific roles, often involving personal data. Their lower digital literacy contributed to hesitations in experimenting with the technology.

Conditions shaping use: Managerial support, GDPR compliance, technological training, and helpdesk resources enabled uptake, while skepticism about accuracy, uneven digital literacy, environmental concerns, and fears of overreliance hindered wider adoption.

Conclusion: The Pilot GenAI shows that GenAI can improve efficiency and quality in higher education, but its value depends on user readiness, cultural conditions, and continued human oversight. Scaling at THUAS should proceed in phases, with differentiated training, strong governance, integration into existing workflows, and open attention to ethics and sustainability.

1 Introduction

The **THUAS Strategic Plan 2023–2028** emphasizes digitalization, co-creation, inclusivity, and research-driven innovation [1]. However, like many other academic institutions, THUAS faces challenges around digitalization in practice. THUAS staff (education/ teachers, researchers and support) differ widely in their levels of digital literacy. Some are confident in experimenting with and embedding new technologies, while others require more guidance and support.

At a time when Generative AI (GenAI) is rapidly gaining attention and many staff members are already exploring its use, THUAS seeks to better understand its impact. The institution aims to base future decisions on evidence, ensuring that any integration of GenAI into professional workflows is responsible, effective, and aligned with institutional goals. In line with this, the **AI Steering Group** commissioned the project **Pilot GenAI** to gather empirical evidence on how Generative AI can support THUAS staff, where its limitations lie, and under what conditions it can be responsibly integrated within the institution.

The following sub-sections introduce GenAI and the research questions guiding this study.

1.1 Generative AI

Generative Artificial Intelligence, or GenAI, refers to models trained on large datasets that can generate new content, such as text, images, or code, when prompted by users [2]. These models recognize patterns in existing data and use probabilistic methods to predict and construct coherent outputs. While AI has been in development for decades, generative models have advanced rapidly since 2017 with the introduction of transformer architectures. Their mainstream adoption accelerated in 2022 following the release of large-scale applications such as ChatGPT [3]. Today, GenAI is applied in many sectors, including business, healthcare, law, and education, making it a fast-growing part of the professional and academic world.

In education, the use of GenAI has sparked debate. On one hand, it supports efficiency by automating repetitive tasks, assisting with writing, and generating educational content. On the other hand, concerns remain regarding over-reliance, potential loss of critical skills, accuracy of outputs, and the ethical use of institutional data [4]. These discussions are directly relevant for THUAS, where many staff members have already begun experimenting with GenAI in their daily professional roles.

1.2 Research Questions

The project Pilot GenAI aimed to understand how the use of GenAI by THUAS staff affects three core dimensions: Usability, Quality, and Productivity.

1. **Usability** refers to how staff experience GenAI in relation to their professional tasks. This includes both ease of use (whether the platform is intuitive and accessible) and usefulness (whether the tool supports staff in completing their tasks more effectively).
2. **Quality** refers to the standard of work produced when completing professional tasks. It reflects both the technical accuracy and the professional relevance of outputs generated with GenAI.
3. **Productivity** refers to the extent to which staff can perform their professional tasks more efficiently when supported by GenAI. It provides insight into whether GenAI helped staff save time, complete tasks faster, or work more effectively.

Usability focuses on perceptions and experience, quality of outcomes, and productivity, which addresses the efficiency of work processes.

The project considered these dimensions across three domains: **education staff/ teachers** (teaching and pedagogical roles), **research staff** (academic and applied research roles), and **support staff** (administrative, policy, and service functions).

> In summary, there are three **research questions** guiding this study:
> 1. How do staff perceive the usability of GenAI?
> 2. In what ways does GenAI influence the quality of staff outputs?
> 3. In what ways does GenAI influence staff productivity?

The answers to these questions help identify both opportunities and limitations for adoption and provide input for a roadmap toward the responsible and sustainable integration of GenAI within THUAS.

The rest of the report is structured as follows: Section 2 outlines the methodology; Section 3 reports the findings; Section 4 describes experiences with the EduGenAI platform; Section 5 discusses the results.

2 Methodology

This section describes how the Pilot GenAI was designed and implemented. It outlines the research design based on the Dutch 3E Framework, the EduGenAI platform used within the pilot, explains the sequence of activities carried out during the pilot, and provides details on participants and ethical considerations.

2.1 Research Design

The Pilot GenAI was designed according to **the Dutch (Evidence-informed Evaluation of EdTech) 3E framework** [5]. The framework was developed by the Learning Technology & Analytics research group at THUAS in collaboration with Npuls. It was created to provide institutions with a structured way of evaluating educational technology tools, combining different levels of evidence to assess both potential and actual impact.

> The Dutch 3E Framework distinguishes three levels of evidence:
> (1) Bronze evidence (theory-based insights to establish expectations),
> (2) Silver evidence (practice-based insights capturing perceptions) and
> (3) Gold evidence (causal evidence for measurable impact).

Each level contributes a different perspective, ranging from theory to practice to causality. Together, these methods provided a comprehensive picture of how GenAI affects the usability, quality, and productivity of work at THUAS.

1. **Bronze –** This level used two methods to establish expectations:

    - A literature review (detailed in Section 3.1.1) was conducted at the start of the project. Relevant international and national studies on the use of GenAI in education, research, and support staff within educational institutions were analyzed. The review identified potential benefits and risks of integrating GenAI into the professional workflows of a knowledge institution.

    - A Theory of Change (ToC) was then developed (Section 3.1.2). The ToC made explicit the assumptions behind the pilot and served both as a set of hypotheses for this study and as a foundation for planning the future scaling and integration of GenAI into THUAS workflows.

2. **Silver –** This level used two methods to capture staff perceptions:

Surveys were conducted twice online. Both surveys followed the Technology Acceptance Model (TAM) [6]. TAM is a widely used survey for studying the adoption of new technologies. It measures two core constructs: (1) Perceived Ease of Use (PEoU), the degree to which users expect the system to be free of effort, and (2) Perceived Usefulness (PU), the degree to which users believe the system will enhance their work.

The first baseline survey (Pre-Pilot TAM) was administered before access to the EduGenAI platform (discussed in Section 2.2) to measure expectations. The follow-up survey (Post-Pilot TAM) was conducted after four weeks of use to capture actual experiences. The responses from both surveys were then analyzed to assess changes in perception resulting from actual platform usage.

- **Interviews** were conducted with 10 participants across the three domains and with varying levels of digital literacy. To eliminate bias from only including enthusiastic participants, staff who had dropped out of the pilot were also contacted and invited to share their perspectives. The interviews explored in detail how participants experienced GenAI, the benefits they observed, and the challenges they encountered.

  These interviews were held online via MS Teams, audio-recorded with consent, transcribed, and anonymized. Interview reports were then prepared and shared with participants for verification of interpretation. Insights were analyzed through thematic coding, identifying recurring patterns and differences across domains. Subsequently, these themes were compared with the other findings to provide richer context and to explain certain observations.

3. **Gold** – This level measured causal effects through an experiment.

- A **micro-RCT** was organized after the four-week access period to EduGenAI. RCT is a method in which participants are randomly assigned to groups to test the effect of an intervention [7]. In this study, Group A (experimental group) completed tasks with GenAI, while Group B (control group) completed the same tasks without GenAI. Randomization was carried out separately within each of the three domains (education, research, and support). A stratified sampling approach was used, based first on participants' level of digital literacy and then on age. Within each domain, participants were then randomly assigned to either the experimental group (GenAI use) or the control group (no GenAI use). A snapshot of participants on the day of RCT is shown in Figure 1.

  It is referred to as micro-RCT because it was conducted over a short time frame (one day in June 2025) with a relatively small number of participants. This method was chosen to obtain causal evidence on productivity and quality, insights that could not be derived from surveys or interviews alone.

  During RCT, each participant completed a task designed to reflect their professional practice. Teachers prepared a lesson plan, researchers drafted a short research abstract, and support staff created a meeting summary or policy update. Outcome measures included task duration, self-assessed quality, and GPT-evaluated quality. The latter were performed by two LLMs (GPT-4.1 and mistral-large-2411). These LLMs were selected because of their strong benchmark performance and complementary approaches. For the evaluation of outputs, LLMs were prompted with structured instructions to assess text quality. Each model was asked to rate outputs against predefined criteria (clarity, structure, relevance, and coherence) on a 1–5 scale. Different prompt sets were tailored to the task type to ensure the evaluation reflected realistic standards for THUAS staff work. Finally, Inter-rater reliability was checked across both models to ensure consistency.

Figure 1. The participants during the RCT day

2.2 EduGenAI Platform

The pilot was conducted using the EduGenAI platform [8], developed by Npuls [9], the national digital transformation programme for higher and vocational education in the Netherlands. Npuls aims to accelerate digitalization in education by developing shared infrastructures, knowledge, and tools for institutions. The EduGenAI platform is part of this effort and is designed to provide a secure and GDPR-compliant environment for exploring the application of generative AI in education. The platform offers an interface similar to other large language models while operating within a controlled institutional context.

The EduGenAI platform was chosen for the pilot based on its availability within the national Npuls initiative, its compliance with legal requirements such as the GDPR, and its suitability for use in higher education settings.

As part of the pilot, the **AI Expert Team** at THUAS developed three **personas** for three domains. A persona in this context is a role-specific configuration of the AI, pre-loaded with instructions that reflect institutional priorities and constraints. Each persona guided the AI to respond in ways relevant to the daily work of staff in that domain. For example, the Education persona included prompts about lesson design, learning outcomes, and feedback for students; the Research persona incorporated prompts about literature reviews, abstracts, and questionnaires; and the Support persona embedded prompts for meeting summaries, communication, and policy texts. These personas also encoded THUAS-specific compliance requirements and ethical boundaries, ensuring outputs aligned with institutional policy and values. The personas were introduced to lower the entry barrier for staff.

2.3 Sequence of Activities

The pilot followed a structured sequence from preparation to reporting (Figure 2). The preparation phase (April–May 2025) included the literature review and the development of the ToC. Participants could register from 15 April 2025, and the pilot formally began with a kick-off on 15 May. The kick-off session provided information about the pilot, after which informed consent was collected and the first survey (Pre-Pilot TAM) was distributed on 16 May. From 15 May to 15 June, participants had access to the EduGenAI platform. The second survey (Post-Pilot TAM) was sent on 16 June, directly after the usage period. On 19 June, a micro-RCT was conducted across all domains. Following the RCT, semi-structured interviews were carried out with staff from different domains and levels of digital literacy, including some who had dropped out. The analysis and reporting phase took place between July and September 2025.

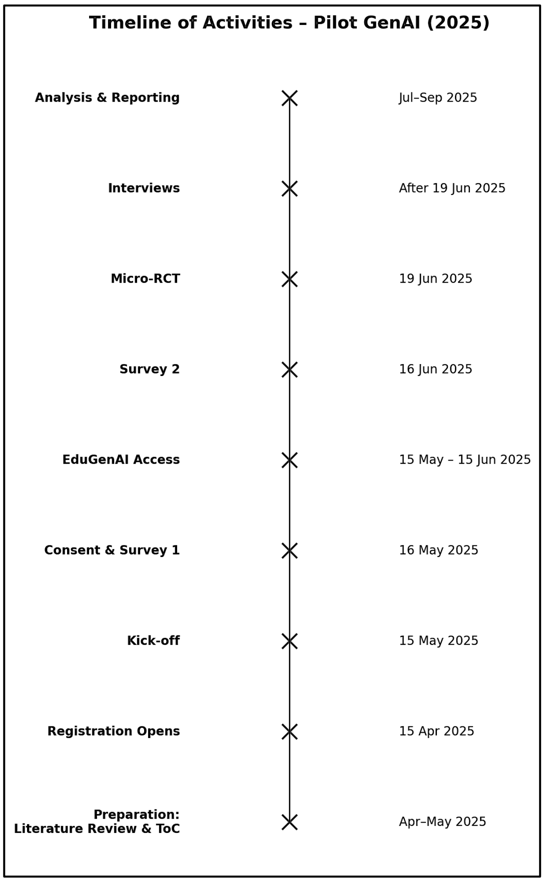


Figure 2. Timeline of Activities

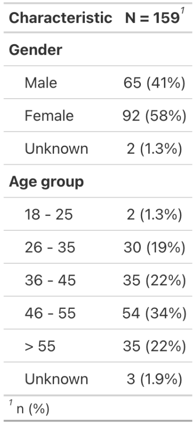
2.4 Participants

The participants in this study were employees of THUAS working in the domains of education, research, and support. Inclusion in the study was restricted to employees who had successfully completed the *HCTL AI Basic Course*, signed the informed consent form, and obtained managerial approval to participate.

In total, 205 employees registered for the pilot, but only 159 participants met the inclusion criteria. This group formed the core sample for the study.

The participants represented a broad cross-section of THUAS staff: 63 from education, 23 from research, and 73 from support. The demographic information (gender and age) is presented in Table 1.

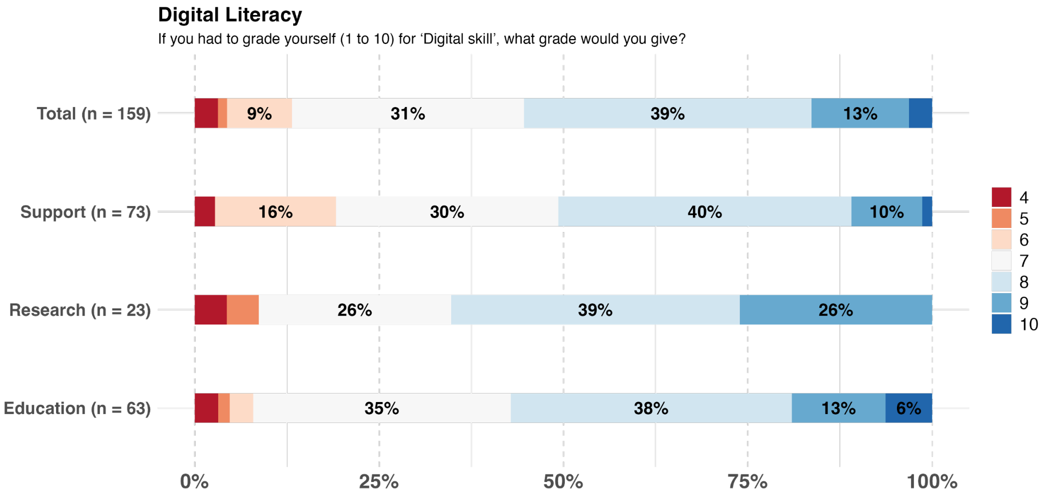
Table 1. Demographic characteristics of participants



Characteristic	N = 159[1]
Gender	
Male	65 (41%)
Female	92 (58%)
Unknown	2 (1.3%)
Age group	
18 - 25	2 (1.3%)
26 - 35	30 (19%)
36 - 45	35 (22%)
46 - 55	54 (34%)
> 55	35 (22%)
Unknown	3 (1.9%)

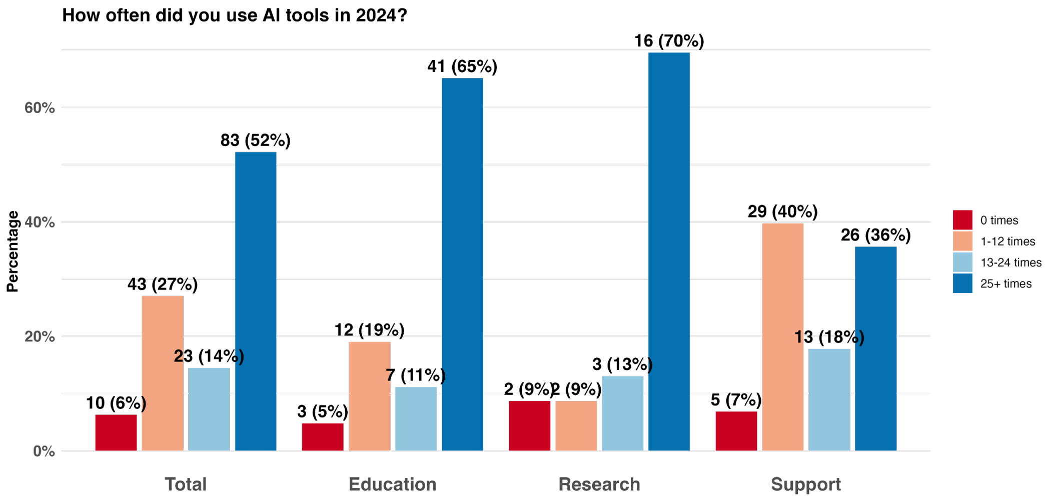
[1] n (%)

Levels of digital literacy were generally high (Figure 3), with most respondents rating themselves (in the registration form) between 7 and 9 on a ten-point scale. Prior experience with AI tools varied (Figure 4), but more than half (52%) reported using AI tools at least 25 times in 2024, indicating a strong baseline of familiarity.



© De HHs, Lectoraat Learning Technology & Analytics

Figure 3. Self-reported digital literacy of participants



© De HHs, Lectoraat Learning Technology & Analytics

Figure 4. Self-reported prior experience with AI tools in 2024

Participation varied across the different phases of the research, as shown in Table 2.

Table 2. Participation across phases of the Pilot GenAI

Phase	Number of Participants
Registered for the pilot	205
Completed informed consent and baseline survey (Pre-Pilot TAM)	159
Participated in RCT	67 (2 excluded: no consent)
Completed follow-up survey (Post-Pilot TAM)	106 (2 excluded: no consent)
Interviewed	10

2.5 Ethical Considerations

Participation in the study was voluntary and based on informed consent. All participants received information about the aims, methods, and data to be collected, and were assured that they could withdraw at any time without consequences. Data collection and storage complied with GDPR regulations, anonymized reporting was applied, and all data will be retained securely for ten years in line with the Dutch Code of Conduct for Scientific Practice. Ethical oversight for the pilot was provided through the internal review procedures of THUAS.

3 Results

The results are presented following the three levels of evidence in the 3E Framework: bronze (literature review and ToC), silver (surveys and interviews), and gold (RCT). Within each level, findings are reported for the three dimensions of usability, productivity, and quality.

3.1 Bronze-level Evidence

The Bronze level of the 3E Framework establishes the theoretical foundation for the pilot. This level included a literature review (Section 3.1.1), which synthesized international and national studies on GenAI in education, and the development of a ToC (Section 3.1.2), which mapped how GenAI might influence staff practices at THUAS. Together, these elements provided the hypotheses and assumptions that guided the empirical research in later stages.

3.1.1 Literature Review

The literature review highlights that while GenAI is increasingly studied in higher education, most findings focus on education and research staff, with support roles notably underrepresented. Despite this, administrative tasks are frequently discussed in a broader sense under different domains.

Usability. GenAI is generally seen as useful and easy to adopt, especially as users gain experience. Automation of routine tasks, content generation, and writing assistance were the most highlighted use cases.

Quality. GenAI is reported to improve output quality across roles, but concerns such as misinformation and academic integrity were frequent. Human oversight remains essential to maintain standards.

Productivity. GenAI is reported to consistently boost productivity by automating routine tasks and freeing time for more strategic and creative work across all staff domains.

This literature review examines recent studies on the impact of GenAI on three outcomes in higher education: perceived usability, quality, and work productivity, across three staff domains - education, research, and support staff. The review includes English-language studies published since 2023, conducted in Europe, and focused on GenAI use in higher education. Out of 195 identified articles, 15 are included after screening. The reviewed literature predominantly focuses on teaching and research staff, with support staff being notably underrepresented. Only one study [10] focuses exclusively on administration and educational support tasks. However, tasks associated with support staff (e.g., administrative tasks) are also often discussed concerning the other two domains.

Across all staff domains, **perceived usefulness and ease-of-use** are critical factors influencing the adoption of GenAI. Included studies focus on how these perceptions affect the adoption of GenAI among the three domains, and how the perceptions change through continued use. According to the TAM surveys, perceived usefulness and ease-of-use play a critical role in shaping future behavioral intentions of use. For teaching staff, perceived usefulness is a strong positive determinant of adoption. Educators report that GenAI helps them accomplish tasks more quickly, improves job performance, and enhances their overall effectiveness in administrative tasks and as a resource for self-learning and intellectual engagement [11–13]. Moreover, as they become more familiar with GenAI, they perceive it as easier to use and more useful [14]. Similar observations can be made among research staff, where perceived usefulness and usability of GenAI are critical drivers of adoption. At the same time, it is noted that experience with GenAI promotes self-efficacy and stronger affirmation of its value [14]. Interestingly, experience with GenAI also exposes its limitations, such as hallucinated citations and unreliable outputs. This often reshapes researchers' perception of its usability [15,16]. For support staff, perceived usefulness is strongly centered on the anticipated efficiency of reducing administrative workload. It was also observed that support staff, compared to teaching and research staff, report significantly more positive views on AI affordances and generally express higher AI self-efficacy[13].

GenAI's impact on **quality of work** is described with more caution because of its potential to improve output and the risks it could pose to integrity and foundational skills. For teaching staff, GenAI can enhance the quality of assessment processes by providing immediate, detailed, and sound feedback, which helps them maintain consistent assessment standards [17,18]. At the same time, GenAI threatens the integrity of assessment practices, requiring teachers to adapt their methods and potentially leading to increased workload for verification [19,20]. For research staff, GenAI can enhance the quality of outputs by improving text clarity and grammar, as well as providing other kinds of editorial support [13,18,21]. However, one notable limitation is the inaccuracy of outputs (hallucinations), particularly regarding references and biographical information. This creates a risk of disinformation and a potential decline in research quality due to "plausible yet flawed results" [13,15,18]. Finally, work quality of support staff is improved by enabling staff to allocate time to higher-order, strategic, and creative activities once GenAI automates repetitive tasks. Furthermore, GenAI is reported to enhance the quality of linguistic and administrative outputs[13,18,20]. Across all three domains, quality maintenance depends on ensuring that humans remain accountable for all GenAI-produced content.

Lastly, improved **productivity** is consistently linked to GenAI largely due to its ability to automate routine tasks, streamline intellectual workflows, and free up time for higher-order thinking. The productivity of teaching staff was enhanced by automating repetitive tasks such as writing announcements, generating assessment components, and managing language tasks (e.g., proofreading text, checking for grammar errors, etc.) [11,19,22]. Similarly, research staff benefits from AI-assisted drafting and technical support, thus streamlining processes that traditionally require significant time and effort [14,15,21]. It also boosts productivity by enhancing the quality of outputs, facilitating idea generation and code interpretation [16,23]. For support staff, GenAI improves productivity by streamlining workflows and automating repetitive tasks such as drafting emails, reports, and process documentation [10,13]. These shifts increase the volume of work that can be completed within the same time frame and enhance staff capacity to contribute to broader institutional goals through more thoughtful and impactful work.

Overall, the studies suggest that GenAI is perceived as a valuable tool for enhancing efficiency, improving output quality, and supporting higher-order work across staff roles. It is observed that the adoption and impact of GenAI are shaped by factors such as role-specific needs, levels of experience, and concerns about reliability.

3.1.2 Theory Of Change

A ToC (Figure 5) is designed for this study based on the literature review. It hypothesizes the mechanisms through which GenAI is expected to influence the usability, quality of work, and productivity of THUAS staff.

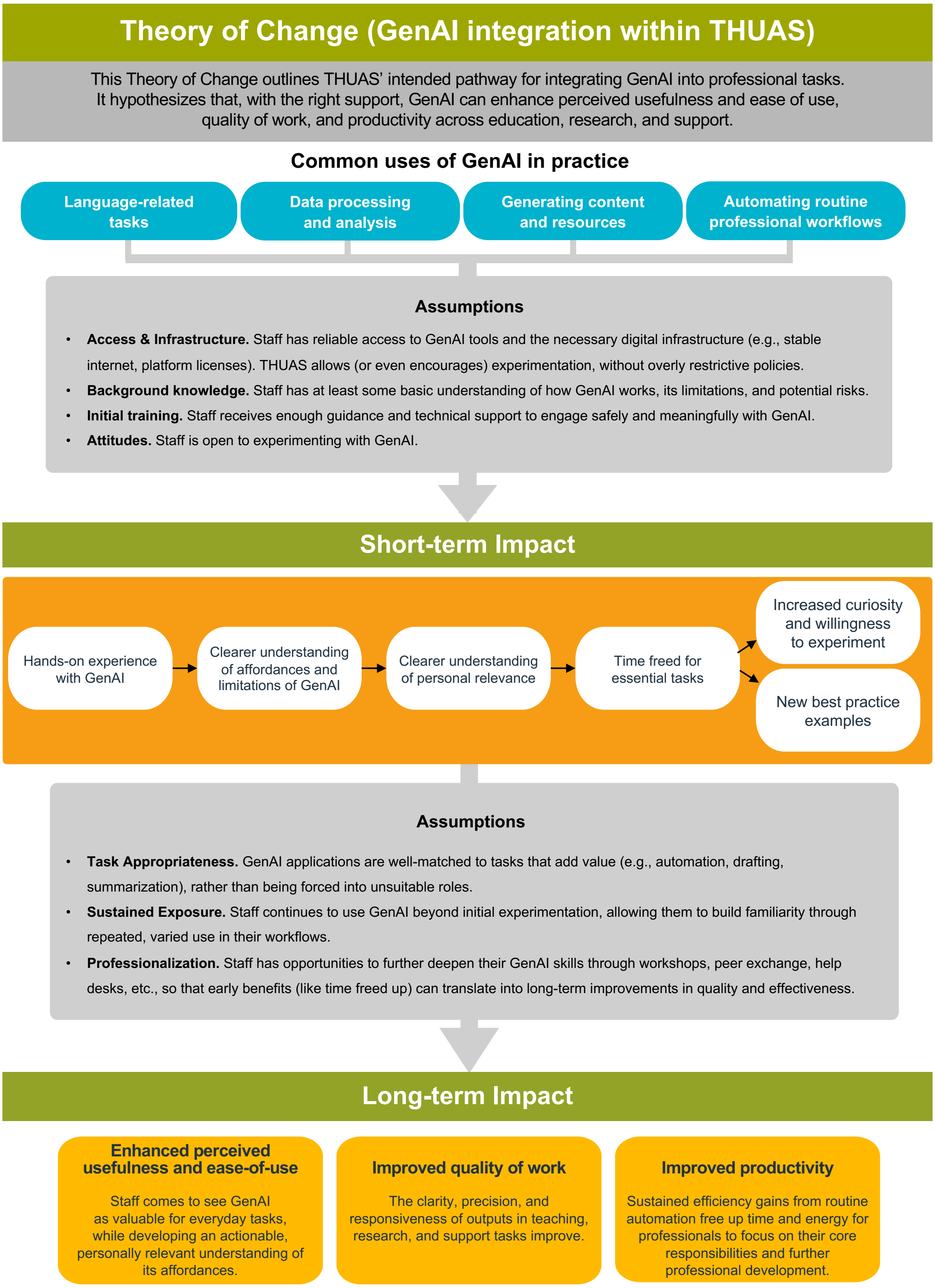
The central hypothesis of this study is that, given access to GenAI, staff will gradually embed it into their daily routines when supported by training, clear guidance, and a culture that allows experimentation without fear of judgment. This process is expected to improve usability (by making GenAI more intuitive and relevant), enhance the quality of outputs (through more structured and polished drafts), and increase productivity (by reducing time spent on repetitive tasks). However, the extent of these effects will likely differ between domains depending on task characteristics, prior experience, and digital maturity.

A ToC is a structured framework that helps clarify how specific activities are expected to lead to desired outcomes, under certain assumptions [24]. A ToC is designed for this study based on the literature review described in the previous section. It hypothesizes how the GenAI adoption by education, research, and support staff at THUAS may contribute to short-term and long-term improvements in perceived usefulness, quality of work outputs, and productivity.

The process begins with embedding GenAI into staff members' routine professional tasks (such as language-related tasks, data analysis, and content creation). For example, in teaching, it could assist with instructional design, grading, and feedback generation. In research, it could support data preparation, literature summarization, and drafting research papers. For support staff, it could streamline scheduling calls, writing correspondence, documenting meeting minutes, etc.

With GenAI embedded in practice, THUAS staff are expected to become more open to experimentation, identify opportunities to improve the quality of their work, and contribute to emerging examples of effective use. This initial exposure could help staff develop a clearer understanding of GenAI's capabilities, general limitations, and, more importantly, its relevance to their work. This could make delegating lower effort and routine tasks to GenAI easier, thus reducing cognitive and operational workload, which would lead to more time for handling their core responsibilities. The short-term impact is expected to translate to long-term impact (improved usability, quality, and productivity) with continued exposure.

There are several assumptions underlying the hypothesized ToC. Firstly, THUAS needs to provide continued access to GenAI tools and support infrastructure, basic AI training, and workshops. Moreover, an encouraging culture of trust, experimentation, and appropriate task alignment is needed to realize the actual benefits of AI. These conditions also reflect the external variables described in the TAM as influencing users' perceptions of usability [6]. Lastly, the staff needs to be open to experimentation with new technology, learn to use it, and share the attained knowledge with colleagues.



Theory of Change (GenAI integration within THUAS)

This Theory of Change outlines THUAS' intended pathway for integrating GenAI into professional tasks. It hypothesizes that, with the right support, GenAI can enhance perceived usefulness and ease of use, quality of work, and productivity across education, research, and support.

Common uses of GenAI in practice

- Language-related tasks
- Data processing and analysis
- Generating content and resources
- Automating routine professional workflows

Assumptions

- **Access & Infrastructure.** Staff has reliable access to GenAI tools and the necessary digital infrastructure (e.g., stable internet, platform licenses). THUAS allows (or even encourages) experimentation, without overly restrictive policies.
- **Background knowledge.** Staff has at least some basic understanding of how GenAI works, its limitations, and potential risks.
- **Initial training.** Staff receives enough guidance and technical support to engage safely and meaningfully with GenAI.
- **Attitudes.** Staff is open to experimenting with GenAI.

Short-term Impact

Hands-on experience with GenAI → Clearer understanding of affordances and limitations of GenAI → Clearer understanding of personal relevance → Time freed for essential tasks → Increased curiosity and willingness to experiment; New best practice examples

Assumptions

- **Task Appropriateness.** GenAI applications are well-matched to tasks that add value (e.g., automation, drafting, summarization), rather than being forced into unsuitable roles.
- **Sustained Exposure.** Staff continues to use GenAI beyond initial experimentation, allowing them to build familiarity through repeated, varied use in their workflows.
- **Professionalization.** Staff has opportunities to further deepen their GenAI skills through workshops, peer exchange, help desks, etc., so that early benefits (like time freed up) can translate into long-term improvements in quality and effectiveness.

Long-term Impact

Enhanced perceived usefulness and ease-of-use

Staff comes to see GenAI as valuable for everyday tasks, while developing an actionable, personally relevant understanding of its affordances.

Improved quality of work

The clarity, precision, and responsiveness of outputs in teaching, research, and support tasks improve.

Improved productivity

Sustained efficiency gains from routine automation free up time and energy for professionals to focus on their core responsibilities and further professional development.

Figure 5. Theory of Change for the Pilot GenAI

3.2 Silver-level Evidence

This section presents results from the silver level of the 3E Framework, which focused on staff perceptions and experiences during the pilot. Evidence was collected through TAM surveys administered before and after the pilot, complemented by interviews. These methods provide insight into usability, quality, and productivity across domains. The findings of TAM surveys are presented in Section 3.2.1 and for interviews in Section 3.2.2.

3.2.1 Surveys

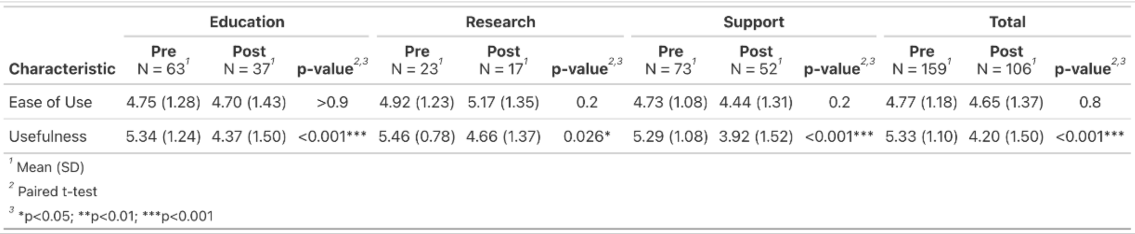
The TAM surveys were conducted online and offered a quantitative view of staff attitudes. The pre-pilot TAM captured participants' expectations, while the post-pilot TAM measured their experiences after four weeks of access to the EduGenAI platform.

Usability

> The overall usability was experienced positively across the institution. Across all domains, perceived ease of use remained stable, confirming that GenAI was not experienced as technically difficult to use. However, perceived usefulness declined significantly, especially in education and support, indicating that staff became more critical once they tested the tool in practice.

TAM is most often used to assess usability or adoption of new technology, through measures of perceived ease of use and perceived usefulness. The items of the survey are measured on the scale of 7. The domain-specific and aggregated results are presented in Table 3.

Table 3. Perceived ease of use (PEoU) and usefulness (PU) before and after the pilot



	Education			Research			Support			Total		
Characteristic	**Pre** N = 63[1]	**Post** N = 37[1]	**p-value**[2,3]	**Pre** N = 23[1]	**Post** N = 17[1]	**p-value**[2,3]	**Pre** N = 73[1]	**Post** N = 52[1]	**p-value**[2,3]	**Pre** N = 159[1]	**Post** N = 106[1]	**p-value**[2,3]
Ease of Use	4.75 (1.28)	4.70 (1.43)	>0.9	4.92 (1.23)	5.17 (1.35)	0.2	4.73 (1.08)	4.44 (1.31)	0.2	4.77 (1.18)	4.65 (1.37)	0.8
Usefulness	5.34 (1.24)	4.37 (1.50)	<0.001***	5.46 (0.78)	4.66 (1.37)	0.026*	5.29 (1.08)	3.92 (1.52)	<0.001***	5.33 (1.10)	4.20 (1.50)	<0.001***

[1] Mean (SD)

[2] Paired t-test

[3] *p<0.05; **p<0.01; ***p<0.001

The mean PEoU score before the pilot was 4.77 (SD = 1.18) and after the pilot 4.65 (SD = 1.37). The difference was not statistically significant, indicating that staff expectations of usability were largely confirmed by their experiences. By contrast, PU declined significantly, from 5.33 (SD = 1.10) in the pre-pilot phase to 4.20 (SD = 1.50) in the post-pilot phase (*p* < .001). While usefulness was still rated relatively high, this shift shows that staff became more critical after practical exposure to GenAI. The domain-level analysis reveals differences:

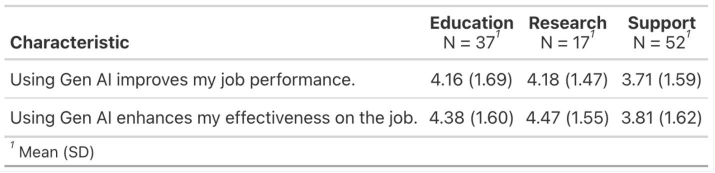
- **Teachers** reported stable ease of use (Pre = 4.75; Post = 4.70; *p* > .900), but a significant drop in usefulness (Pre = 5.34; Post = 4.37; *p* < .001).
- **Researchers** showed slightly higher ease of use (Pre = 4.92; Post = 5.17; *p* = .200) and a modest but significant decline in usefulness (Pre = 5.46; Post = 4.66; *p* = .026).
- **Support staff** showed more divided results. Ease of use declined slightly (Pre = 4.73; Post = 4.44; *p* = .200), while usefulness fell sharply (Pre = 5.29; Post = 3.92; *p* < .001).

Quality

> The quality of work was generally perceived as moderate across domains. Teachers and researchers reported moderate to positive effects of GenAI on quality, while support staff gave lower scores.

The post-pilot TAM survey asked participants to rate their agreement with two statements related to quality (performance and effectiveness) as shown in Table 4. These self-reported perceptions provide a domain-level perspective on how GenAI influenced the quality of work.

Table 4. Domain-specific quality perceptions.



Characteristic	Education N = 37[1]	Research N = 17[1]	Support N = 52[1]
Using Gen AI improves my job performance.	4.16 (1.69)	4.18 (1.47)	3.71 (1.59)
Using Gen AI enhances my effectiveness on the job.	4.38 (1.60)	4.47 (1.55)	3.81 (1.62)

[1] Mean (SD)

Across all domains, mean scores for quality ranged from 3.71 to 4.47, indicating moderate perceptions overall. Domain-level analysis reveals differences:

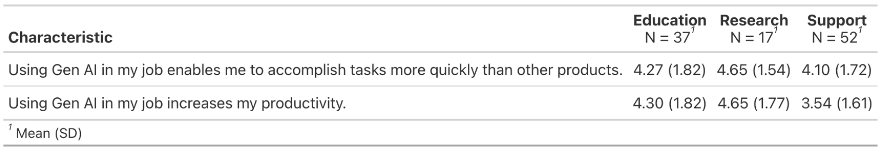
- **Teachers** reported moderate to positive scores (4.16 and 4.38).
- **Researchers** reported the highest values across domains, averaging 4.18 for performance and 4.47 for effectiveness.
- **Support staff** gave lower ratings, with mean scores of 3.71 for performance and 3.81 for effectiveness.

Productivity

> The productivity was generally perceived as somewhat improved but not universally high. Researchers consistently reported the strongest productivity benefits, teachers indicated moderate improvements, and support staff reported the least impact.

Productivity was assessed through two TAM items on speed and overall productivity (Table 5).

Table 5. TAM-based productivity perceptions by domain.



Characteristic	Education N = 37[1]	Research N = 17[1]	Support N = 52[1]
Using Gen AI in my job enables me to accomplish tasks more quickly than other products.	4.27 (1.82)	4.65 (1.54)	4.10 (1.72)
Using Gen AI in my job increases my productivity.	4.30 (1.82)	4.65 (1.77)	3.54 (1.61)

[1] Mean (SD)

Across all domains, mean scores for productivity ranged from 3.54 to 4.65. Domain-level analysis reveals differences:

- **Teachers** reported mean scores of 4.27 for speed and 4.30 for productivity.
- **Researchers** provided the highest ratings, with both items averaging 4.65.
- **Support staff** scored 4.10 for speed and 3.54 for productivity, the lowest among all domains.

3.2.2 Interviews

The interviews provided qualitative insights into how staff experienced GenAI in practice. Unlike the surveys, which captured broad patterns, the interviews offered detailed accounts of staff reflections on usability, productivity, and quality in their daily work.

Usability

Perceptions of usability were shaped by digital literacy, task appropriateness and prior experience with GenAI. Teachers and researchers found clear applications in writing-focused tasks, while support staff often struggled to connect the tool to their diverse and context-dependent workflows.

Usability experiences were closely tied to digital literacy and prior exposure to GenAI. Staff with limited digital skills often needed additional guidance to formulate prompts and interpret outputs, whereas confident users experimented independently. First-time users described curiosity mixed with hesitation, while experienced users often compared EduGenAI to tools like ChatGPT, treating it as an extension of their established routines.

"I had never used GenAI before, but I was curious and open to experimenting."
(Interview, Support domain)

"I've been using ChatGPT almost daily, especially for checking the structure and clarity of texts."
(Interview, Research domain)

Teachers highlighted its value for lesson design, learning outcomes, and idea generation, though some noted outputs were too generic and lacked sensitivity to student contexts.

"It's a helpful brainstorm partner but doesn't replace my professional judgment."
(Interview, Education domain)

Researchers described the platform as easy to use for summarization, structuring, and refining drafts. Their main concerns related to the reliability of outputs and fabricated references.

"It's like having a language coach. I use it to fine-tune, not to replace my writing."
(Interview, Research domain)

Support staff showed the widest variation. They found it useful for many routine tasks but struggled to see relevance for their context-specific tasks, especially where personal data was involved. Uncertainty about what was "allowed" also limited exploration.

"It's usable, but I didn't know what to ask it. My job is too context-specific sometimes."
(Interview, Support domain)

Quality

Quality gains were observed across domains but remained conditional. Teachers and researchers benefited most from structured drafts, while support staff experienced uneven outcomes tied to task type and data quality.

Staff linked quality gains to GenAI's ability to provide structured drafts and clarity in outputs, but also stressed the need for professional review.

Teachers frequently highlighted GenAI's usefulness in producing structured drafts but stressed that these outputs required refinement to ensure subject-specific accuracy.

"If you don't know the topic, you can be satisfied too soon. You need to refine your questions."
(Interview, Education domain)

Researchers consistently noted improvements in early drafts, questionnaires, and academic writing. At the same time, they emphasized the importance of verifying content for accuracy and correct referencing.

"It really helped me with the first concept of the questionnaire… it gave me a head start."
(Interview, Research domain)

Support staff reported mixed results. Some appreciated clearer summaries, while others stressed that output quality depended heavily on input quality. Tasks requiring nuanced contextual knowledge were harder to support effectively.

"Make an excerpt of this text… it works. But if the data under the AI is weak, garbage in is garbage out."
(Interview, Support domain)

Productivity

Researchers most strongly reported productivity gains, followed by teachers, while support staff showed the most mixed experiences, balancing time savings against the effort of learning and adapting to new workflows.

Interview responses highlighted that GenAI could save time in routine or repetitive tasks, though the extent of perceived gains varied across domains.

Teachers noted that GenAI reduced preparation time and helped structure lesson materials quickly, especially when tasks were repeated. Still, they underlined that outputs required adaptation to fit pedagogical aims.

"When I had to do something for the third or fourth time, it helped me reframe it quickly."
(Interview, Education domain)

Researchers described significant time savings in academic writing, questionnaire design, and coding support. For some, GenAI accelerated tasks that would typically take weeks.

"Usually [research tasks] take me a month… with GenAI, it took me a week easily."
(Interview, Research domain)

Support staff were more divided. Some described clear time savings in professional work such as writing meeting summaries, while others emphasized that the tool demanded extra effort to learn new workflows, frame prompts, and refine outputs.

"A few minutes later, I had my meeting summarized… I thought, yeah, this saves me a lot of time."
(Interview, Support domain)

“Oh no, me with GenAI? I’m not skilled enough… I did all the blocks separately — I should have combined them in one good prompt.”
(Interview, Support domain)

3.3 Gold-level Evidence

The gold-level evidence provides causal insights into GenAI's effects on staff work. This evidence was generated through a micro-RCT. The next sub-section presents the findings of the micro-RCT in detail.

3.3.1 RCT

The RCT was conducted to measure how GenAI affected quality and productivity when staff performed domain-specific tasks.

Quality

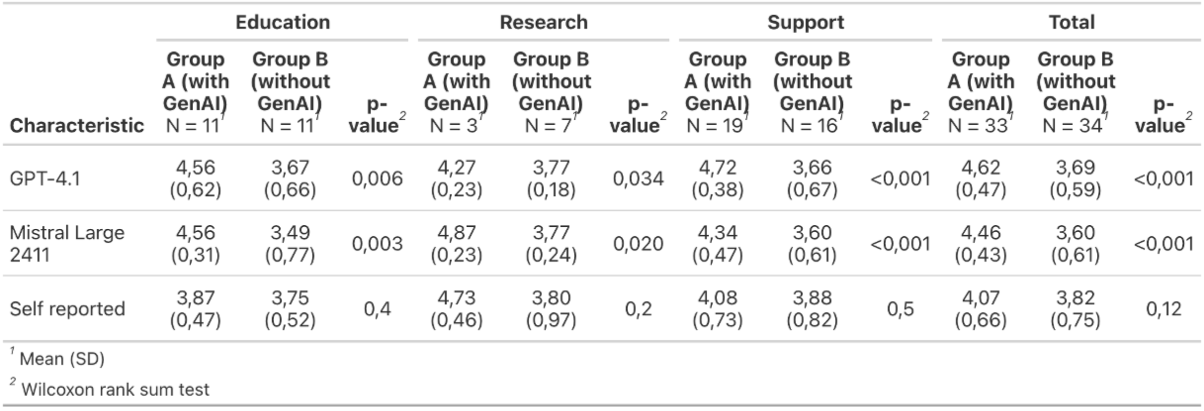
The quality of outputs was assessed in two ways:

1. GPT-evaluated quality (LLM: GPT-4.1 and mistral-large-2411), and
2. Self-reported quality by participants.

Overall, across all domains, GPT-4.1 consistently rated GenAI outputs as higher quality (M = 4.62 vs. 3.69; *p* < .001), while self-reported scores showed no significant difference (M = 4.07 vs. 3.82; *p* = .12).

The domain-specific and aggregated results are presented in Table 6.

Table 6. Quality scores across measures (GPT-based evaluations and self-reports)



	Education			Research			Support			Total		
Characteristic	**Group A (with GenAI)** N = 11[1]	**Group B (without GenAI)** N = 11[1]	**p-value**[2]	**Group A (with GenAI)** N = 3[1]	**Group B (without GenAI)** N = 7[1]	**p-value**[2]	**Group A (with GenAI)** N = 19[1]	**Group B (without GenAI)** N = 16[1]	**p-value**[2]	**Group A (with GenAI)** N = 33[1]	**Group B (without GenAI)** N = 34[1]	**p-value**[2]
GPT-4.1	4,56 (0,62)	3,67 (0,66)	0,006	4,27 (0,23)	3,77 (0,18)	0,034	4,72 (0,38)	3,66 (0,67)	<0,001	4,62 (0,47)	3,69 (0,59)	<0,001
Mistral Large 2411	4,56 (0,31)	3,49 (0,77)	0,003	4,87 (0,23)	3,77 (0,24)	0,020	4,34 (0,47)	3,60 (0,61)	<0,001	4,46 (0,43)	3,60 (0,61)	<0,001
Self reported	3,87 (0,47)	3,75 (0,52)	0,4	4,73 (0,46)	3,80 (0,97)	0,2	4,08 (0,73)	3,88 (0,82)	0,5	4,07 (0,66)	3,82 (0,75)	0,12

[1] Mean (SD)

[2] Wilcoxon rank sum test

1. **GPT-evaluated quality.** According to GPT-4.1 (Figure 6), outputs from Group A (with GenAI) achieved significantly higher scores (M = 4.62, SD = 0.47) compared to Group B (without GenAI) (M = 3.68, SD = 0.57; *p* < .001). Mistral-large produced similar results, with Group A (M = 4.46, SD = 0.43) scoring higher than Group B (M = 3.60, SD = 0.59; *p* < .001). Inter-rater reliability between GPT-4.1 and mistral-large was good (ICC = 0.807), showing strong consistency between the two LLMs.

   As GPT-4.1 is benchmarked as the more advanced model, its results were used as the main reference for further analysis.

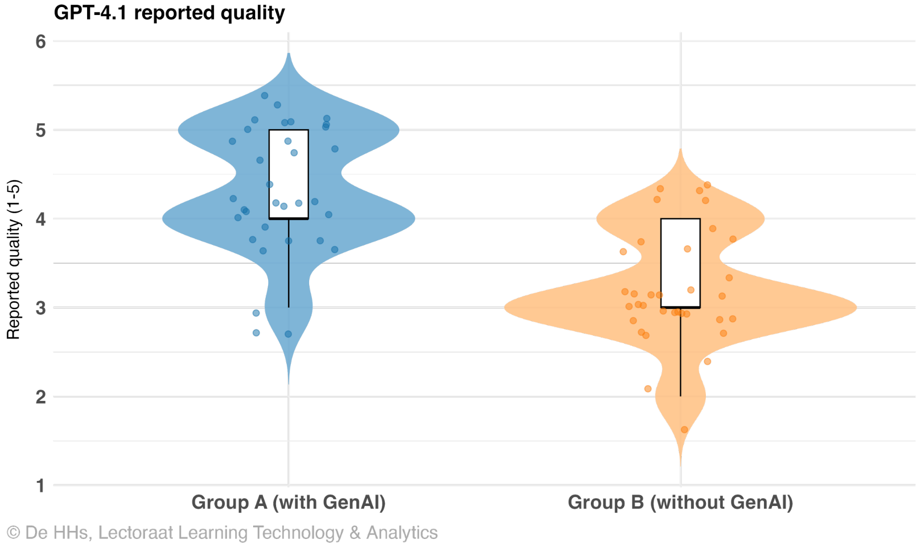


Figure 6. GPT-4.1 evaluated quality scores across groups.

2. **Self-reported quality.** Participants' own ratings did not show a significant difference between groups. Group A reported M = 4.07 (SD = 0.66), while Group B reported M = 3.84 (SD = 0.76; *p* = .17).

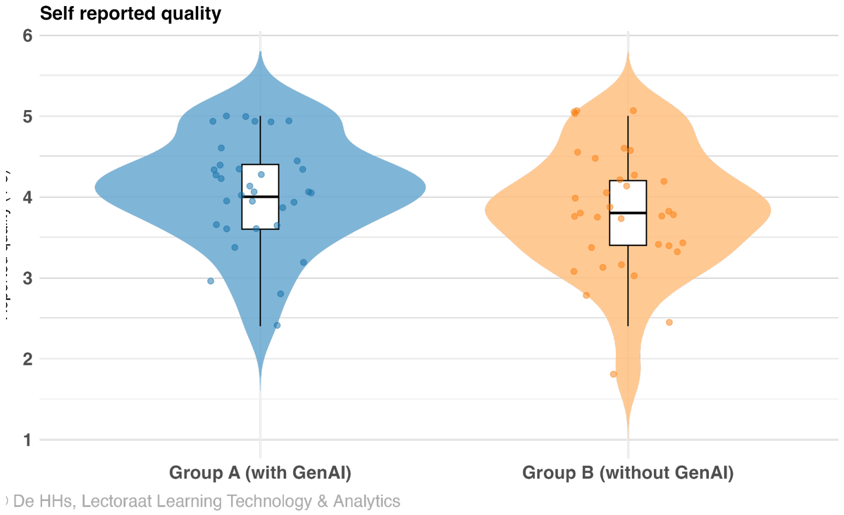


Figure 7. Self-reported quality scores across groups.

Across both LLMs, tasks completed with GenAI were rated significantly higher in quality than those completed without it. However, this improvement was not reflected in participants' self-reports.

The RCT results show clear domain-specific patterns in quality outcomes.

- In **education**, GenAI-supported outputs were rated significantly higher by GPT-4.1 (M = 4.56 vs. 3.67; *p* = .006) and Mistral-large (M = 4.56 vs. 3.49; *p* = .003), though self-reports showed no significant difference.

- In **research**, GenAI outputs again received higher scores (GPT-4.1: M = 4.27 vs. 3.77; *p* = .034; Mistral-large: M = 4.87 vs. 3.77; *p* = .020), while self-assessments did not differ significantly.
- The largest differences appeared in **support**, where GPT-4.1 rated GenAI outputs much higher (M = 4.72 vs. 3.66; *p* < .001), confirmed by Mistral-large (M = 4.34 vs. 3.60; *p* < .001), but again, self-reports showed no significant change.

Agreement between human and GPT. The agreement between participants' self-evaluations and GPT-4.1 scores was low (ICC = 0.184, not significant). Several factors may explain this divergence. *First*, the rubric used in the RCT was described in broad terms, leaving room for subjective interpretation. *Second*, self-assessments are often influenced by personal bias, as some participants underrate their own work while others are more generous. By contrast, the LLMs consistently focused on structure, clarity, and coherence.

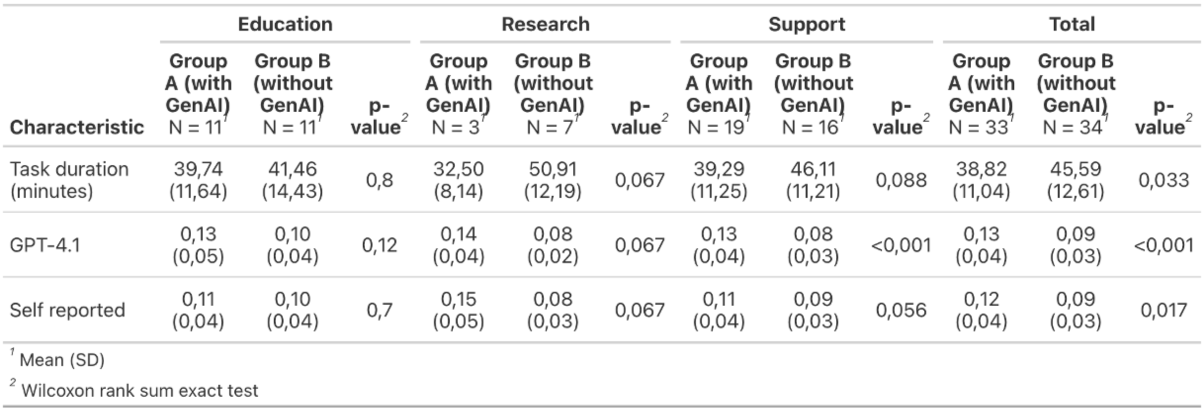
> Human self-assessments and automated evaluations measured different quality aspects and did not align closely, underlining the need to consider both perspectives.

Productivity

> The productivity of participants' outputs was assessed in two ways:
>
> 1. **Task duration**, measuring the time required to complete a task.
>
> 2. **Productivity scores**, calculated as: $\frac{\textit{Quality score}}{\textit{Task duration}}$
>
> Productivity gains were observed when staff used GenAI. Tasks were completed faster, and when quality was taken into account, productivity scores were significantly higher across both self- and GPT-based measures.

The domain-specific and aggregated results are presented in Table 7.

Table 7. Productivity scores across measures



	Education			Research			Support			Total		
Characteristic	**Group A (with GenAI)** N = 11[1]	**Group B (without GenAI)** N = 11[1]	**p-value**[2]	**Group A (with GenAI)** N = 3[1]	**Group B (without GenAI)** N = 7[1]	**p-value**[2]	**Group A (with GenAI)** N = 19[1]	**Group B (without GenAI)** N = 16[1]	**p-value**[2]	**Group A (with GenAI)** N = 33[1]	**Group B (without GenAI)** N = 34[1]	**p-value**[2]
Task duration (minutes)	39,74 (11,64)	41,46 (14,43)	0,8	32,50 (8,14)	50,91 (12,19)	0,067	39,29 (11,25)	46,11 (11,21)	0,088	38,82 (11,04)	45,59 (12,61)	0,033
GPT-4.1	0,13 (0,05)	0,10 (0,04)	0,12	0,14 (0,04)	0,08 (0,02)	0,067	0,13 (0,04)	0,08 (0,03)	<0,001	0,13 (0,04)	0,09 (0,03)	<0,001
Self reported	0,11 (0,04)	0,10 (0,04)	0,7	0,15 (0,05)	0,08 (0,03)	0,067	0,11 (0,04)	0,09 (0,03)	0,056	0,12 (0,04)	0,09 (0,03)	0,017

[1] Mean (SD)

[2] Wilcoxon rank sum exact test

Task duration. As shown in Figure 8, participants in the GenAI group (Group A) completed tasks in significantly less time (M = 38.8 min, SD = 11.04) compared to the control group without GenAI (Group B, M = 45.9 min, SD = 12.61; *p* = .033). This confirms that GenAI reduced the time needed to complete tasks.

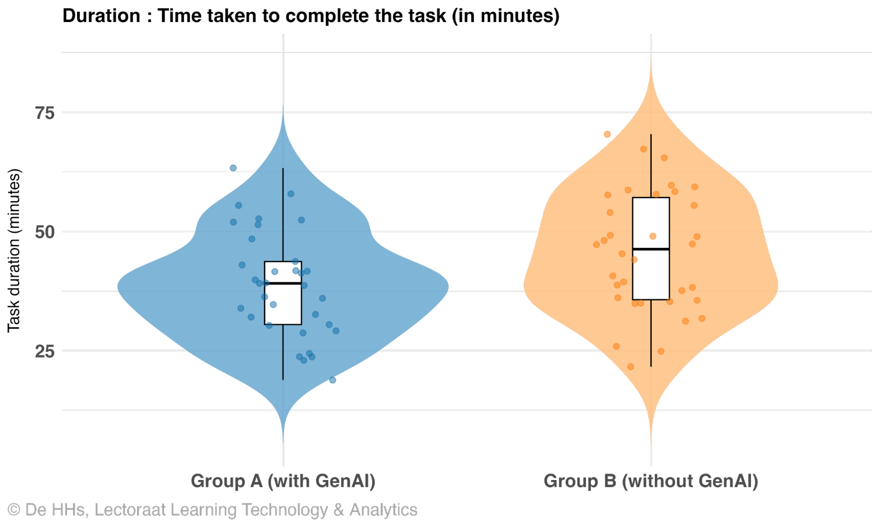


Figure 8. Task duration (minutes) across groups

Productivity scores. Task duration alone does not fully capture productivity, as quality also matters. Therefore, productivity was calculated using both GPT-evaluated quality and self-reported quality.

- **GPT-evaluated productivity.** Calculated as GPT-4.1 quality score divided by task duration. Group A again scored significantly higher than Group B (p < .001), confirming from an independent benchmark that GenAI increased productivity (Figure 9).

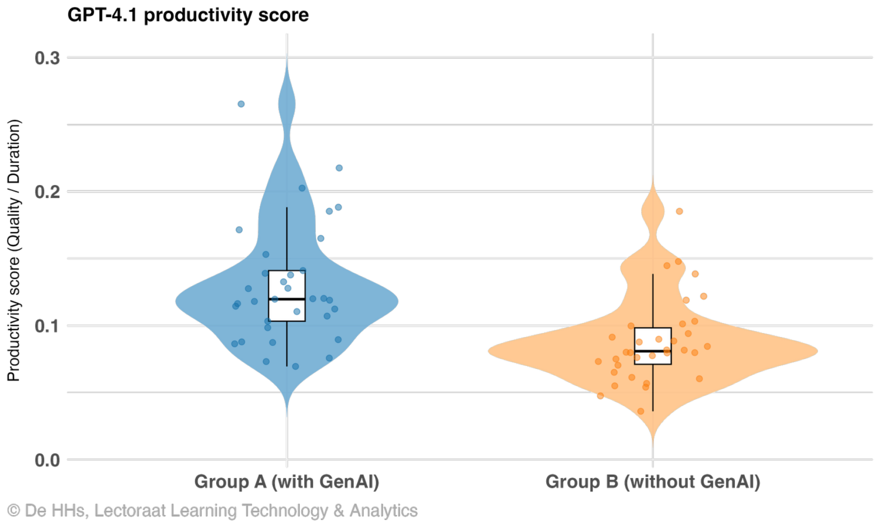
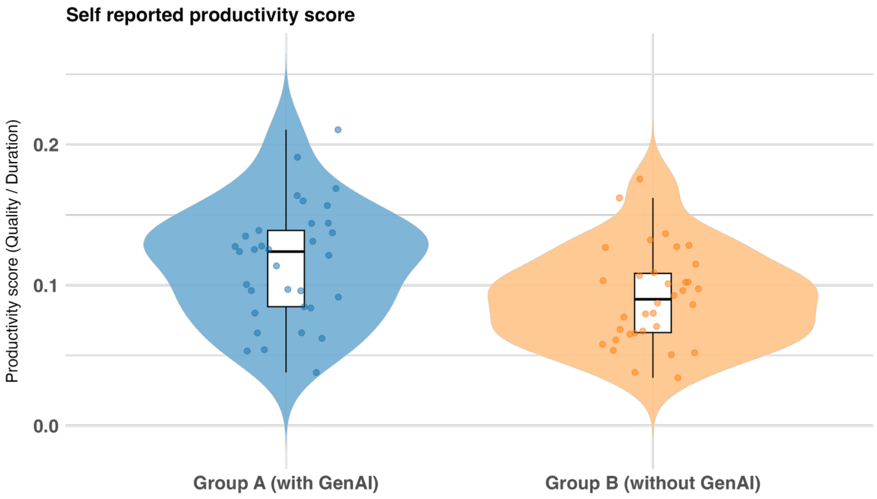


Figure 9. GPT-evaluated productivity scores across groups.

- **Self-reported productivity.** Calculated as self-reported quality divided by task duration. Group A reported higher productivity than Group B (p = .017). Figure 10 shows that participants felt more productive when using GenAI.



© De HHs, Lectoraat Learning Technology & Analytics

Figure 10. Self-evaluated productivity scores across groups.

Domain-specific results.

- **Teachers** in Group A completed tasks slightly faster (M = 39.7 min, SD = 11.6) than Group B (M = 41.5 min, SD = 14.4), though the difference was not significant (*p* = .800). GPT-productivity was higher in the GenAI group (M = 0.13, SD = 0.05 vs. 0.10, SD = 0.04; *p* = .120), but again not significant.
- **Researchers** achieved the strongest gains. Group A completed tasks much faster (M = 32.5 min, SD = 8.1) than Group B (M = 50.9 min, SD = 12.2; *p* = .067). GPT-productivity (M = 0.14 vs. 0.08; *p* = .067) and self-reported productivity (M = 0.15 vs. 0.08; *p* = .067) were higher, though sample sizes were small.
- **Support staff** in Group A completed tasks faster (M = 39.3 min, SD = 11.2) than Group B (M = 46.1 min, SD = 11.2; *p* = .088). GPT-productivity was significantly higher with GenAI (M = 0.13 vs. 0.08; *p* < .001), and self-reported productivity also trended higher (M = 0.11 vs. 0.09; *p* = .056).

GenAI support led to faster task completion across all domains, with productivity scores significantly higher when quality was considered. Gains were strongest in research, moderate in education, and mixed for support staff, where measured improvements were clearer than self-perceived ones.

4 Experiences with the EduGenAI Platform

The EduGenAI platform was valued for its safety, personas, and accessible design. However, its impact was limited by unclear functions, the need for better onboarding, and weaker integration compared to widely used external platforms.

Participants reflected on their experiences with the EduGenAI platform regarding safety, usability, and integration with their daily work. They reported mixed views that reflected strengths and areas where improvement is needed. Many participants described the platform as GDPR compliant and safe to use within the institutional environment, which gave them confidence compared to other existing tools. Several users saw the interface as straightforward, and the use of familiar terminology made it more intuitive. Features such as personas also helped participants to get familiar with the platform quickly.

“The personas made it easier for me to start designing lessons. It saved me a lot of time thinking where to begin.”
(Interview, Education domain)

Others found parts of the interface unclear and relied on trial and error to learn how to use the system. Several participants expressed a wish for clearer onboarding and embedded guidance. Some experienced users preferred other platforms such as ChatGPT or Copilot, citing familiarity and smoother integration into their daily use software. Compared to those tools, EduGenAI was sometimes considered less convenient because it required switching systems. Some hesitated to input sensitive internal data despite the GDPR compliance.

“It’s usable, but ChatGPT is always open in my browser — switching to EduGenAI felt less convenient.”
(Interview, Research domain)

5 Discussion

This section reflects on the pilot's findings and considers what they mean for the use and scaling of GenAI at THUAS. The discussion combines evidence from literature, surveys, interviews, and the RCT with broader reflections on institutional and cultural factors.

5.1 Findings by Dimension: Usability, Quality, and Productivity

The pilot revealed different perspectives across the three dimensions of usability, quality, and productivity.

Usability was experienced positively overall. Ease of use remained stable across the pilot period, consistent with the literature noting that LLM interfaces are generally intuitive. However, perceived usefulness declined once staff engaged with EduGenAI in practice. This pattern suggests that while first impressions were optimistic, hands-on use led to more critical evaluations as staff explored the platform’s affordances and limitations. Interviews confirmed that while personas and familiar terminology lowered entry barriers, participants still needed time to experiment with prompts and adapt outputs. This suggests that usability is not only a matter of technical design but also of user guidance and confidence in applying GenAI to daily tasks.

Quality was where GenAI had its most apparent effect. Across domains, outputs produced with GenAI were rated higher by GPT-based evaluations. Staff, however, assessed quality more cautiously. Teachers and researchers described GenAI as helpful for producing structured drafts and repetitive tasks where speed and clarity were valued. At the same time, they stressed that these drafts lacked subject depth and required refinement. Support staff were more divided, reflecting the difficulty of applying generic outputs to their highly context-specific tasks. The literature on GenAI in higher education echoes these findings: studies frequently report gains in efficiency and structure, but also highlight recurring problems with factual accuracy, hallucinations, and incomplete referencing. This is why many authors stress the need for systematic human review and refinement before outputs can be used in professional practice. The divergence observed here between staff self-assessments and model-based quality scores reflects the same issue: while textual outputs may look polished, professional standards demand contextual accuracy, trustworthiness, and accountability that GenAI alone cannot guarantee.

Productivity showed mixed results. Surveys indicated moderate benefits, particularly for researchers and teachers, but support staff were more cautious. The RCT confirmed that tasks were completed faster with GenAI, and productivity scores were significantly higher when quality was factored in. However, staff did not always perceive these gains themselves, highlighting the difference between measured and experienced productivity. This tension is also described in earlier studies, where the cognitive effort of learning and verifying AI outputs offsets time savings. At THUAS, this manifested in participants balancing the speed of generating drafts against the effort needed to refine, contextualize, and validate them.

5.2 Findings by Domain: Education, Research, and Support staff

The pilot revealed differences in domain perspectives:

Teachers highlighted GenAI's role in reducing preparation time and generating lesson structures, aligning with literature that positions AI as a support tool for repetitive and time-intensive tasks. However, they consistently noted that pedagogical judgement was essential to avoid overly generic outputs. The decline in perceived usefulness in surveys suggests that teachers became more critical once they tested outputs in practice, balancing efficiency with concerns about subject depth and individual student needs.

Researchers reported the most consistent benefits. GenAI was valued for drafting abstracts, structuring texts, and speeding up literature reviews. These uses resonate with studies showing that GenAI can accelerate academic writing while raising questions about reliability and referencing. The RCT confirmed that research tasks were completed faster and rated higher quality with GenAI support. Still, concerns remained about factual accuracy and transparency of sources.

Support staff were the most divided. While some appreciated GenAI for summarizing meetings or drafting communications, others found connecting to their varied and context-specific roles harder. Survey data showed the sharpest decline in perceived usefulness, and interviews revealed hesitation linked to uncertainty about what tasks were "allowed," time to explore, and fears of over-reliance. Literature on AI in administrative roles similarly notes that benefits are less direct when tasks require context-specific knowledge or sensitive data.

5.3 Conditions Supporting or Hindering Use

Differences in outcomes were shaped by digital literacy, task appropriateness, institutional support, and personal attitudes toward technology.

Several conditions shaped how THUAS staff engaged with GenAI during the pilot.

Digital literacy was highest among researchers and lowest among support staff, which was reflected in their confidence and perceptions of using GenAI.

Task appropriateness was another essential factor. Teachers and researchers, often with more prior experience, could readily connect GenAI to professional tasks such as lesson design, writing support, and analysis. Support staff, by contrast, struggled to identify direct applications, reflecting the more context-specific nature of their work and the use of personal data that makes GenAI harder to apply.

Institutional support also played a role. Managerial encouragement, training workshops, helpdesk assistance, and clear manuals helped staff to build confidence. They further valued the assurance of GDPR compliant platform for safe use. Beyond these practical supports, they emphasized that usability was also cultural. They needed an environment where experimentation felt safe. As one participant reflected:

> *"AI can make our work faster, but only if people feel safe enough to explore it without being afraid of doing something wrong."*
> (Interview, Research domain)

Attitudes also influenced readiness. Some staff were skeptical about the accuracy of GenAI outputs. Uncertainty about best practices and ethical boundaries added to these challenges. A number of participants voiced concern about the environmental costs of large-scale AI use. Others worried about possibly losing writing skills or becoming too dependent on the technology.

> *"I worry that if I use it too much, I will stop thinking for myself."*
> (Interview, Support domain)

> *"I think about the energy GenAI uses. That matters to me."*
> (Interview, Education domain)

5.4 Limitations of the Study

Despite the strengths of the design, the study reports certain limitations.

Firstly, the TAM survey used was version 4[6]. While this version was considered suitable for a short pilot, more comprehensive versions are available. For longer-term research, testing other versions that capture a broader range of user perceptions would be valuable.

Secondly, the scope of the RCT was narrow, as it was conducted over a short period with a limited number of participants. While it provided valuable causal insights, its results cannot fully represent the long-term use of GenAI or its integration into complex workflows. We attempted to balance this limitation by combining other quantitative and qualitative methods, but further research with larger groups and over more extended time frames is needed.

Thirdly, there was a degree of self-selection bias. Participants volunteered for the pilot, which may mean they were more curious, motivated, or digitally confident than the broader staff population. Even though dropouts were also interviewed to reduce bias, this limitation should be considered when interpreting the results.

Finally, the study was conducted in the specific context of THUAS. The findings may not directly translate to other institutions with different levels of digital maturity, workflows, or organizational cultures. Future work should aim for findings that can be reproduced or reused across different contexts, as this would strengthen their generalizability and value.

List of Abbreviations

- **3E Framework** – Evidence-Informed Evaluation of EdTech Framework
- **AI** – Artificial Intelligence
- **FZ/IT** – Facilitaire Zaken & Information Technology
- **GenAI** – Generative Artificial Intelligence
- **GDPR** – General Data Protection Regulation
- **HR** – Human Resources
- **ICC** – Intraclass Correlation Coefficient
- **LLM** – Large Language Model
- **LTA** – Learning Technology & Analytics
- **PEoU** – Perceived Ease of Use
- **PU** – Perceived Usefulness
- **RCT** – Randomized Controlled Trial
- **TAM** – Technology Acceptance Model
- **ToC** – Theory of Change
- **THUAS** – The Hague University of Applied Sciences

References

1. The Hague University of Applied Sciences. Inquiry-based learning with impact. Strategic plan 2023-2028. Accessed September 30, 2025. https://www.thuas.com/sites/hhs/files/documents/AboutTHUAS-Organisation-instellingsplan-2023-2028-english.pdf

2. Sengar SS, Hasan A Bin, Kumar S, Carroll F. Generative artificial intelligence: a systematic review and applications. *Multimed Tools Appl*. 2024;84(21):23661-23700. doi:10.1007/s11042-024-20016-1

3. ChatGPT. Accessed September 30, 2025. https://chatgpt.com

4. Giannakos M, Azevedo R, Brusilovsky P, et al. The promise and challenges of generative AI in education. *Behaviour & Information Technology*. 2025;44(11):2518-2544. doi:10.1080/0144929X.2024.2394886

5. Garg M, Baker T. *The Dutch 3E Framework: Evidence-Informed Evaluation of EdTech*.; 2025. doi:10.5281/zenodo.15070789

6. Lewis JR. Comparison of four TAM item formats: effect of response option labels and order. *J Usability Stud*. Published online 2019.

7. Braga LH, Farrokhyar F, Dönmez Mİ, et al. Randomized controlled trials – The what, when, how and why. *J Pediatr Urol*. 2025;21(2):397-404. doi:10.1016/j.jpurol.2024.11.021

8. EduGenAI. Accessed September 30, 2025. https://npuls.nl/edugenai

9. Npuls. Accessed September 30, 2025. https://npuls.nl/

10. Kutty S, Chugh R, Perera P, et al. Generative AI in higher education: Perspectives of students, educators and administrators. *Journal of Applied Learning and Teaching*. 2024;7(2):47-60. doi:10.37074/JALT.2024.7.2.27

11. Castillo-Martínez IM, Flores-Bueno D, Gómez-Puente SM, Vite-León VO. AI in higher education: a systematic literature review. *Front Educ (Lausanne)*. 2024;9. doi:10.3389/feduc.2024.1391485

12. McDonald P, Hay S, Cathcart A, Feldman A. *Apostles, Agnostics and Atheists: Engagement with Generative AI by Australian University Staff*.; 2024. doi:10.5204/rep.eprints.252079

13. Viruel SR, Rivas ES, Palmero JR. The Role of Artificial Intelligence in Project-Based Learning: Teacher Perceptions and Pedagogical Implications. *Educ Sci (Basel)*. 2025;15(2). doi:10.3390/educsci15020150

14. Andersen JP, Degn L, Fishberg R, et al. Generative Artificial Intelligence (GenAI) in the research process - A survey of researchers' practices and perceptions. *Technol Soc*. 2025;81. doi:10.1016/j.techsoc.2025.102813

15. Fecher B, Hebing M, Laufer M, Pohle J, Sofsky F. Friend or foe? Exploring the implications of large language models on the science system. *AI Soc*. 2025;40(2):447-459. doi:10.1007/s00146-023-01791-1

16. Kallunki V, Kinnunen P, Pyörälä E, Haarala-Muhonen A, Katajavuori N, Myyry L. Navigating the Evolving Landscape of Teaching and Learning: University Faculty and Staff Perceptions of the Artificial Intelligence-Altered Terrain. *Educ Sci (Basel)*. 2024;14(7). doi:10.3390/educsci14070727

17. Jukiewicz M. The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Think Skills Creat*. 2024;52. doi:10.1016/j.tsc.2024.101522

18. Sáez-Velasco S, Alaguero-Rodríguez M, Delgado-Benito V, Rodríguez-Cano S. Analysing the Impact of Generative AI in Arts Education: A Cross-Disciplinary Perspective of Educators and Students in Higher Education. *Informatics*. 2024;11(2):37. doi:10.3390/informatics11020037

19. Marchena Sekli G, Godo A, Carlos Véliz J. Generative AI Solutions for Faculty and Students: A Review of Literature and Roadmap for Future Research. *Journal of Information Technology Education: Research*. 2024;23:014. doi:10.28945/5304

20. Stan MM, Dumitru C, Bucuroiu F. Investigating teachers' attitude toward integration of ChatGPT in language teaching and learning in higher education. *Educ Inf Technol (Dordr)*. Published online 2025. doi:10.1007/s10639-025-13396-w

21. Enang E, Christopoulou D. Exploring Academics Intentions to Incorporate ChatGPT into Their Teaching Practices. *JOURNAL OF UNIVERSITY TEACHING AND LEARNING PRACTICE*. 2024;21(8). doi:10.53761/rn5y5614

22. Cambra-Fierro JJ, Blasco MF, López-Pérez MEE, Trifu A. ChatGPT adoption and its influence on faculty well-being: An empirical research in higher education. *Educ Inf Technol (Dordr)*. 2025;30(2):1517-1538. doi:10.1007/s10639-024-12871-0

23. Tran TM, Bakajic M, Pullman M. Teacher's pet or rebel? Practitioners' perspectives on the impacts of ChatGPT on course design. *High Educ (Dordr)*. Published online 2024. doi:10.1007/s10734-024-01350-7

24. Belcher BM, Bonaiuti E, Thiele G. Applying Theory of Change in research program planning: Lessons from CGIAR. *Environ Sci Policy*. 2024;160:103850. doi:10.1016/j.envsci.2024.103850

Use of AI in the Project

AI tools were used in several stages of the research process. In the analysis phase, AI-assisted tools were used to transcribe and structure interview data and to help compare outputs in the RCT through large language model–based evaluations. In the reporting phase, GenAI was used to support drafting and refining sections of the text and also for translating the report into Dutch (the original version of the report is written in English). At the same time, the research team made all substantive decisions, and every AI-assisted output was reviewed, edited, and validated by human researchers.