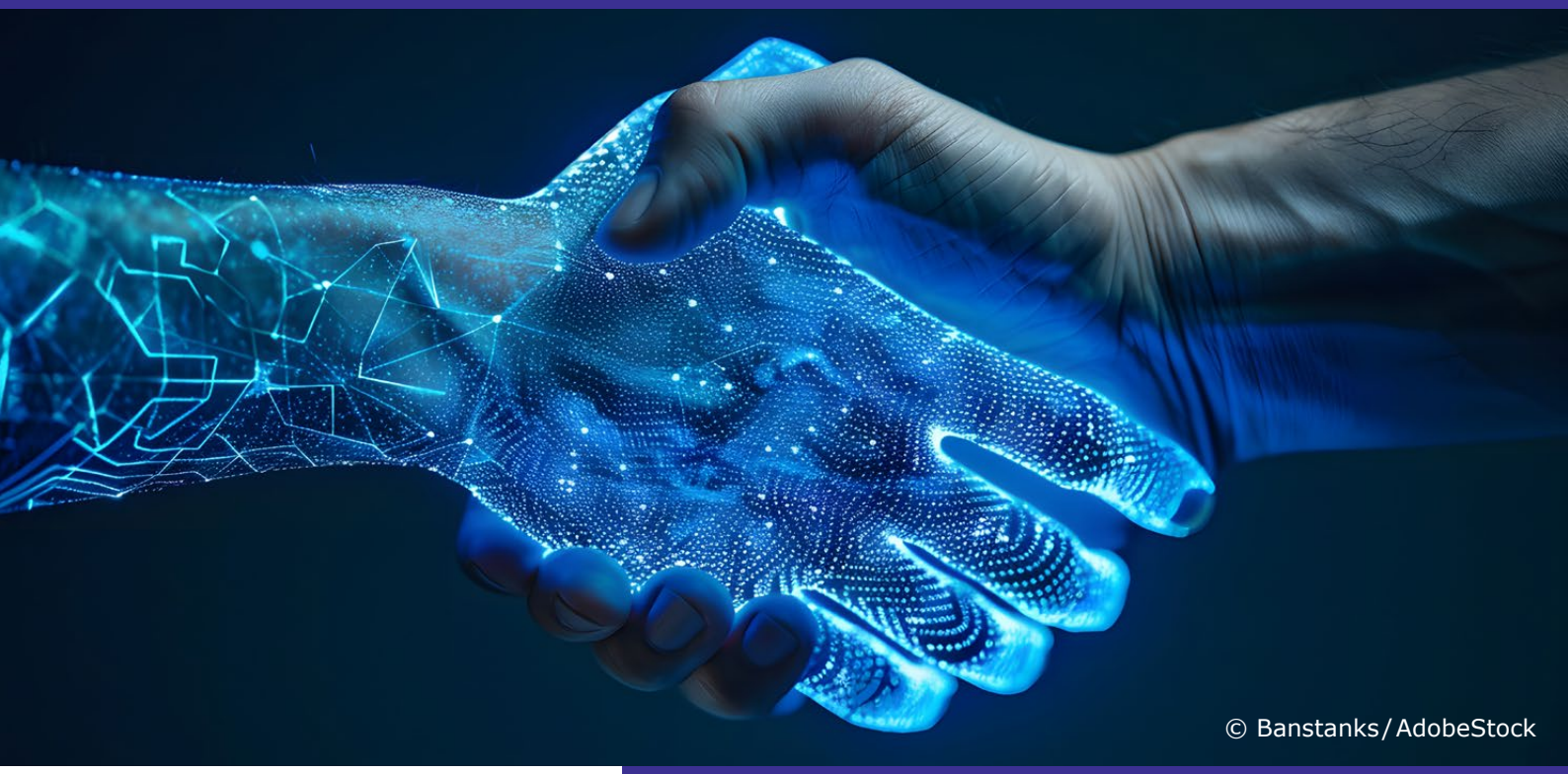
ESCP
BUSINESS SCHOOL

IT ALL STARTS HERE

BERLIN | LONDON | MADRID | PARIS | TURIN | WARSAW



© Banstanks / AdobeStock

LIGHTS - Technology

Human, AI, or both? Evidence from a standardized classroom task

ESCP Impact Paper No.2026-42-ENG

Houman KHAJEHPOUR
Universität Potsdam

Gonçalo PINA
ESCP Business School

ESCP RESEARCH INSTITUTE OF MANAGEMENT (ERIM)

ESCP Impact Paper No.2026-42-ENG

Human, AI, or both? Evidence from a standardized classroom task

Houman Khajehpour*
Gonçalo Pina**

ESCP Business School

Abstract

We show that the ability to evaluate AI output, knowing when to keep, edit, or discard it, is the defining professional skill in a management context. Using a class experiment, we compare human-only, AI-only (ChatGPT GPT-5 with Deep Research, January 2026), and human-AI collaboration on a standardized macroeconomic consulting task across 11 Master in Management groups. Generative AI reduced drafting and formatting time by roughly 40% on average. However, the time saved was largely absorbed by human remediation: average total AI-assisted production time is comparable to that of fully manual work. Looking at quality of output, AI-only and human-AI quality were similar on average and slightly above that of human-only output, yet AI-only outcomes were more dispersed. Competitive advantage, in the classroom and in the workplace, increasingly derives from the ability to audit, validate, and selectively trust AI-generated outputs, not from AI use per se. We translate the experimental evidence into a practical framework: when to trust AI output, when to edit it, and when to discard and restart.

Keywords: Artificial Intelligence, consulting, performance, human-AI collaboration.

*Universität Potsdam
**Professor, ESCP Business School

Human, AI, or both?
Evidence from a standardized classroom task

Introduction

Artificial intelligence is reshaping how analytical and knowledge work is performed. AI tools can now draft reports, build presentations, and execute data analysis faster than human teams. They produce results that managers often find difficult to distinguish from human-authored work (Bick et al., 2026; Yang et al., 2023). Recent experimental evidence shows that ChatGPT can raise productivity in professional writing tasks, reducing average time by 40% while improving output quality by 18% (Noy & Zhang, 2023). Yet speed is not the whole story. When AI output contains errors, including misidentified data, incorrect visualizations, or misrepresented trends, the human time required to remediate those errors can offset the initial efficiency gains.

In academic and professional settings alike, the rapid adoption of AI has outpaced our understanding of where it succeeds and where it fails. AI is associated with greater output volume, but also with the risk of reduced analytical depth and critical thinking (Vieriu & Petrea, 2025; Bick et al., 2023). These risks are not confined to classrooms. Many workplace AI-based tasks involve the same cycle of producing, checking, and correcting analytical outputs. The ability to detect AI errors matters as much as AI's ability to generate content.

Despite the growing use of generative AI across professional and academic settings, evidence on the precise time and accuracy trade-offs of AI assistance in specialized, data-intensive tasks remains scarce. Existing studies typically compare AI-only and human-only output in isolation, or rely on self-reported productivity gains, leaving the remediation cost, that is, the human time required to bring AI output to professional standards, largely unmeasured.

This paper makes three contributions. First, we report a within-task classroom experiment in which the same 11 Master in Management groups produced a fully human version, a fully AI-generated version, and an AI-human collaboration of an identical macroeconomic consulting deliverable, allowing a direct comparison rather than across-subject inference. Second, we measure generation and remediation time separately and show that, although AI generation is on average 40% faster than fully manual work, the time saved is largely absorbed by the human remediation needed to bring AI output to professional standards. Our findings show that the total time required for the AI-assisted task, including both AI generation and human remediation, is only slightly lower than that required for manual work: the mean time is the same, while the median time is 13% lower. Third, we measure quality by grading the three different outputs. We show that AI-only and human-AI work had similar average quality (mean grades of 85 and 84 out of 100, respectively, against 81 for human-only) but with wider dispersion, with AI-only outcomes ranging 50-100 vs. 70-100 for human-only. We also show human remediation will usually improve a poor AI deliverable, but, in at least one case, can also worsen it.

The central skill the task elicits is therefore the ability to evaluate AI output, knowing when to keep, edit, or discard it. In this task, this means assessing whether the AI retrieved the right data, applied the correct transformations, and produced results that pass a domain-specific knowledge check. Crucially, managers need to be able to evaluate quickly if the AI-output error is easy enough to fix, or deep enough that rebuilding from scratch would be

more efficient. We translate these questions into a practical decision framework applicable to managers and professionals working with AI in any analytical context.

The remainder of the paper is structured as follows: Section 2 describes the assignment design and requirements; Section 3 compares the time spent on human versus AI-generated tasks; Section 4 evaluates the quality and accuracy of the resulting outputs; and Section 5 concludes with a practical framework for professionals and educators on how to evaluate, trust, and remediate AI-generated analytical work.

The assignment: Foreign Expansion Plan – Global Macro Trends

The core of this study is the "Foreign Expansion Plan" assignment, a macroeconomic consulting task integrated into the Master in Management (MiM) class "Global Macro Trends" at ESCP Business School's Berlin campus. Students were required to act as analysts evaluating the macroeconomic environment of two prospective markets to guide a planned business expansion. The fundamental objective was to demonstrate an understanding of economic data and indicators through searching, downloading, and plotting time-series variables to support a strategic recommendation.

Groups were tasked with choosing two countries and using the Federal Reserve Economic Data (FRED) database to analyze 25 years of data or the maximum available period of quarterly or annual data. The variables of interest included Real Gross Domestic Product (GDP) growth (year-on-year percentage changes), unemployment rates, inflation rates (percent change from year ago), exchange rates relative to the USD, and risk measurements such as government bond spreads relative to German Bunds.

Groups were composed of three to six members and asked to record the time taken to complete each part of the project. The fundamental requirement of the assignment involved data visualization. For each graphical representation, the instructions explained what was to be plotted and students were asked to accompany each plot with "active titles", titles that summarized a strategic insight rather than just labeling the plot (e.g., "Spain's labor market shows persistent weakness compared to Italy").

The assignment workflow was divided into distinct phases. In "Part A - Human Only", students performed all data collection and plotting manually without any AI assistance. In "Part B - AI Only", they provided the same instructions to ChatGPT with the "Deep Research" option enabled to generate a parallel slide deck. This comparative approach was designed to assess differences between students and artificial intelligence in complex data environments.

For the final stages, the team was split into two sub-teams for Parts C and D. Sub-team 1 was responsible for "Part C - Evaluation," conducting a manual critical assessment of the AI output's correctness, focusing on units, transformation, and frequency mismatches. Sub-team 2 handled "Part D - Human Fixes AI's presentation," using human labor to remediate the AI's errors from "Part B - AI Only" and bring the final deck to professional standards. By construction, the time recorded for "Part D" is the sum of the AI generation time in "Part B" and the human remediation time, and is therefore the AI-assisted analogue of Part A. Time is counted as billable hours. If a team of 4 took 30 minutes to produce Part A, that is counted as 120 minutes. Groups had to submit human decks, AI decks, evaluation sheets, and refined final presentations.

In the next sections, we evaluate all groups' submissions, assessing them along two critical dimensions: time and work quality.

The promise of AI: Efficiency

The general outcomes regarding task completion time revealed a considerable, though highly variable, disparity between manual and AI-assisted creation. Across the 11 student groups, human manual effort (Part A) ranged from 90 minutes to an outlier of 480 minutes for one team, with a median of around 194 minutes. In contrast, the AI generation phase (Part B) was significantly faster, often taking less than 100 minutes, with the longest time being 240 minutes. This aligns with broader research suggesting that using ChatGPT in professional writing tasks raised productivity and reduced the average time taken by 40% (Noy & Zhang, 2023).

The positive side of the time data was a clear efficiency gain for administrative and formatting tasks. Teams like Group 1 were able to produce an initial draft in just 60 minutes using AI, compared to 167 minutes manually, which is a 64% decrease in creation time. AI can be highly effective at quickly organizing the summaries and structuring presentations. Offloading routine drafting frees the preparation time normally required for deeper analysis.

The remediation stage, however, highlighted the drawbacks of using AI. For many groups, the time saved in generation was lost in correction. Group 9 spent 230 minutes on the AI generation but required an additional 185 minutes of human work to fix the output, effectively doubling the task duration. Furthermore, Group 3 faced difficulties when producing output, spending 240 minutes on AI generation because the model initially failed to plot any graphs correctly or missed the indicators, which required repeated prompts and iterations.

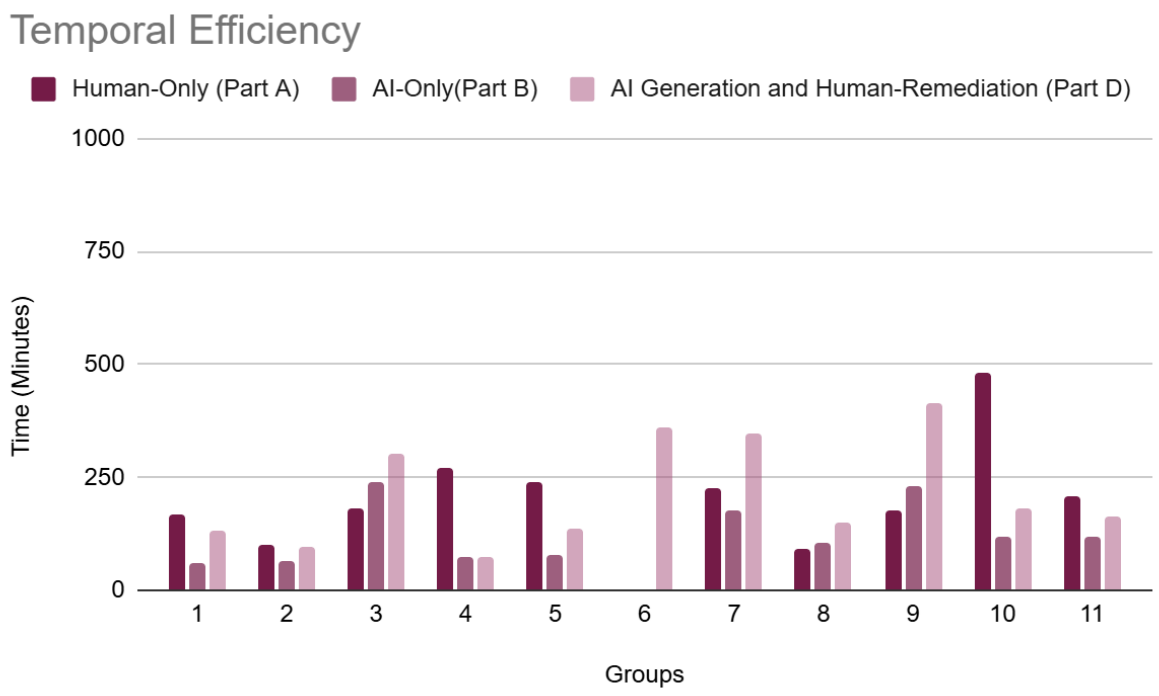


Figure 1, Comparison of Time-on-Task for three parts of the task (Part A = fully manual; Part B = AI generation only; Part D = Part B + human remediation, comparable to Part A).

To summarize, although AI demonstrates the capacity to significantly surpass human speed in processing tasks and generating drafts, it shifts human efforts toward remediation and verification. The experiment showed that the apparent efficiency of AI generation may be

illusory if the output is not immediately accurate, as the need for recalculation and the sensitivity to prompts can offset the initial efficiency. Consequently, the complete production time is highly dependent on how many errors the AI produces, since correcting them requires significant intervention by humans to reach expected standards.

Output quality: fallible precision

Across the 11 groups, the mean grade was 81 for Human-Only, 85 for AI-Only, and 84 for Human-Fixes-AI. Only one team (Group 4) scored 100 on every methodology. The most informative pattern, however, is the dispersion: human-only grades ranged from 70 to 100, AI-only from 50 to 100, and Human-Fixes-AI from 60 to 100. AI-only therefore had both the highest ceiling and the lowest floor. When AI got the task right, it produced polished, data-accurate decks faster than any human team, but when it failed (smoothed curves, missing plots, opaque series codes), it required substantive rebuilding rather than light editing. Human remediation rescued AI output in some cases (Group 9 went from 80 on AI-only to 90 after human fixes) but worsened it in at least one (Group 8 went from 80 to 60). The most valuable competency the experiment elicits is therefore not human-AI collaboration but evaluation of AI output.

The best AI-only submissions collected the correct economic indicators from FRED and correctly implemented all required graphs. In contrast, human teams were generally better at creating active titles that conveyed strategic insights and conclusions rather than mere descriptive labels. Furthermore, human teams demonstrated stronger reasoning when implementing and interpreting the task. Group 4, for example, executed all manual and AI-generated parts correctly, serving as a model for precise data extraction and task execution.

Figure 2, Distribution of Grades across Methodologies

However, the evaluation also revealed significant mistakes in performance and missing components in other AI outputs. In some cases, AI identified the task and sources but was unable to transform the data into a data-driven visualization, instead producing smooth waves and curves that did not accurately represent the correct trends. Specific failures included a lack of transparency in the risk assessment plots, as well as an inability to plot exchange rate graphs correctly (Groups 7, 8, and 9). These results highlight the risk of misinterpretation and limited reliability in specialized visualizations, even when the model presents its findings with high confidence.

The human decks also contained unexpected technical errors that mirrored common pitfalls. Additionally, several human-only teams encountered difficulties with unit transformations, such as Group 6's inflation graph, which displayed absolute index points instead of the required year-on-year growth. This resulted in a distorted vertical axis scale ranging from -200 to 1,800. Risk assessment plots were a weakness for human teams as well, with Group 2 and Group 9 attempting to combine bond yields and share prices on dual-axis charts that obscured meaningful comparisons. Finally, a notable mistake occurred in the evaluation phase of one cohort; while the AI correctly utilized exchange rate levels, the human team in Group 7 plotted a "Percent Change from Year Ago" in their manual slides and incorrectly judged the AI's output to be wrong on the basis of their own faulty analysis.

In summary, the evaluation phase (Part C) highlighted that while AI can in some cases perfectly execute macroeconomic analysis in isolation, it remains inconsistent in doing so. The human errors were often due to oversight or misapplication of proxies, while AI errors were foundational failures in understanding specific commands or misinterpreting data-driven visualization requirements. This confirms that professional standards in consulting can only be maintained through a rigorous human-AI collaboration process where human domain expertise is used to cross-validate every automated output.

A practical decision framework for working with AI

This paper presented a within-task classroom experiment that quantifies, on a single standardized deliverable, the time and quality trade-offs of human-only, AI-only, and human-AI work. The evidence is two-sided. Generative AI compressed drafting and formatting time by roughly 40% on average (range 30–64% across groups) and, in successful cases, produced data-accurate decks on the first attempt. In other cases, it failed at high-precision steps, for example, missing indicators, misreading data series, or producing smoothed curves that did not represent the actual trend. Crucially, human-only work was not error-free either: groups displayed inflation as index points instead of year-on-year growth, and combined incommensurable series on dual-axis charts. These results should be interpreted with the study's limitations in mind: 11 groups from a single program, one type of data-intensive task, and a specific AI model in a controlled classroom setting. Broader generalization will require replication across different task types, organizational contexts, and AI tools.

These findings carry implications well beyond the business school classroom. In any setting where analysts, consultants, or professionals use AI, they must be trained to recognize that AI's polished prose can mask weaknesses in the underlying data, particularly in financial and economic reporting. This observation is consistent with Bick et al. (2023). Every AI-generated output, whether in a student assignment or a professional deliverable, should be treated as a fallible draft requiring cross-validation by someone with domain expertise. For educators, the core competency to develop is the ability to critically evaluate AI output before acting on it. For managers and professionals, competitive advantage increasingly derives from the ability to audit, validate, and selectively trust AI-generated outputs, not from AI use per se. In our experiment, the most consistent strong results came from pairing AI generation with human evaluation and refinement, confirming that the ability to evaluate and refine AI output is a core professional competency.

Our data support a practical decision framework for working with generative AI on analytical tasks. Keep the output when the AI has shown it can retrieve the correct data, apply the right transformations, and produce results that pass a domain knowledge check,

as happened for the strongest groups in our experiment, who obtained accurate decks with no remediation required. Edit the output when the structure and narrative are sound, but minor defect corrections, such as specific data points, units, or visualizations, are needed. The human remediation time is usually lower than rebuilding from scratch. Discard and restart when foundational errors undermine the entire output. In our experiment, several groups spent more time fixing AI output than they would have spent working manually. Communicating and documenting what works and what does not has the potential to improve efficiency further. Training professionals to judge quickly whether to keep or discard an AI output can likewise reduce the time lost to unproductive remediation. More broadly, the goal of integrating AI into professional and academic work is not to replace human effort but to reallocate it toward judgment, validation, and framing. In the human-AI collaboration model that emerges from our data, AI absorbs the heavy lifting of layout and formatting while human expertise remains the source of data integrity, strategic narrative, and accountability for the final result. Knowing when to delegate to the model and when to override it will be a defining skill for the next generation of managers.

References

Bick, A., Blandin, A., & Deming, D. J. (2026). The rapid adoption of generative AI. *Management Science*.

Bick, M., Breaugh, J., Dong, C., Pina, G., & Waldner, C. (2023). AI and the future of academic writing: Insights from the ESCP Business School Prompt-o-thon workshop in Berlin (ESCP Impact Paper No. 2023-24-EN). ESCP Business School. https://academ.escpeurope.eu/pub/IP%20N%C2%B02023-24-EN_Bick_Breaugh_Dong_Pina_Waldner.pdf

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187-192.

Vieriu, A. M., & Petrea, G. (2025). The impact of artificial intelligence (AI) on students' academic development. *Education Sciences*, 15(3), 343. https://doi.org/10.3390/educsci15030343

Yang, C. L., Allen, L., Restrepo-Amariles, D., & Troussel, A. (2023). GPT in the loop: Evidence from the field. Available at SSRN: https://ssrn.com/abstract=4658765 or http://dx.doi.org/10.2139/ssrn.4658765

Appendix

Group	Human-Only (Part A)	AI-Only(Part B)	Human fix	AI Generation and Human-Remediation (Part D)
1	167	60	70	130
2	100	65	30	95
3	180	240	60	300
4	270	75	N/A	75
5	240	78	60	138
6	N/A	N/A	360	360
7	225	170	175	345
8	90	105	45	150
9	175	230	185	415
10	480	120	60	180
11	208	120	45	165

Table 1: Temporal Results of Class Assignment for Each Group (note that groups had different sizes ranging 3-6 students)

Group	Human-Only	AI-Only	Human Fixes AI's presentation
1	90	100	100
2	80	100	80
3	70	50	70
4	100	100	100
5	90	50	90
6	80	90	80
7	80	90	90
8	70	80	60
9	70	80	90
10	80	100	80
11	80	100	90

Table 2: Grades Results for Each Part