Leveling Up or Leveling Down? The Impact of Large Language Models on Student Performance in Higher Education

Oana Vuculescu, Franziska Günzel-Jensen, Lars Frederiksen and Carsten Bergenholtz[1]
Department of Management, BSS, Aarhus University

[1]corresponding author: cabe@mgmt.au.dk

25.09.2024: This is a pre-print. Please do not share without permission.

Abstract
The rapid adoption of Large Language Models (LLMs) like ChatGPT-4 is transforming teaching and assessment practices at higher educational institutions. Our study investigates the impact of LLMs on student performance in an open-ended exam scenario. While existing literature suggests that LLMs generally enhance performance across various tasks and contribute to a democratizing effect—especially benefiting lower-performing individuals—our research presents a more nuanced picture. Through a mixed methods approach, we conducted an experimental lab-study (N=146) with business school students analyzing an Organizational Behaviour case study. Students first solved a task without having access to LLMs, and subsequently randomized access to ChatGPT-4 in the second task. Our findings reveal an "equalizing effect," where low-performing students significantly improved their performance with LLM assistance, and high-performing students experienced a decline in performance, bringing them to the level of their lower-performing peers. Qualitative follow-up interviews (19) highlighted that high performers struggled to effectively integrate LLM outputs into their work, mirroring the challenges faced by low performers who often resorted to simple copy-paste strategies. These results underscore the need for a deeper understanding of how LLMs can be leveraged to benefit all learners without inadvertently disadvantaging high achievers.

1. Introduction

The landscape of higher education has been transformed by technological innovations over the past decades. This has led to profound changes in teaching and assessment methods. From the adoption of calculators, and the ubiquity of internet resources to the integration of digital communication platforms like Zoom during the COVID-19 pandemic, universities and students have continually adopted and adapted to new technologies (Salomon et al. 1991, Lodge et al. 2023). The latest technological advance that impacts higher education is the advent of Language Models (LLMs) like ChatGPT-4. These technologies are rapidly incorporated into educational systems and adapted by students (Malmström et al. 2023), prompting extensive discussion (Lodge et al., 2023; Mollick & Mollick 2023, Larson et al. 2024). This ongoing discussion encompasses a range of issues including academic integrity (Cotton et al. 2023), morality (Moser et al., 2022), practical challenges (Yan et al.,

2023), and the potential effects on educational outcomes and student performance (Extance 2023, Larson et al. 2024).

Recent experimental studies indicate that LLMs generally enhance performance across various tasks when individuals have access to these technologies (Dell'Acqua et al. 2023; Noy & Zhang 2023; Brynjolfsson et al. 2023, Merali 2024). This improvement is evident in fields as diverse as creativity, coding, customer service, and case analysis (Brynjolfsson et al. 2023; Doshi & Hauser 2023; Dell'Acqua et al. 2023, Cui et al. 2024). The benefits of LLMs seem especially significant for those initially performing at lower levels, suggesting a democratizing effect, enabling a broader range of users to achieve higher performance (Noy & Zhang 2023; Brynjolfsson et al. 2023; Choi and Schwarcz 2023, Merali 2024).

However, in educational settings, experiments find that the impact of LLMs varies. For instance, Nie et al. (2024) indicate small positive effects in scenarios such as introductory coding classes, whereas an investigation into using LLMs in legal tasks showed mixed results (Choi and Schwarcz 2023). The varying impact on performance of students across different studies may be attributed to the nature of the tasks: LLMs tend to support performance better for solving well-defined problems with clear starting points and end goals. However, their efficacy may diminish in tasks that demand higher levels of problem formulation (Otis et al. 2023) particularly in more complex scenarios (Choi and Schwarcz 2023) where solutions are less straightforward and not predefined (Boussioux et al. 2024). This is relevant in higher education examinations, such as case analyses which are often applied at business schools, which can be ambiguous and open-ended. The degree to which students benefit from LLM assistance in exam situations remains an open question, particularly when comparing high and low performers.

Given these considerations, we ask: What is the performance impact of allowing the use of LLMs on the performance of high- and low-performing individuals in open-ended exam settings within a higher educational setting? Our study contributes to the existing body of literature by exploring these questions through an embedded mixed methods study. Unlike previous studies, our research not only measures the quantitative impact of GAI on student performance in an experimental study but also

integrates qualitative follow-up interviews with experiment participants to understand the underlying mechanisms at play.

In our experimental study, where students at a business school analyzed an Organizational Behaviour case study, we find that low performing participants increase their performance when allowed to work with LLMs while high-performing participants decrease in their performance. This seems to constitute a specific, undesirable form of an "equalizing effect", compared to the notion of a "democratizing effect" highlighted in prior studies (Noy & Zhang 2023; Brynjolfsson et al. 2023; Dell'Acqua et al. 2023). Whereas the democratizing effect generally implies that performance differences are reduced by raising both low and high performers to a uniformly high level, an equalizing effect narrows the performance gaps by disproportionately elevating lower performers, without benefiting—or even hindering—the progress of higher performers. Our qualitative study adds further nuance by showing how high-performing participants are hindered as they report to struggle to integrate LLM outputs in their text production and workflow. This is similar to low-performers, who feel overwhelmed by the task and often engage in a simple copy-and-paste strategy.

2. Hypothesis development

Our research question focuses on a typical exam situation at a Nordic business school setting. In these exams, questions are generally open-ended, requiring students to analyze a given scenario, identify and prioritize relevant information and angles, and structure a coherent answer to the question. Often a range of different approaches can be used. The challenge thus constitutes a more open-ended problem than many of the cited studies in this emerging field of LLM studies, where fairly detailed questions are to be addressed (such as Dell'Acqua et al. 2023). On the other hand, an exam situation is less open-ended and less ambiguous compared to the situations entrepreneurs face (Otis et la. 2023), since the challenge takes departure in a given, lengthy case and some open-ended questions are provided.

We argue that obtaining access to ChatGPT-4 is, on average, going to improve the performance of students. This is due to a range of reasons. For one, since ChatGPT-4 is trained extensively on a range of datasets, including academic literature and real-world examples relating to the topic

(Organizational Behavior) the chatbot is likely to offer relevant information (Dell'Acqua et al. 2023). Getting access to the chatbot offers cognitive support, by helping the students to structure and synthesize relevant information (Noy & Zhang 2023; Lodge et al. 2023). The chatbot is also able to produce relatively coherent and fluent writing (Noy & Zhang 2023), in particular if compared to non-native English speakers (Herbold et al. 2023). While LLMs did not provide a consistent positive average effect when dealing with more complex, essay-based law scenarios (Choi and Schwarcz 2023), we note that when analyzing a case in organizational behavior, one is not required to rely on deep insight into complex legal documents and precedence, nor do we expect students to cite such documents. Furthermore, while Otis et al. (2023) did not find a positive effect of providing access to a LLM, their challenge was very open-ended. Individuals had to figure out what their problem was, how to frame it and how to interpret the answers received. In an educational setting where students are to analyze a case, they are provided insight into a given problem-scenario, which they then have to make sense of. Overall, we expect that students, on average, will benefit from accessing a state-of-the-art chatbot. Our primary hypothesis is as follows:

H1: When analyzing a business case, getting access to ChatGPT-4 improves performance.

The studies reviewed generally imply that lower performing individuals get a bigger performance boost, compared to high performers (e.g. Dell'Acqua et al. 2023; Choi and Schwarcz 2023; Doshi and Hauser 2023). We expect a similar pattern in our data for the following reasons. Low performers may not know where to begin and can therefore benefit from getting initial assistance (Lodge et al. 2023) from a chatbot that, as mentioned, has obtained relevant information on the topic. ChatGPT-4 has also demonstrated an ability to structure a response to educational questions, and is able to write fairly convincingly, especially compared to the English proficiency of non-native English writers, in a Nordic University (Herbold et al. 2023). This could benefit the weaker performers in particular. Furthermore, we note the fairly detailed nature of the challenges in some of the reviewed studies (e.g. Dell'Acqua et al. 2023; Dosi and Hauser 2023). ChatGPT-4 generally thrives on contextual details, and if details are lacking it provides answers of a lower quality, compared to challenges with high levels of detail (Mollick 2024). This point aligns with business schools' efforts to create exam questions that a LLM cannot easily answer. At our educational institution, we have in a range of

business-oriented courses seen the limits of state-of-the-art models (i.e. ChatGPT-4 in Summer/Autumn 2023), when the chatbot does not know the specific expectations of a given requested case analysis. This setup might imply that LLM output does not immediately exceed an average performance level and would require additional work or iterative prompting to lead to higher performance (see also Choi and Schwarcz 2023). Therefore, LLM assistance is more impactful for weaker performers, compared to better performers.

Finally, akin to an exam situation, we are grading on a closed scale, meaning that a very good performance produced without the chatbot, cannot be improved much upon, if at all, when gaining access to the chatbot. We are thus both expecting the low performers to improve more than the average, and the high performers to improve less, if at all. Our secondary hypothesis is therefore as follows:

H2: When analyzing a business case, getting access to ChatGPT-4 improves performance more for low-performers, compared to high-performers.

3. Methods

To enhance our understanding of the effect of integrating LLM on student performance, we employ an embedded mixed methods design (Creswell et al., 2011), integrating qualitative interviews within a quantitative experimental study. By embedding qualitative interviews, we aim to capture in-depth insights into the participants' experiences and perceptions that influence their performance. As such, we combine quantitative and qualitative methods in a “test and explore” approach (Wellman et al., 2023). The objective of the experimental trial is to assess the performance impact of allowing the use of LLM on participants’ performance, particularly among high- and low-performers, in exam settings in business schools. Concurrently, the qualitative component seeks to explore the underlying reasons behind observed performance changes. This integrated approach not only enriches our understanding of the quantitative outcomes but also provides a more comprehensive view on the underlying mechanisms of LLMs impact, guiding future improvements and implementations.

3.1 Quantitative study: Experimental design

We conducted a randomized controlled trial to assess the impact of using ChatGPT-4. The study setup, hypotheses and analytical approach are pre-registered (https://aspredicted.org/G8P_XDP). We received ethical approval from Aarhus University (BSS-2023-086). Data was collected in October and November 2023 following recruitment from a pool of participants at the university.

The final sample was composed of 146 participants (73 in each condition[1]), almost exclusively students, with an average age of 24.1, where 85 participants identified as female, as 55 as male (see Table 1). The data was collected in the local social science lab in batches of at most 24 participants at a time. English fluency was the only inclusion criteria. In total participants spent approximately 70 min in the experiment. They were paid a flat rate (about 19 Euros) and a potential performance bonus; the top 20% best performing participants had a 10% chance of winning about 53 Euros in a lottery.

3.1.1 Experimental task: Procedure

Initially, participants watched a 13-minute video that presented an organizational behavior case, representing a typical business school challenge and teaching material. Core issues relate to extrinsic and intrinsic motivation, group vs individual incentives and creating a collective identity. Participants were then asked two attention checks; recalling the number of female quotes in the video and the name of the case company. No participants failed the attention check.

Task 1: Baseline performance: In their first writing task, all participants were asked to answer a question about the given case study. They were informed that they have 22 minutes for this part of the experiment and were given access to a transcript of the video. This first task constitutes the baseline performance in the experiment, subsequently used to distinguish high vs. low performers in the qualitative and quantitative analysis. Following this first task, participants were asked to fill in a short survey related to efficacy and enjoyment.

[1] Although we had an even distribution of participants, a handful of participants in the treatment group did not use the chatbot. We have decided to focus our subsequent analyses on differences between those who used the chatbot, and those who didn't, rather than on 'intention to treat', that is the treatment allocation. Results obtained using the treatment allocation distinction are virtually identical.

Task 2: Treatment and control: In task 2 participants received a new question[2] and were again informed they had 22 minutes to respond. Some participants were now randomly given access to a React web app interface linked to a GPT-4 API. The web app was similar to OpenAI's ChatGPT-4 interface (see Figure 1). All interactions (prompts and answers) were recorded for each participant. Being that participants were only asked to solve one task the interface did not allow for the creation of multiple threads. All interactions were visible, and accessible through the interface. The system prompt was: 'I am a helpful assistant. Ask me anything'. The temperature parameter was left at the default value of 0.8 which is appropriate given that the task is not well defined, and thus generated responses are expected to be diverse, but not random (the upper limit is 2). As in their prior task, participants had access to the transcript of the video, but could not copy from it[3].

For logistic reasons the randomization was carried out per experimental session, not participants. Treatment groups were strongly encouraged to use the chatbot. The setup explicitly imitated an exam situation, where the control group was informed that this was an exam setting where accessing the internet was allowed and the treatment group was informed that among available resources, they also have access to ChatGPT-4. Receiving this kind of information was typical for students at the focal university. It was thus not straightforward for participants to infer if they were in a control or treatment condition.

Post-task survey: Following the completion of the second task, participants were asked to answer a survey, including a set of demographic questions, questions about how they experienced the tasks, and questions pertaining to their experience with cases. In addition, participants in the treatment group were asked about their prior experience with ChatGPT and LLMs more broadly. Finally, all participants were asked to rate how optimistic they are in relation to the use of AI technologies in higher educational settings (see Table 1 for an overview of means and standard deviations, and the Measures section for further details on the variables).

---

[2] The question that participants were to answer in task 2: "Keeping in mind that scientific work is often more successful when undertaken in teams, please discuss what the Cure Factory Copenhagen should consider when motivating teamwork."

[3] This was a deliberate choice being that we expected participants in the treatment group would have an unfair advantage if they were simply able to copy paste the entire case. This, in addition, captures the real-life exam where students are asked to deal with considerably longer cases which cannot be parsed easily by current state of the art LLMs.



Figure 1. GPT-4 interface.

3.1.2 Measures

Task 1 and 2 evaluation: Performance in task 2 constitutes our dependent variable. On average, participants produced about 270 words, across both conditions. Participants' performance in both tasks 1 and 2 were assessed by two of the authors, who have extensive experience in teaching and grading organizational behavior exams. All tasks were rated by both, on a scale from 1 to 10. Interrater agreement was 90.5% and 90.9% respectively, for task 1 and 2. In other words, on a 1 to 10 scale, the assessors differed on average 0.85 (task 1) and 0.82 (task 2). If assessors disagreed on a score, they discussed the submission and agreed on a rating. The assessment of task 2 constitutes the dependent variable. The grading criteria followed the principles of how exams in an Organizational Behavior bachelor level course would be graded at the given university; ability to identify and prioritize the key challenges implied by the exam case and question, usage of relevant terminology and clear and consistent language and ability to use relevant insights to clearly answer the question.

Case study experience: Participants were asked, on a scale from 1 to 5, to assess if this case-based task reflects something they have done in the past.

Felt enjoyment: Participants were asked to assess how much they enjoyed the task, on a scale from 1 to 10 (following Noy and Zhang 2023). This was asked for both task 1 and task 2.

Felt skilled: Participants were asked to assess how skilled they felt while completing the task, on a scale from 1 to 10. This was asked for both task 1 and task 2.

Felt effective: Participants were asked to assess how effective they felt while completing the task, on a scale from 1 to 10. This was asked for both task 1 and task 2. The last two variables were categorized as self-efficacy by Noy and Zhang (2023).

ChatGPT experience: Participants were asked to assess if they have previously engaged with chatbots such as ChatGPT, Bing, Google Bard or others (following Doshi and Hauser 2023, Noy and Zhang 2023)

Useful aid: Participants were asked to assess how useful their aid was; either the Internet or ChatGPT-4 + the Internet, depending on experimental condition (following Noy and Zhang 2023).

Effort: Participants were asked to assess how much effort they put into the task, on a scale from 1 to 5 (adapted from Ryan et al. 1983)

Optimistic AI productivity: On a scale from 1 to 10, we asked participants how optimistic they are about AI making higher education students more productive (adapted from Noy and Zhang 2023).

AI impact: On a scale from 1 to 10, we asked participants how they feel about the impacts of future advances in AI on higher education (adapted from Noy and Zhang 2023).

Education: We asked participants about their field of study and level of education.

Mother tongue: We asked participants what their mother tongue is (5% stated English).

Table 1

	Treatment				Control			
Measured after	First task		Second task		First task		Second task	
Gender			45F/24M				40F/31M	
Age			23.6	3.86			24.6	6.38
Variable	M	SD	M	SD	M	SD	M	SD
Case Experience	NA	NA	2.913	1.369	NA	NA	3.084	1.295
Felt effective	5.913	2.349	6.797	2.304	5.577	1.887	6.366	2.051
Felt enjoyment	6.565	2.172	7.057	2.120	6.098	2.224	6.549	2.068
Felt skilled	5.724	2.064	6.217	2.215	5.394	1.744	6.408	1.761
ChatGPT experience	NA	NA	5.666	2.888	NA	NA	5.169	3.018
Task evaluation/Grade	5.623	1.791	5.478	1.819	4.633	2.205	4.323	1.895
Useful aid	NA	NA	3.289	1.295	NA	NA	2.323	1.317
Effort	NA	NA	3.768	0.876	NA	NA	3.591	0.803
Optimistic AI productivity	NA	NA	6.927	2.475	NA	NA	6.239	2.411

AI impact	NA	NA	6.434	2.245	NA	NA	6.084	2.334

Table 1. Descriptive statistics. For some question data has been captured after both the first and second task. M signifies the mean, while SD signifies the standard deviation. NA means the question was not asked at that time. Data has been divided into treatment and control conditions.

3.1.3 Preliminary data analysis

From the complete dataset one participant was dropped due to technical issues in recording their data. Two more participants were excluded because their answers did not match the minimum requirements set by the pre-registration. Finally, after inspecting the records of the chatbot, two more participants were flagged because they had found a technical workaround and copy pasted the entire transcript[4]. Thus, the final sample size is 141 participants.

A correlation table (Table 2) showcases the correlation between all collected controls and the main dependent variable, performance in task 2.

[4] Results are very robust to these exclusions.

Table 2

Correlations with confidence intervals

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Felt enjoyment														
2. Felt skilled	.61** [.49, .70]													
3. Felt effective	.63** [.51, .72]	.70** [.60, .78]												
4. Felt enjoyment 2	.58** [.45, .68]	.41** [.27, .54]	.36** [.20, .49]											
5. Felt skilled 2	.36** [.20, .49]	.51** [.38, .63]	.36** [.21, .50]	.68** [.58, .76]										
6. Felt effective 2	.29** [.13, .44]	.42** [.27, .55]	.42** [.28, .55]	.62** [.51, .71]	.69** [.59, .77]									
7. Age	.05 [-.11, .22]	.02 [-.14, .19]	.08 [-.09, .24]	-.04 [-.20, .13]	-.08 [-.25, .08]	-.11 [-.27, .06]								
8. Case Experience	.06 [-.10, .23]	.20* [.04, .36]	.03 [-.13, .20]	.10 [-.07, .26]	.15 [-.02, .31]	-.01 [-.18, .16]	-.00 [-.17, .16]							
9. ChatGPT experience	.06 [-.11, .22]	.07 [-.10, .23]	.11 [-.05, .27]	.15 [-.02, .31]	.15 [-.02, .31]	.15 [-.01, .31]	-.21* [-.36, -.04]	.02 [-.15, .18]						
10. Useful aid	.02 [-.15, .18]	.07 [-.10, .23]	.18* [.01, .33]	.30** [.15, .45]	.23** [.06, .38]	.48** [.34, .60]	-.06 [-.22, .11]	.04 [-.12, .21]	.28** [.12, .43]					
11. Effort	.07 [-.10, .23]	-.04 [-.21, .12]	.03 [-.14, .19]	.04 [-.12, .21]	.01 [-.16, .17]	-.01 [-.18, .15]	.15 [-.02, .31]	.08 [-.08, .25]	-.05 [-.22, .11]	.09 [-.07, .25]				
12. Optimistic AI productivity	-.01 [-.18, .15]	.02 [-.15, .18]	-.02 [-.18, .15]	.29** [.13, .44]	.25** [.09, .40]	.20* [.03, .35]	-.10 [-.26, .06]	.10 [-.07, .26]	.44** [.30, .57]	.23** [.07, .38]	.09 [-.07, .26]			
13. AI impact	.01 [-.16, .17]	.03 [-.14, .19]	.10 [-.07, .26]	.23** [.07, .38]	.21* [.05, .36]	.21* [.04, .36]	-.08 [-.24, .09]	.04 [-.13, .20]	.36** [.21, .50]	.22** [.06, .37]	.02 [-.14, .19]	.78** [.70, .83]		

14. Task evaluation/Grade 1	.07	.18*	.09	-.16	-.13	-.05	-.03	.00	.02	.03	.14	-.08	-.15	
	[-.10, .23]	[.02, .34]	[-.08, .25]	[-.32, .00]	[-.29, .04]	[-.21, .12]	[-.19, .14]	[-.16, .17]	[-.14, .19]	[-.14, .19]	[-.02, .30]	[-.24, .09]	[-.30, .02]	
15. Task evaluation/Grade 2	.14	.18*	.14	-.05	-.01	.00	.07	.09	.14	.09	.01	.04	-.02	.46**
	[-.03, .30]	[.01, .34]	[-.03, .30]	[-.21, .12]	[-.18, .15]	[-.16, .17]	[-.10, .23]	[-.08, .25]	[-.03, .30]	[-.08, .25]	[-.16, .18]	[-.13, .20]	[-.18, .15]	[.32, .58]

Note. Values in square brackets indicate the 95% confidence interval for each correlation. Questions concerning enjoyment, skill and effectiveness were asked after both tasks and variables 4-6 thus refer to answers following completion of task 2.

3.2 Qualitative study methods

In order to gain deeper insights into the performance effects of engaging with chatbots, we initiated qualitative interviews with participants from the treatment group. This approach was chosen to capture nuanced insights into their experiences. To maximize the accuracy of their recollections, interviews were conducted within 24 hours post-experiment. Practically speaking, following the experiment, all participants from the treatment group were invited and we interviewed participants on a first-come-first-served basis. We stopped adding further participants to the interviews once new data began to echo earlier observations (Glaser & Strauss, 1967; Lund, 2014). Altogether we interviewed 18 participants, all of them students at the Business School; seven of which belonged to the low performer group (labeled L1 - L7 in the findings section) and eleven of which belonged to the high performer group (labeled H1 - H11 in the findings section).

The interview process was designed to foster open dialogue. These online interviews spanned a range of topics, including prior experience in working with LLMs, the participants' experience with tasks 1 and 2, their objectives for using LLMs in task 2, and their subjective assessments of the outcomes of tasks 1 and 2. The interview also explored preferences for working independently versus with chatbots. We expressly ensured all interviewees' anonymity to increase candor. The interviews lasted on average 25 minutes. All interviews were recorded and transcribed verbatim (173 single-spaced pages).

The interview data was coded in an abductive manner (Linneberg & Korsgaard, 2019), grounding the initial deductive coding in existing literature and experimental outcomes. This was followed by inductive analysis to deepen our understanding of the mechanisms influencing the varied performance outcomes observed among participants, particularly the factors contributing to the performance improvements in low performers and the performance decline in high performers.

4. Results: Experimental study

Figure 1 presents the distribution of performance scores for task 2 across control (red) and treatment (blue) conditions. The distribution indicates that participants in the treatment condition tended to obtain higher scores.

Figure 1

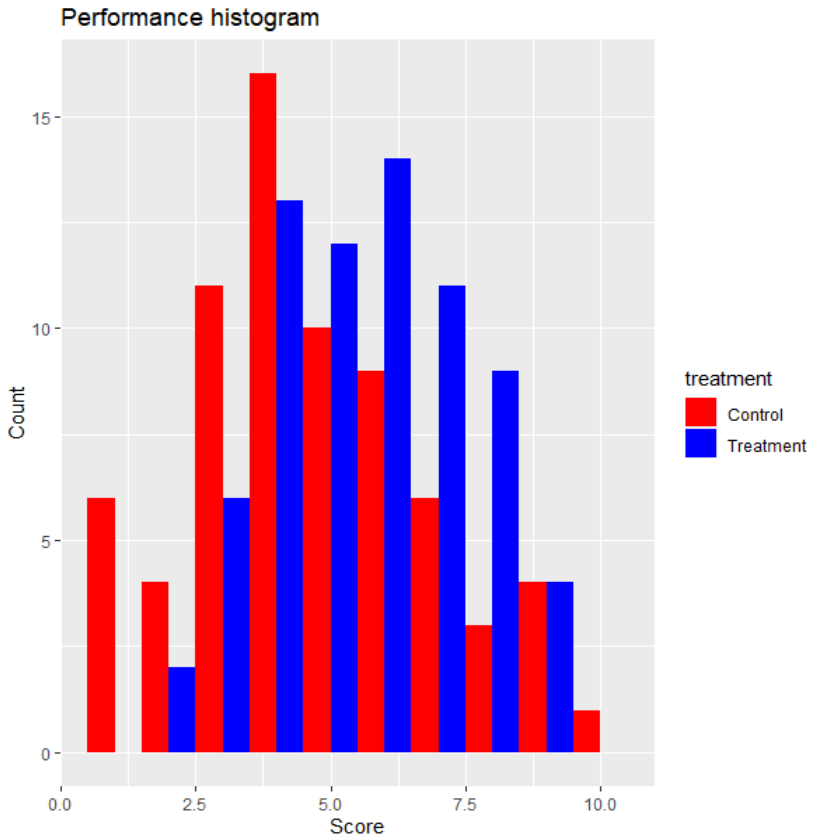


Figure 1: Performance in task 2 in control (red) vs treatment (blue) condition.

The first pre-registered hypothesis states that having access to ChatGPT-4 will improve performance. To test this, we first conducted a paired sample t test. We find support for our hypothesis (p=0.0014), i.e. that participants in the control group display a lower performance as compared to those in the treatment group (see also Figure 2.A). We then ran an OLS regression where we first included demographic controls (age, gender) and subsequently our independent variable (Table 3)[5]. Model 3 also includes the grade in the first period as a control, being that we expect performances to be correlated. In subsequent models we include additional controls. None of the demographic variables have a significant impact on the dependent variable. In addition, the additional task-related variables (such as case experience, perceived enjoyment, perceived skill or effectiveness) were also found to be poor predictors of our dependent variable. A similar trend pertains to the more general perspectives on AI (use of LLMs, AI impact, perception of AI in education).

For our second hypothesis, we anticipated that low performers (measured as performance in task 1)

[5] Mother tongue is not included since only 7 participants self-categorized as native English speakers.

would benefit more from using the chatbot compared to high performers. This is indeed confirmed by a paired t test ($p<0.0001$). Strikingly, high performers do not benefit at all from the use of chatbots ($p=0.866$). On the contrary, for all high performers we see a reduction in performance from task 1 to task 2 ($p<0.001$).

Table 3. OLS regression

	Dependent variable:		
	Task evaluation, period 2		
	(1)	(2)	(3)
Task evaluation, period 1			0.372***
			(0.071)
Treatment[6]		1.204***	0.823***
		(0.312)	(0.295)
Gender(M)	0.206	0.335	0.234
	(0.332)	(0.319)	(0.292)
Age	0.019	0.030	0.033
	(0.031)	(0.029)	(0.027)
Constant	4.414***	3.479***	1.707**
	(0.764)	(0.767)	(0.780)
Observations	141	141	141
R^2	0.006	0.103	0.253
Adjusted R^2	-0.009	0.083	0.231
Residual Std. Error	1.924 (df = 138)	1.834 (df = 137)	1.679 (df = 136)
F Statistic	0.389 (df = 2; 138)	5.241*** (df = 3; 137)	11.523*** (df = 4; 136)
Note:			*p<0.1; **p<0.05; ***p<0.01

To further explore these results, we fit a regression model where we include an interaction effect between a dummy variable (Low performers', coded 1 if performance in task 1 was less than or equal

[6] This variable does not just signify 'belonging to a treatment group' but whether participants in fact used the chatbot. Results are virtually unchanged for treatment allocation.

to 5, and 0 otherwise[7]) and the treatment. Results are shown in Table 4 and support the figure above: the difference between treatment and control seems to be driven by the gains in performance among low performers, while those who performed well in task 1 do not benefit from using the chatbot.

Table 4. Interaction effect

	Dependent variable:	
	Task evaluation, period 2	
	(1)	(2)
Low performers (yes)	-1.160***	-1.940***
	(0.306)	(0.396)
Treatment	0.787**	-0.186
	(0.307)	(0.443)
Low performers (yes)* Treatment		1.788***
		(0.600)
Constant	5.272***	5.767***
	(0.276)	(0.316)
Observations	141	141
R^2	0.152	0.203
Adjusted R^2	0.139	0.186
Residual Std. Error	1.777 (df = 138)	1.728 (df = 137)
F Statistic	12.323*** (df = 2; 138)	11.643*** (df = 3; 137)
Note:		*p<0.1; **p<0.05; ***p<0.01

[7] Results are qualitatively the same if we use 6 as a threshold, although of course, the interaction effect is somewhat reduced. Using 4 and 7 as threshold the interaction effect becomes insignificant at the conventional level of 0.05 but the direction remains unchanged.

Figure 2: Performance comparison in task 2

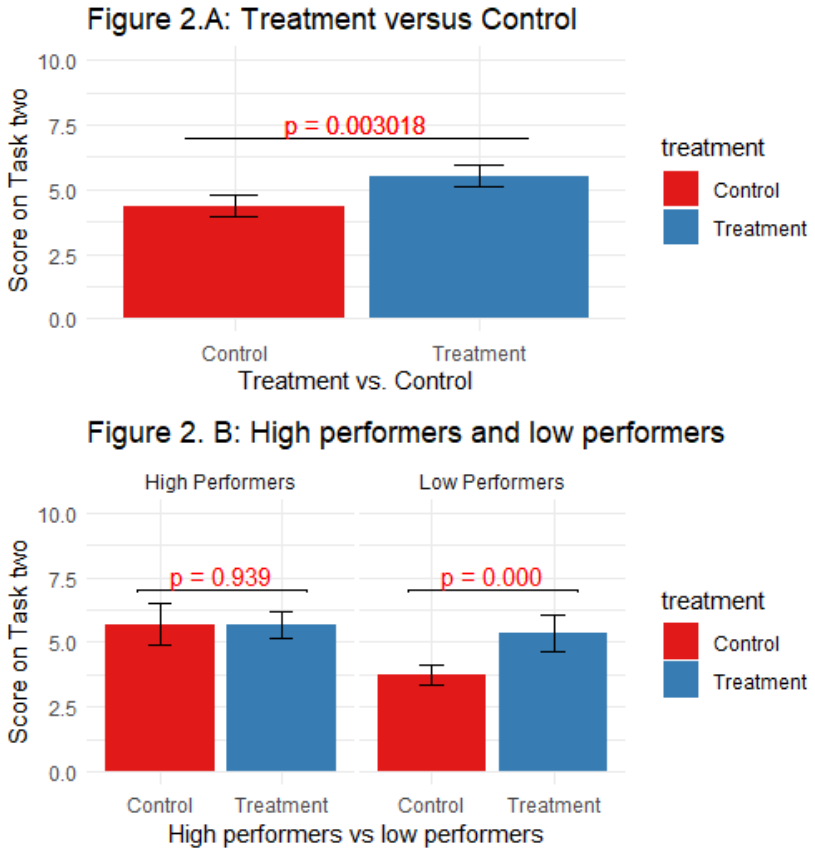


Figure 2: Figure 2a compares performance on task two across treatment (blue) and control (red). Figure 2b contrasts the performance in task two of those who scored in the bottom half versus the top half in the first task, with treatment (blue) and control (red) groups.

4.1 Exploratory results

Using the React webapp data that we collected, we constructed two additional measures: the number of prompts a participant has used[8], as well as the average number of words per prompt. Results show that good performance in task 2 is correlated with the number of prompts used (p: 0.039, Pearson correlation: 0.27) as well as the average number of words used pr. prompt (p: 0.075, Pearson correlation: 0.24, log transformed). As illustrated in Figure 3 below, the vast majority of participants used relatively few, and short prompts. Only 25% of participants had prompts averaging more than 26 words[9], while the median number of words per prompt was 15.

[8] After eliminating prompts like 'Hi' or 'Thank you!'

[9] For participants who used more than one prompt we have used the average number of words. Results are virtually unchanged when taking the maximum number of words across their prompts.

Figure 3: Frequency of prompts and nr. of words used

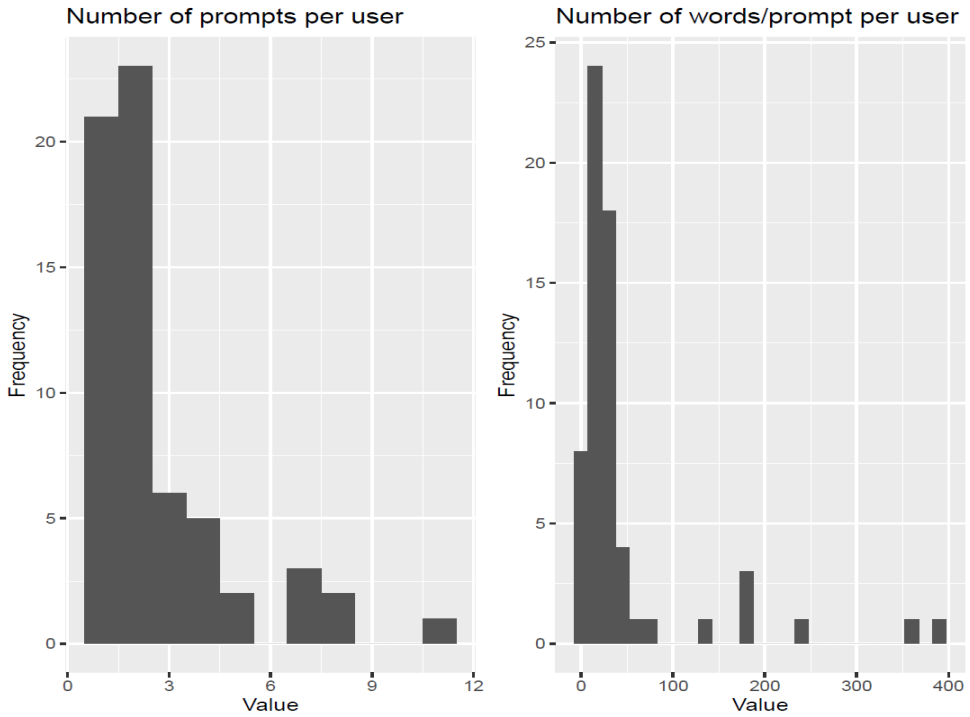


Figure 3: The figure on the left shows the number of prompts per user. The figure on the right shows the number of words per prompt per user.

4.2 Results: Qualitative study

4.2.1 Low performers

Low-performing participants encountered *significant challenges* with the initial task 1, expressing feelings of being "overwhelmed with task 1" (L3) and stating, "it was a bit hard (...) to begin and to do it" (L6). The sentiment of unease was further echoed by other participants who felt "a bit nervous" (L1), "frustrated" (L3) or "stuck" (L5), leading to a sense of disorientation, that one respondent described as, "I felt lost" (L6). Many expressed dissatisfaction with their outputs stating, "I'm not proud of what I've written there." (L6).

The introduction of the chatbot for the second task marked a significant turning point for these participants. They reported an *ease of entry into the task*, with one noting that the chatbot could provide a "helicopter view" (L3) and another stating that the chatbot "jumpstarted my process" (L6). Working with the chatbot was an experience like having access to "a catalyst" (L6) and many reported to use it "for everything" (L2).

The respondents used a *duplication approach* when it came to working with the output provided by the chatbot. They either “copy and modify it slightly” (L3) or simply “summarized [(the chatbot’s output]) it in my answer” (L5) The participants faced challenges in assessing the quality of the chatbot-generated output, students often followed their “intuition” (L3) or drew upon their own work experience (L1) i.e. “I also work in teams every day” (L5), not providing ground for significant changes of the chatbot output.

The consensus among low performers was a strong preference for working with the chatbot, attributing this to an improved workflow and that “I had more to say” (L1) and the “efficiency” (L4) it brought to the task. Furthermore, respondents expressed that a *sense of support* the chatbot offered was likened to “the feeling when your father is in your back, it gives you that feeling like somebody is with you” (L6), providing a safety net that made students feel more confident. Working with the chatbot made participants “feel more relaxed” (L1) and provided the feeling that they had a "plan B, plan C" (L6) at their disposal. In addition, they felt a heightened confidence from the “good grammar and rich vocabulary” (L4) and in general perceived “polished language” (L2) provided by the chatbot.

4.2.2 High performers

In contrast to the low performers, most of the high performers articulated that they found *joy* in engaging with task 1, valuing the thinking required to navigate through the complexity of the task and many feeling they delivered “a good answer” (H11).

The introduction of a chatbot in the second task was met with initial enthusiasm by many, who understood its potential to enhance “structure and efficiency” (H3) in their work as well as to provide “some inspiration” (H1), while another viewed it as a way to initiate the task “without a plan at the beginning” (H3). Despite these positive voices, some high performers also reported a sense of pressure to utilize the chatbot, due to concerns over a lack of performance and peer perception: “If I’m not working with the chatbot but the others are, I’ll not perform as good and will look bad” (H7).

High performers, in contrast to low performers, aimed for a *co-creation approach*, trying to blend their own ideas with the chatbot’s contributions with the aim of “teamwork makes dream work” (H10). However, this approach was not without its challenges. The sheer volume of information

provided by the chatbot was often overwhelming: “the chatbot started listing quite a lot of stuff” (H8), leading to “information overload” (H7). This disrupted the workflow of the high performers, who found themselves unable to effectively evaluate the well-sounding chatbot suggestions and integrate their own thoughts. As a result, many participants ended up with case answers that closely mirrored the chatbot’s output, similar to the *duplication approach* seen among low performers. One participant acknowledged, “I was simply modifying the way that the chatbot presented the output” (H6), while another noted the chatbot played a dominant role in their submission: “basically it’s 99% ChatGPT” (H1).

Despite recognizing the chatbot’s potential to “work more efficiently” (H5) and elevate their written language to a more “scientific level” (H4), many high performers ultimately expressed a preference for working without the chatbot. Working with the Chatbot raised significant concerns in this group of respondents regarding ownership and personal achievement. Participants questioned whether the final output could truly be considered their own creation, with one stating, “I didn’t feel like the author” (H7). This detachment from the process and outcome led to feelings of diminished accomplishment, as reflected in sentiments like “I feel I have accomplished something in task 1, but not task 2” (H8); “I don’t think I learned (while performing task 2)” (H5), and “I didn’t challenge myself” (H1).

5. Discussion

Our experimental study assesses the impact of LLMs on student performance , specifically within a business case analysis exam situation in a business school. Initially, we hypothesized that access to an LLM like ChatGPT-4 would notably improve performance more for lower-performing students compared to their higher-performing peers. Our results supported this hypothesis but also revealed an unexpected substantial decline in the performance of high performers, to the same level as the task 2 performance of low performers.

The observed decline in performance among high performers suggests an 'equalizing effect,' differing from the more positive 'democratizing effect.' A democratizing effect implies that both low and high performers experienced an uplift, with low performers seeing a more significant boost (Noy & Zhang, 2023; Brynjolfsson et al., 2023; Doshi and Hauser 2023). However, our findings reveal a different

dynamic where equalization results from both performance gains among low performers and decreases among high performers. This indicates the risk of an imbalanced democratization process.

The nature of the tasks in the various studies is likely to help explain the different dynamics. In Dell'Aqua et al. (2023), participants in the study received a fairly detailed challenge, divided into 18 separate and detailed questions. A user could simply copy-paste the relevant information to the ChatGPT-4 interface and obtain a fairly high-quality answer, and a one-shot approach was sufficient (see also Choi and Schwarcz 2023). In fact, Dell'Aqua and colleagues (2023) note that not further modifying the chatbot output, on average, led to the best result. In contrast, in our setup high performers did not improve when getting access to the chatbot, indicating a different dynamic. As Otis et al. (2023) argues, the nature of the problem shapes how frictionless it is to integrate LLMs. When an initial one-shot approach leads to an average level output, LLMs like ChatGPT-4 can substantially support lower-performing students in open-ended case analysis involving well-established concepts such as intrinsic and extrinsic motivation or individual vs. team dynamics. In our case, when the chatbot received the question posed in task 2, it produced an answer that consisted of about 11 items that were relevant and structured, potentially improving low performers output. It did generally not involve hallucinations or outright flaws. Yet, the chatbot's response lacked depth, context and complexity. It also did not at all manage to prioritize which of these perspectives were more important. Therefore, high-performing students that aim for higher quality output may struggle to integrate the relatively long chatbot-based output with their own ideas, since they did not have an effective working process to interact with and exploit the technology (Lodge et al. 2023).

We doubt that the particular equalizing effect will disappear when even more advanced LLMs emerge (e.g. ChatGPT-5). While benchmarks (Chiang et al. 2024) show that e.g. ChatGPT-4o (Autumn 2024) outperforms ChatGPT-4 (from Autumn 2023), it is not our experience that the performance improvement would substantially change the equalizing effect we identify.

To counter the equalizing effect, students need to establish work processes that enable them to leverage the advantages of LLMs, which also entails understanding under which circumstances (for example time pressure) or for which tasks (for example researching niche topics that LLMs know little about) working with LLMs like ChatGPT-4 is disadvantageous (Mollick 2024). This necessitates

that universities proactively design educational strategies that enable students to enhance their (metacognitive) skills in working with LLMs. Such initiatives should aid students in assessing the appropriate contexts for utilizing LLMs, as well as recognizing situations where their use may not be suitable. The effectiveness of traditional educational approaches in this new context remains uncertain. Just as on-the-job learning shifted due to the integration of intelligent machines (Beane 2019), it is plausible that innovative educational methodologies may need to be developed to meet these emerging demands as the introduction of LLMs will change teaching and research itself too.

Linked to this HEI responsibility is the finding that more detailed and frequent prompts lead to better performance. This finding implies that students who are able to ask more nuanced questions or provide longer, more nuanced prompts are likely to engage more effectively with the LLM, achieving the co-creation approach that is likely to be effective (Mollick 2024; Lodge et al. 2023). Prompting is important, not as a technical skill, but as a requirement to actively engage with the task. When students explicate their thoughts in longer prompts, writing the prompt becomes a way of thinking through the task. This might especially be important for non-native speakers who often do not experience that they can express their thoughts in a first draft eloquently. When writing the prompt in a more colloquial way, knowing that nobody sees their writing, prompting might enable fruitful thinking. In any case, it is clear that the clear majority of students, at the time of data collection, still needed additional, extensive training and experience with how to engage with LLMs in a meaningful and effective way.

5. 1 Limitations

The data was collected in Autumn 2023 at a Nordic Business School and relied on the use of ChatGPT-4 to analyze a case in a particular time-constrained scenario. The conclusions might be limited to similar technological and educational setups as well as the particular sample and point in time we relied upon. We could not detect any relation between the performance outcome and any of our collected individual variables, such as demographics, prior experience with LLMs or attitude towards LLMs. While Denmark was an early adopter of the technology, students generally did not have the opportunity to gain proficiency in using tools like ChatGPT-4. This lack of proficiency is likely related to the fact that at the time, the university still banned the use of LLMs in exam-related

work[10]. More advanced versions of large language models or more sophisticated models of implementation might substantially change the quality of the outcome produced by a chatbot. However, given that our main findings are shaped by students being overwhelmed by the output of the chatbot, we think this limitation is less severe.

The motivation and ability of students to engage with LLMs, both in terms of how to prompt it as well as how to iteratively engage with a LLM is likely to change over time, since students in general improve their ability to work with new technology. Students that have received more in-depth and advanced education on how to engage with chatbots might be better suited for using it, irrespective of their expertise on a given topic or their ability to frame a case analysis. The limitation thus also speaks to the future challenge of how to ensure an effective integration of LLMs into educational realms; how to use LLMs should be a part of educational offerings, either implicitly in most courses or in distinct courses.

In our setup we aimed to emulate a fairly typical on-site exam scenario at a Nordic business school, where students under time pressure are to answer fairly open-ended questions. If students had more time to cope with the chatbot output, as they would in a semester- or even week-long project, they might have had the opportunity to process and integrate the output. Our findings thus most directly speak to such open-ended on-site exam situations, rather than multiple choice scenarios (cf. Choi and Schwarcz 2023), well-defined challenges (cf. Dell'Acqua et al. 2023) or longer project durations.

Notably Dell'Acqua et al. (2023) identified a challenge ChatGPT-4 could not assist with, and one might speculate if our setup emulated that particular, atypical scenario. We doubt this to be the case, since our case description did not contain any internal inconsistency (as Dell'Acqua et al. 2023). Rather, it was designed to represent a typical exam case, akin to what bachelor students would face in an Organizational Behavior exam. The overall implication of ensuring that students have the competencies and motivation to engage with LLMs still hold, yet the particular scope and nature of the challenge might vary.

Finally, the fact that almost all of our participants were non-native English speakers (about 95%), might have shaped their approach and ability (Herbold et al. 2023) to engage with the chatbot, as well

[10] This ban was lifted in 2024.

as how to use the chatbot output to analyze the given case. The participants could engage with the chatbot in their respective mother tongue, but the final submitted product had to be in English. However, prior experimental studies tended to rely on native or at least very proficient English speakers (e.g. Choi and Schwarcz 2023; Doshi and Hauser 2023; Noy and Zhang 2023, Prather et al. 2024) and our sample of student participants that were not native English speakers thus opens up further generalization and theorizing.

Overall, while the above listed factors limit the potential to generalize our findings, we do believe the mechanisms uncovered speak to a broad range of populations and situations.

6. References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Beane, M. Learning to work with intelligent machines (2019) *Harvard Business Review* 97.5: 140-148.

Boussioux, Leonard, et al. (2023) The crowdless future? generative ai and creative problem solving. *Organization Science* (2023).

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative AI at work* (No. w31161). National Bureau of Economic Research.

Chiang, Wei-Lin, et al. (2024) "Chatbot arena: An open platform for evaluating llms by human preference." *arXiv preprint arXiv:2403.04132*

Choi, J. H., & Schwarcz, D. (2023). AI Assistance in Legal Analysis: An Empirical Study. *Available at SSRN 4539836*.

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in education and teaching international*, *61*(2), 228-239.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage

Cui, Zheyuan and Demirer, Mert and Jaffe, Sonia and Musolff, Leon and Peng, Sida and Salz, Tobias, The Effects of Generative AI on High Skilled Work: Evidence from Three Field Experiments with Software Developers (September 03, 2024). Available at SSRN: https://ssrn.com/abstract=4945566 or http://dx.doi.org/10.2139/ssrn.4945566

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Doshi, A. R., & Hauser, O. (2023). Generative artificial intelligence enhances creativity. *Available at SSRN*.

Extance, A. (2023). ChatGPT has entered the classroom: how LLMs could transform education. *Nature*, *623*(7987), 474-477.

Glaser, B., & Strauss, A. (1967). *The Discovery of Grounded Theory: Strategies of Qualitative Research*. Wiedenfeld and Nicholson.

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, *13*(1), 18617.
Koivisto, M., & Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific reports*, *13*(1), 13601.
Larson, Barbara Z., et al. "Critical Thinking in the Age of Generative AI." *Academy of Management Learning & Education* ja (2024): amle-2024.
Lodge, J. M., Yang, S., Furze, L., & Dawson, P. (2023). It's not like a calculator, so what is the relationship between learners and generative artificial intelligence? *Learning: Research and Practice*, *9*(2), 117-124.
Lund, C. (2014). Of What is This a Case? Analytical Movements in *Qualitative Social Science Research*. Human Organization, 73(3), 224–234.
Malmström, H., Stöhr, C., & Ou, A. W. (2023). Chatbots and other AI for learning: A survey of use and views among university students in Sweden. *Chalmers Studies in Communication and Learning in Higher Education*, *1*(10.17196).
Merali, Ali. "Scaling Laws for Economic Productivity: Experimental Evidence in LLM-Assisted Translation." *arXiv preprint arXiv:2409.02391* (2024).
Mollick, E. (2024). *Co-Intelligence*. Random House UK.
Mollick, E., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. *arXiv preprint arXiv:2306.10052*.
Moser, C., Den Hond, F., & Lindebaum, D. (2022). Morality in the age of artificially intelligent algorithms. *Academy of Management Learning & Education*, *21*(1), 139-155.
Nie, A., Chandak, Y., Suzara, M., Malik, A., Woodrow, J., Peng, M., ... & Piech, C. (2024). *The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement but Increased Adopters' Exam Performances* (No. qy8zd). Center for Open Science.
Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187-192.
Otis, N., Clarke, R. P., Delecourt, S., Holtz, D., & Koning, R. (2023). The uneven impact of generative AI on entrepreneurial performance. *Available at SSRN 4671369*.
Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of ai on developer productivity: Evidence from github copilot. *arXiv preprint arXiv:2302.06590*.
Prather, James, et al. "The Widening Gap: The Benefits and Harms of Generative AI for Novice Programmers." Proceedings of the 2024 ACM Conference on International Computing Education Research-Volume 1. 2024.
Ryan RM, Mims V, Koestner R (1983) Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. J. Pers. Soc. Psychol. 45(4):736–750.
Salomon, Gavriel, David N. Perkins, and Tamar Globerson. "Partners in cognition: Extending human intelligence with intelligent technologies." *Educational researcher* 20.3 (1991): 2-9.
Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American psychologist*, *26*(2), 145.
Skjott Linneberg, M; Korsgaard, S (2019) Coding qualitative data: a synthesis guiding the novice. Qualitative Research Journal, 19(3): pp. 259-270.
Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043), 776–778.
Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic literature review. British Journal of Educational Technology (early release).