Will Artificial Intelligence Get in the Way of Achieving Gender Equality?*

Daniel Carvajal† Catalina Franco‡ Siri Isaksson§

March, 2024

Abstract

The promise of generative AI to increase human productivity relies on developing skills to become proficient at it. Women and men may use AI tools differently, which could result in productivity and payoff gaps in a labor market increasingly demanding knowledge in AI. Thus, it is important to understand any gender differences in AI-usage among current students. We conduct a survey at the Norwegian School of Economics collecting use and attitudes towards ChatGPT, a measure of AI proficiency, and responses to policies allowing or forbidding ChatGPT use. Three key findings emerge: first, female students report a significantly lower use of ChatGPT compared to their male counterparts. Second, male students are more skilled at writing successful prompts, even after accounting for higher ChatGPT usage. Third, imposing ChatGPT bans widens the gender gap in intended use substantially, while allowing ChatGPT closes the gender gap. We provide insights into potential factors influencing the AI adoption gender gap and highlight the role of appropriate encouragement and policies in allowing female students to benefit from AI usage, thereby mitigating potential impacts on later labor market outcomes.

JEL Codes: I24; J16; J24; O33
Keywords: Artificial intelligence, ChatGPT, gender, education, technology adoption

*We would like to thank Andrea Bocchino, Finn Casey, Nisvan Erkal, Iver Finne, María Recalde, Erika Povea, Stig Tenold, Heidi Thysen, Bertil Tungodden and Tom Wilkening. Audiences at the ASSA 2024 meeting, Bergen-Berlin Behavioral Economics Workshop, FAIR, King's College, NHH, U. of Strathclyde and U. of Melbourne provided useful comments and feedback. This study was approved by the NNH IRB (NHH-IRB 42/22) and pre-registered in the AEA RCT registry.

†FAIR at NHH Norwegian School of Economics, Bergen, Norway. Email: Daniel.Carvajal@nhh.no

‡Center for Applied Research (SNF) and FAIR at NHH Norwegian School of Economics, Bergen, Norway. Email: Catalina.Franco@snf.no

§FAIR at NHH Norwegian School of Economics, Bergen, Norway. Email: Siri.Isaksson@nhh.no

1 Introduction

Within a year of its release, ChatGPT has already left a mark. Companies have expressed interest in candidates with knowledge of how to use the tool (CNBC, 2023), and new well-paid jobs as "prompt engineer" are quickly emerging (WSJ, 2023). Recent studies indicate that the use of artificial intelligence (AI) tools such as ChatGPT provides substantial productivity gains across domains. For instance, allowing access to AI tools improved output quality in professional writing tasks among online workers by 18% (Noy and Zhang, 2023), increased solutions to issues in real-life customer support tasks per hour by 14% (Brynjolfsson et al., 2023), and reduced the time developers used to complete a coding task by 56% (Peng et al., 2023). Although exact economic impacts are hard to predict and depend on the policies adopted (Brynjolfsson and Unger, 2023), AI proficiency is likely to shape labor market paths and success in the near future. Therefore, it is crucial to assess the adoption and use of these new technologies by students facing this fast-paced labor market, particularly amidst the current heated debate on whether to allow or ban the use of ChatGPT.

This paper focuses on a potential disparity in adoption and proficiency in ChatGPT use based on gender, a side overlooked so far in the debate. Gender likely plays a role in AI adoption based on previously documented gender disparities in internet usage (the so-called "gender digital divide") (Bimber, 2000; OECD, 2018), in technology-related career choices (Buser et al., 2014, 2017; Cimpian et al., 2020), and in confidence regarding skills in male-dominated tasks and the prevalence of stereotypes (Bordalo et al., 2016). Using a survey experiment on university students in Norway, we find substantial gender differences in both adoption and proficiency of ChatGPT usage. We also identify potential explanatory factors influencing this gender gap.

The rapid growth and unprecedented capabilities of ChatGPT and other generative AI technologies have raised concerns among educational institutions, prompting calls for regulatory measures regarding its use. Varying policies have been proposed ranging from outright bans to embracing and incorporating AI tools in the learning process. Those arguing that it should be banned cite fears about students submitting inauthentic and potentially

plagiarized work, substituting the development of critical and problem-solving skills as students get easy and quick answers, and information privacy concerns. Supporters of embracing ChatGPT believe that generative AI is here to stay and should be incorporated into the classroom to guide students to make productive use of it.[1] We believe that gender should be a crucial aspect in the debate on whether to ban or allow the use of AI by students, as differential responses may unintentionally create gender-biased policies. With this aim, our survey also provides evidence of large differences in how female and male students would respond to university bans of the tool, and shows that imposing university bans on the use of ChatGPT widens the gender gap in intended use substantially.

We conducted a preregistered anonymous survey experiment with 531 students at the Norwegian School of Economics (NHH) in November 2023.[2] Participants in the study were recruited in class from the first and third year cohorts of the bachelor's program, as well as from the master's program. We collected student self-reports on current ChatGPT usage and measured prompting skills. We preregistered the hypotheses that perceptions about ChatGPT, preferences about its use, and their experience and exposure could constitute potential factors that influence adoption and skills, and collected measures accordingly for each factor.[3] Crucial to understanding potential differential responses to policies that allow or forbid ChatGPT, we included a vignette experiment describing a course that students would hypothetically be enrolled in. Keeping all other information constant, the description randomly displayed whether the professor explicitly allowed or forbade the use of ChatGPT in the course, and the students were asked to report their intended use of ChatGPT throughout the course.

We report three main findings. First, female students are much less likely to use ChatGPT than male students. A higher proportion of women than men report having heard of ChatGPT but do not use it (10.8% vs. 2.4%), and having used it only a few times (31.5% vs. 22.8%). Furthermore, the proportion of women who report using it all the time is almost half

[1] See Lo (2023) for a review of the advantages and disadvantages of ChatGPT use in education.

[2] NHH is the most selective higher education institution in Norway. We consider anonymity to be important because we want to obtain truthful responses and not responses that reflect what students think should be correct if they knew that we were matching the survey responses to their academic record.

[3] Besides reporting raw gender gaps, we add these factors as well as other baseline variables to examine the extent to which they explain the gaps.

that of men (26.6% vs. 44.8%). Overall, the raw gap in high ChatGPT use (occasionally or all the time) is 16.8 percentage points (pp) or 30% over a base of 57.7% among women. This estimate does not reflect gender differences in course selection, as the curriculum is largely fixed within the NHH cohorts. In addition, including year in college and admission grade, as well as measures of risk and time preferences, reduces the gender gap to 9.4 pp. That is, these baseline characteristics explain about 43% of the raw gap. Adding the full set of potential factors influencing adoption, capturing perceptions, preferences, and experience/exposure related to ChatGPT further reduces the gap to 1.7 pp, which is not statistically different from zero.

Second, men are more skilled at ChatGPT prompting than women even after controlling for baseline use. We measure proficiency by asking students to write a prompt that we then feed into ChatGPT to assess whether it gets the correct answer.[4] Although there is no significant gender gap in the time spent writing the prompt (132 seconds on average), we find gender gaps of about 0.3 standard deviations (SDs) in the variables measuring the number of characters written (141 and 179 on average for women and men, respectively), and success rate of the prompt (25% for women and 36% for men on average).[5] Adding the full set of controls, including the potential factors influencing ChatGPT use, does not reduce the gap substantially. However, controlling for the number of characters and the keywords most correlated with successful prompts reduces the gap to a non-statistically significant 2.6 pp.

Third, the gender gap in the intended use widens if ChatGPT is banned and closes if allowed. About 80% of both female and male students randomized into the professor "allows" ChatGPT treatment state that they would use it in the course. However, women in the professor "forbids" ChatGPT treatment are 38 pp less likely to use it than women in the "allows" treatment. A gender gap equal to 20 pp opens up, since men's intended use is much less likely to be affected by the bans. The gender gap in intended use and the within-gender reaction to the policy is virtually unaffected by adding the full range of control variables and potential factors influencing adoption mentioned above. Hence, we conclude that other aspects

[4]Large Language Models (LLMs) provide different answers each time a prompt is submitted, which in some cases might be correct or not. Therefore, we run each prompt over 100 times and collect how many times the prompt gives the correct answer.

[5]The fractions of women and men with at least a 50% success rate are 25% and 37%, respectively.

that might vary by gender, such as rule-following behavior, obeying the authority, or having trust in the professor's recommendation since they know what is best for students, must be behind the differences in intended use after the policy. Most importantly, this shows that banning ChatGPT in the classroom might have large unintended consequences by putting female students behind their male peers in AI adoption. Although the effects of generative AI on learning and how prepared adopting and non-adopting students will be for the labor market are still unknown,[6] our results inform educational institutions that are currently enacting AI policies on the expected gendered responses to such policies.

Finally, we discuss additional descriptive findings in light of the existing literature on gender differences in choices. Using self-reported admission grades,[7] we look into differences in use and reactions to the hypothetical "allow/forbid" policy by admission grade quintile. While men across all grade quintiles have similar usage rates (gravitating around 80%), women in the top quintiles are much less likely to be currently using ChatGPT (around 40%) relative to women in the bottom two quintiles (over 80%) and than men in any quintile. The responses to the policy are quite similar across admission grade quintiles for men, while only similar across quintiles for women in the "allow" treatment. In the "forbid" treatment, women's intended adoption rates are much smaller in the top quintiles.[8]

We note the resemblance between our findings by admission grade quintile and previously documented patterns of top women, in particular, exhibiting behaviors most dissimilar from men. For example, men are less sensitive to the grade they obtain in a principles class when deciding whether to major in the same field as that class. Women are much more sensitive, with only the women earning the highest grades in the principles class declaring a major in the same field (Rask and Tiefenthaler, 2008; Ost, 2010; Goldin, 2015; Avilova and Goldin, 2018; Kugler et al., 2021; Ugalde, 2022). Niederle and Vesterlund (2007) find that

[6]For example, the use of ChatGPT could hinder critical and problem-solving skills or help develop them further. Acquiring AI skills at the expense of critical and problem-solving skills will probably not help students in the labor market. On the other hand, if acquiring AI skills does not hinder other labor-market relevant skills, differential adoption will likely cause slow adopters to fall behind by missing learning and career opportunities.

[7]Admission grades are based on high school GPA and retakes of courses for students who do not get in on their first attempt. 273 of 514 students provided valid admission grade responses, which prevents us from doing a full heterogeneity analysis. There are no gender differences in the likelihood of reporting the admission grade nor in the reported grade.

[8]We also see that women who do not use ChatGPT at baseline have much larger reactions to the "forbid" policy than women who already use it occasionally or all the time and than men in all usage categories.

women in the top performance quartile are willing to compete to a similar extent as men in lower quartiles, and Coffman (2014) finds that expert women are less likely to speak up. We document similar patterns in a completely new domain: using ChatGPT and responses to policies on ChatGPT use, a skill that is becoming increasingly relevant for labor market success. Crucially, we contribute to the previous literature by showing that top women may be willing to adopt behaviors in which there are ex-ante gender differences through a change in policies or recommendations that alter how the behavior is portrayed. If women are disproportionately affected by negative portrayals of what it means to choose a major when one's grade was not among the best or what it means to enter a competition when one's chances of winning are not the highest, they may simply opt out.

Our findings suggest that more positive portrayals (it is okay/allowed to apply/compete even if you fail) by an authority figure (e.g., a professor) may go a long way in closing gender gaps in choices. Moreover, we believe that gender differences in rule-following or trusting advice from authority opens up a new agenda of research in the topic of gender and behavior.

An additional implication of our results is that recent findings on how using AI recruitment tools increases gender diversity in the workplace (Avery et al., 2023) may be attenuated by women not having the requirements to apply for the increasing number of jobs that require AI skills. If women develop AI skills to a lesser extent than men while in college, as we document, the prospect of increasing gender diversity with debiased recruitment (Pisanelli, 2022; Awad et al., 2023) may be harder to attain.

2 Setting and Research Design

2.1 Participants and recruitment

Participants in our survey are recruited from the first and third year of the bachelor's and master's programs at NHH. The school offers a five-year program consisting of three years of a bachelor's program in economics and business administration followed by two years of a master's program in either economics and business administration or international management. Education is free and students who are admitted into the bachelor's program au-

tomatically get a slot for the master's programs and typically continue with the master's, but can leave after completing the three years of the bachelor's program only.[9]

The bachelor's program at NHH is the most popular program in Norway listed as the first choice of most applicants to higher education. In 2023, it was listed as a first choice by 2,170 applicants who competed for 500 slots.[10] 50% of admitted students come straight from high school (first-time admission) and the other 50% usually retake some subjects or do some activities after graduating from high school that grants them higher admission points to be more competitive in the admission process. The 2023 admission cutoffs for the first-time admission and regular admission were 55.6 and 59.5, respectively. For reference, grades in Norway go from 1 to 6, and GPAs are calculated from high school grades and the score in five to six exams taken throughout high school (Landaud et al., 2023). The cutoffs, calculated by multiplying the GPA by 10, illustrate that successful applicants in both admission categories typically achieve scores close to a perfect 6 in every school and exam subject.

In the bachelor's program students take 4 subjects every semester, for a total of 24 of which only 6 are elective.[11] Subjects in the master's programs involve 6 subjects and a master's thesis, where at least 3 of the 6 subjects must be selected from a list of mandatory subjects. We believe that the small role of elective courses, particularly in the bachelor's program, make a strong case that our results are not simply driven by gender differences in the choice of subjects that are more or less amenable to the use of ChatGPT.

Students participating in the survey were recruited during lecture hours of two of the mandatory courses of the bachelor's program: a first-year and a third-year course, as well as one of the core courses in the master's program. The survey experiment was preregistered in the AEA RCT Registry (AEARCTR-0012452) and the data was collected subsequently in November 2023. The survey was anonymous and implemented in the classroom using a QR code. Students lasted on average 7 and a half minutes responding the survey.[12]

[9]The bachelor's program is taught in Norwegian, while the master's programs are taught in English.
[10]Almost 5,000 applicants listed the NHH program in any rank of their list. There were 62,757 higher education slots in Norway in 2023 (Direktoratet for høyere utdanning og kompetanse, 2023).
[11]There are no electives in the Autumn semester of the first year (where half of our sample is recruited from), and one elective thereafter except in the last semester of the program in which students can choose two electives.
[12]On average, women spend 7.7 minutes and men 7.3 minutes. The difference is not statistically significant.

2.2 Anonymity and Participant Incentives

In considering the best format to administer the survey, we weighed the prospect of linking student responses to their future academic performance with the potential for misrepresentation of ChatGPT use and experimenter-demand effects if students knew that the survey was not anonymous. Since this is the first study documenting patterns in student use of ChatGPT, we opted for anonymity as we put the highest value on truthful responses.

Related to anonymity, incentivizing the prompting exercise and second-order beliefs questions would have required collecting some personal information to provide incentives. We also opted for conducting the survey in the classroom to avoid students getting external help (from someone else or from ChatGPT) to get the prompt correct.

Validation exercises have found strong similarities in the use of hypothetical and unincentivized measures relative to incentivized elicitations and real-world behavior across different domains (Hainmueller et al., 2015; Brañas-Garza et al., 2021, 2023; Enke et al., 2022; Falk et al., 2023). At the same time, there has been an increase in the use of unincentivized measures in economics research (Ameriks et al., 2020; Bernheim et al., 2022; Stango and Zinman, 2023; Almås et al., 2023; Andre et al., 2022). Given the restrictions in our scenario and the concerns over potential effects of incentives on reporting actual capabilities, we opted for the use of unincentivized questions.

2.3 Survey design

The survey consists of four sections: background characteristics, a hypothetical vignette experiment, a prompting skills task, and a questionnaire on the use and attitudes about ChatGPT, presented to the respondent in that order. The questionnaire is in Appendix C.

Background characteristics. First, participants were asked questions on demographic and academic background, including gender, whether the student is from Norway, risk and time preferences measured through survey questions following Falk et al. (2018). Students were given the possibility of reporting or not their admission grade, with 273 students reporting valid responses out of the 514 respondents (53% of the sample).

Use and attitudes about ChatGPT. Participants indicated self-reported use of ChatGPT.

Our baseline use outcome is obtained from the question "*How familiar are you with ChatGPT?*", with choices corresponding to *low use* if the participant indicated: "not heard about it", "heard about it but not using it myself", or "used it a few times", which indicates none or limited use; and *high use* if the participant indicated "use it occasionally", or "use it regularly", which indicates continuous use. Participants also selected the types of tasks they "typically ask ChatGPT to help with".

Prompting skills measure. To measure a participant's skill in the use of ChatGPT, we presented participants with an image of the "Ebbinghaus illusion",[13] and asked them to write in a text box the query/prompt they would provide to ChatGPT to arrive at the correct official name of the visual phenomenon represented by the image. We use three outcome measures based on the prompting exercise: time spent writing the prompt, the number of characters written, and the success rate of the prompt, given by the proportion of ChatGPT answers that mention the official name out of over 100 queries made, for each prompt.

Potential factors influencing usage. We elicited attitudes of respondents regarding ChatGPT, which we preregistered and classified into three categories of primary factors affecting ChatGPT usage: (i) preferences, (ii) perceptions, and (iii) exposure/experience. In terms of preferences, we aim to measure potential utilitarian costs or benefits associated with ChatGPT usage and examine the role of persistence in the use of technology. Concerning perceptions, we focus on four key areas: perceived usefulness of ChatGPT, whether ChatGPT usage is considered cheating, trust in the accuracy of information provided by ChatGPT, and confidence in one's abilities to use the technology. Lastly, we explore the exposure/experience factor, analyzing how prior exposure to ChatGPT might influence its adoption.

Hypothetical vignette experiment. Participants were presented with a hypothetical scenario, describing a course the participants are hypothetically enrolled in. The course description indicates how it is evaluated and we experimentally vary a statement of whether the professor explicitly allows or forbids the use of ChatGPT in the course as follows:

[13]The Ebbinghaus-Titchener illusion (Titchener, 1901) is represented by two circles of the same size that are surrounded by a different context each: the first circle is surrounded by small circles and the second circle is surrounded by big circles. When most observers view these figures, the context affects perceptions of size. The image used is presented in Appendix C.

"Imagine you are enrolled in a course on Environmental Policy and Economic Impact. This course explores the intersection of environmental regulations, economic incentives, and their effects on industry practices and sustainability. The professor explicitly allows/forbids the use of ChatGPT during coursework. It is an 8-week course with final evaluation given by a final in-person written exam."

Subsequently, the respondent was asked: *"Given this scenario, how likely are you to use ChatGPT throughout the course?"*, where the choice consists of indicating intended use in a 5-point scale from "Very unlikely" to "Very likely".

Participants stratified by gender were randomly allocated into one of two treatment conditions: (i) when the professor explicitly *allows* the use of ChatGPT, and (ii) when the professor explicitly *forbids* the use of ChatGPT. This allows us to causally study the effects of the allow/forbid policy on intended use. A second layer of randomization was the type of evaluation of the course, where the evaluation could be either an in-person exam or a home exam.[14]

2.4 Sample characteristics

Almost 55% of our sample is male, which is close to the historical male student representation at NHH of about 60% (Hirshman and Willén, 2022). 54% and 40% of the sample are in the first and third year of the bachelor's program, respectively. Men are statistically more willing to take risks and give up something beneficial today in order to benefit more from that in the future (Falk et al., 2018) than women in the sample. While only 53% of the sample provided a valid answer for their admission grade, there are no gender differences in the likelihood of reporting the grade or in the grade itself. On average, the admission grade is 5.6 (median equal to 5.7) for both men and women, and the distributions look quite similar.

[14]Respondents that were presented with the home exam scenario were asked a second question: *"Given this scenario, how likely are you to use ChatGPT during the final exam?"* This way, respondents would differentiate the use of ChatGPT throughout the course and during the exam in order to make the measures comparable across different evaluation scenarios. We are not using this layer of randomization in this draft.

2.5 Empirical Strategy

We use two main econometric specifications. For the outcomes related to baseline use and prompt success rate described in Section 2.3, we focus on estimating the gender gap using an indicator for whether the participant is a male student:

$$y_i = \alpha_0 + \alpha_1 \text{Male}_i + X_i\gamma + \varepsilon_i \tag{1}$$

We measure the gender gap through the coefficient α_1. In our main results table, we present the raw gap along with a series of controls X_i including baseline use (for the success rate outcome only), background characteristics, and preferences, perceptions, and experience as described above.

Our second econometric specification involves estimating the gender gap for the policy reaction to allowing/forbidding ChatGPT in the hypothetical course presented in the vignette experiment:

$$y_i = \beta_0 + \beta_1 \text{Male}_i + \beta_2 \text{ChatGPT forbidden}_i + \beta_3 \text{Male}_i \times \text{ChatGPT forbidden}_i + X_i\gamma + \epsilon_i \tag{2}$$

The outcome y_i is equal to 1 for students who state that they are likely or very likely to use ChatGPT during the course. The coefficient β_1 provides the gender gap when ChatGPT is allowed, β_2 represents the policy response (from allowed to forbidden) among women, and β_3 measures the differential change in the policy response for men relative to women. Similarly as in specification 1, we add different types of controls that help us understand the influence of the preregistered factors on our results.

3 Results

3.1 Female Students Are Less Likely to Use ChatGPT Than Male Students

We begin by analyzing the responses to the survey question *"How familiar are you with ChatGPT?,"* which contains 5 answer options: not heard, heard but not use, used few times, use

occasionally, and use all the time. Figure 1 shows the proportion of responses in each category split by gender, with the height of the bars adding up to 100% within gender. Women are much more likely to be represented in low use categories. 11.2% of women while 2.5% of men state that they have heard about ChatGPT but do not use it. 31.9% of women and 23% of men have used it a few times. Only 1 out of 514 students answered not to have heard about it. In contrast, men are overrepresented in the use all the time category with 44.3% of men relative to 25.4% of women. The proportions in the use occasionally category are similar with 31.5% of women and 29.8% of men.

We statistically estimate the gender gap in use through specification 1, where the outcome is a binary measure indicating a high use if the participant responds use occasionally or use all the time. Overall, the raw gender gap in high use at baseline is estimated at 17.2 pp or 30% over a base of 56.9% of women using ChatGPT occasionally or all the time (see Column 1 in Panel A of Table 1).

To understand the overall gender gap in use, it may be insightful to plot this variable by a measure of relative academic skill.[15] For example, given a level of skill, students may use ChatGPT less or more depending on how they think it complements or substitutes their own skills. Figures 4a and 4b show, for women and men separately, the fraction of students reporting a high use by quintile of the admission score distribution.[16] The fraction of men with high ChatGPT use (Figure 4b) is between 73% in the highest quintile up to 84% in the middle quintile, so it is quite homogeneous across quintiles. In contrast, the fraction of women with high ChatGPT is strongly and negatively correlated with admission grade quintile. In the bottom two quintiles, the fraction of women with high baseline use is similar to the fraction of all men, while for the three top quintiles, the fraction of women with high baseline use is below 45% (Figure 4a). A regression estimating the correlation between the raw admission grade and the high baseline use indicator yields a negative and significant coefficient for both men and women, but it is almost four times larger for women (-0.378) than for men (-0.097).

[15] Admission grades tend to be correlated with college GPA, which in turn increases hiring interest by employers (Kessler et al., 2019).

[16] Quintiles are calculated pooling men's and women's admission grades.

Our results are in line with previous findings suggesting a correlation between women's choices according to their position in the skill distribution in choices based on laboratory tasks (Niederle and Vesterlund, 2007; Coffman, 2014) and on the grade in a principle's class determining what college major students enroll in (Rask and Tiefenthaler, 2008; Ost, 2010; Avilova and Goldin, 2018; Kugler et al., 2021; Ugalde, 2022). In our setting, we find that top women engage less in ChatGPT use, a behavior that may be perceived as showing that they are not as qualified as they are.[17] If female students interpret ChatGPT as such, they may care more about how they will be perceived by employers down the line given the evidence that student beliefs about hiring decisions affect important decisions such as which college major to pursue (Ugalde, 2022). Women may opt for not using ChatGPT to avoid giving the wrong signal to fellow students, professors, or employers. However, we show in Section 3.4 that institutional policies on ChatGPT use can affect intended use and that women are as likely to intend using it as men under certain scenarios, which suggests that needing to prove themselves in college and to employers must not be playing a first-order role in the gender gap in use.

3.2 Male Students Are More Skilled at Prompting than Female Students

As mentioned before, proficiency in AI tools like ChatGPT is becoming an increasingly important skill for labor market success. We documented in the previous result that female students, in particular top women, are using ChatGPT to a lesser extent than male students. Lower use can directly impact proficiency since acquiring it probably results from continued use with a tool. We show that male students are more skilled at writing successful ChatGPT prompts than female students even after accounting for baseline use.

Figure 2 shows standardized versions of three outcomes measuring prompt quality: Time spent, success rate out of 108 runs of the same prompt on ChatGPT, and number of characters written. These variables are standardized relative to the mean and standard deviation of men in the sample. The stars next to each gap visualization correspond to the statistical difference in the raw gap.

[17] For example, Williams (2014) find that a majority of female scientists report in a survey that they feel the need to provide more evidence of their competence than others to prove themselves to their colleagues.

On average, men spend 132 seconds writing their prompt, and women spend less time, but not statistically significantly so. The average success rate recording the fraction of times that the prompt obtains the desired answer (Ebbinghaus illusion) is 36% for male students and lower by 11 pp or 0.26 SD for women (see also Table 1, Column 1 in Panel B). Lastly, female students write about 0.3 SD fewer characters in their prompt relative to a mean of 179 characters among male students. The success rate and number of characters seem to be strongly and positively correlated, as shown in Figure A1.

Table 1, Panel B, Column 1 quantifies the raw gap in prompt success rates. On average, women have a success rate of 24.9%, meaning their prompt gives the correct answer about 27 times out of 108 ChatGPT runs. The gender difference is estimated at 11.1 pp, which means that men get the correct answer about a third of the time on average.

As expected, in Figures 5a and 5b, students at the top of the admission grade distribution have higher success rates with their prompts regardless of gender. In the top quintile of the distribution, students have success rates of about 41-42%, which is in stark contrast to the overall average of 31%. As with the high baseline use outcome, men have more homogeneous success rates across quintiles than women. Even though women in quintile 1 have the highest ChatGPT baseline use, their success rate (17%) is the lowest among all and almost half of the success rate for men in the same quintile (30%), who have similar levels of baseline use.

3.3 Potential factors influencing adoption and skills

As discussed in section 2.3, we preregistered and measured three main categories of potential factors influencing adoption and use of ChatGPT: (i) preferences, (ii) perceptions, and (iii) experience or exposure. The results are summarized in Figure 7.

3.3.1 Gender Differences in Preferences

As preferences, we consider three factors. For the first two, we ask students to indicate their agreement or disagreement with the following two statements: *"I think ChatGPT is enjoyable to use"*, and *"I think ChatGPT is difficult to use"*, representing a utilitarian benefit and cost from using ChatGPT, respectively. The choices range on a 5-point scale from strongly agree

to strongly disagree. Figure 7b shows the percentage of students that indicated disagreement with the claim that ChatGPT is difficult to use, and agreement with the claim that using ChatGPT is enjoyable. While 63% of women find ChatGPT not difficult to use, and 68% find it enjoyable to use; in both cases, the percentage of men is higher than women, by 7 pp for disagreement that it is difficult to use, and by 12 pp for agreement that it is enjoyable to use, the latter being significant at the 1% level. This suggests that men have stronger preferences for the use of ChatGPT, as they find it more enjoyable (higher utilitarian benefit), and less difficult (lower utilitarian cost) to use than women.

We also measured what we refer to as "persistence", where the participants were asked *"If ChatGPT does not provide the desired answer on your first attempt, how many additional attempts do you typically make?"* with four options ranging from "One more try" to "I keep until satisfied".[18] We find that 58% of female students indicate that they attempt two more tries or more, compared to 73% of male students, a difference significant at a 1% level. This indicates that men are more persistent in attempting to obtain desired results than women, which can lead to gender differences in the use of ChatGPT, as men would be more eager to maintain longer "conversations" with ChatGPT for a specific query. Moreover, the gap in persistence could generate differences in skill, as men can learn more from the increased prompting experience.

We now aim to understand the relationship between the gender differences in responses to survey questions related to preferences in the use of ChatGPT, and the self-reported adoption and skills of the technology. In our regression analysis, incorporating preferences factors helps explain a significant part of the gender gap in ChatGPT use, with the gender gap coefficient, previously at 17.2 pp, now being 5.1 pp and not statistically significant (see Table 1, Panel A, Column 3). However, the same exercise on the success rate of the prompt (Panel B of Table 1) keeps the gap at around 10 pp, a similar level to the initial gap. This suggests that our measures of preferences seem to capture part of the gender gap in use, but not in their ChatGPT skills.

[18]We also allow participants to indicate that they do not use it, and these participants are excluded from the analysis of these covariates.

3.3.2 Gender Differences in Perceptions

We also consider belief-based motives that can affect behavior in our setting, which we categorize as perceptions. We consider four relevant sets of perception over the use of ChatGPT: (i) perceptions over its use as cheating, (ii) perceptions of its usefulness, (iii) trust in its accuracy in providing information, and (iv) confidence in one's own skills using it. The perceptions are illustrated in Figure 7a, which shows the percentage of participants who align with a series of statements, representing the different sets of perceptions.

First, students might not adopt the technology if they perceive its use is unethical/cheating. To measure this, participants were asked to indicate agreement or disagreement with the following two statements: *"Using ChatGPT as an aid to solve assignments in a course is equivalent to cheating"*, and *"Using ChatGPT as a learning aid in a course is equivalent to cheating"*, with options ranging on a 5-point scale from strongly disagree to strongly agree. Figure 7a shows the percentage indicating either strongly disagree or somewhat disagree with the statement. While the majority of participants disagree with considering the use of ChatGPT as equivalent to cheating, there are important gender differences, with around 13 pp more men disagreeing relative to women, a significant difference at the 1% level. However, it is important to highlight that when the use of ChatGPT is as a learning aid, 83% of participants disagree with the use being equivalent to cheating, relative to a 58% disagreeing when the use is as an aid to solve assignments. Moreover, a related statement to those on cheating in our survey is *"It is easy for professors to identify if a student has used ChatGPT"*, which measures the perceived risk of getting caught using ChatGPT. Figure 7a shows that while 43% of women disagree with the statement, the proportion of men that disagree is weakly higher (51%).

A highlighted factor in previous work on the "gender digital divide" in driving gender differences in the use of the internet corresponded to the perceptions of the usefulness of the technology in different tasks (OECD, 2018). We capture perceptions of usefulness of ChatGPT by asking students to indicate *"What do you believe are the main advantages of using ChatGPT in coursework?"*. Figure 7a shows the percentage of students that indicated each statement as an advantage of using ChatGPT. While almost no one sees no advantages of

using ChatGPT, there are strong gender differences in perceptions of usefulness. Only 17% of women believe it can improve their grades in a course, whereas 32% of men believe it can, almost double the proportion. There are also strong differences in the perception that it increases accuracy or work quality, with 29% of women and 42% of men holding this belief. Additionally, slightly fewer than half of female students (48%) believe that ChatGPT improves the learning of course methods, whereas the majority of men (63%) hold this belief. The differences mentioned are significant at the 1% significance level. However, in terms of saving time, there are no strong gender differences in perceptions, with most men (86%) and women (80%) believing it is a main advantage of ChatGPT. Altogether, these results show that men perceive ChatGPT as more useful than women, consistent with previous findings in other technology-related settings.

There could also be potential differences in trust in the accuracy of the information provided by ChatGPT, affecting the perceived benefits of using the technology. To capture this, we presented participants with a screen capture of a real prompt and answer submitted to and by ChatGPT, respectively, where the participant was asked whether they trust that the information provided by ChatGPT was accurate, using a 4-point scale from "Completely trust" to "Completely distrust".[19] Figure 7a shows that there are no differences in trust in the accuracy of information provided by ChatGPT, where a majority of men and women (63%) indicated either "Somewhat trust" or "Completely trust".

Finally, confidence in their skills in using the technology might affect women's willingness to engage with ChatGPT, as it has been shown in previous research using male-dominated settings (Coffman et al., 2023). To measure confidence, we take advantage of the prompting task the students performed, and asked them, *"How confident do you feel that the query you just provided will make ChatGPT get the information you need?"*, with choices within a 4-point scale ranging from "Not confident at all" to "Extremely confident". We observe important differences in confidence by gender. 60% of women and 80% of men indicate some level of confidence in their prompt. Moreover, as represented in Figure A2a, around 40% of

[19]The query asked to ChatGPT in the example provided was the following: "What is the poverty rate in Denmark". The participants were later asked, *"Based on this response from ChatGPT, how much do you trust that the poverty rate reported is accurate?"* (see Appendix C).

men indicate feeling very or extremely confident in their own prompt being correct, relative to only 17% of women.

Overall, men have more positive perceptions towards ChatGPT than women, and these seem to play a key role in explaining the gender gap in ChatGPT use and prompting skills. The gap in both outcomes vanishes once we use the measures of perceptions as controls (see Table 1, Column 4). In particular, the perceptions of usefulness and confidence seem to be particularly important in explaining both gaps. This is represented in Table A1, where we observe how the gap changes after controlling for the different sets of perceptions. Column 1 in Panel A shows that controlling for usefulness reduces by half the gender gap in baseline use. Column 5 in Panel B shows that the variable that has the most explanatory power for the gender gap in success rate is how confident they feel that their prompt will provide the correct answer. Adding the level of confidence by itself reduces the gender gap in success rates to 4.1 pp, which is no longer statistically significant. Figure A2a shows that there are indeed large gender differences in the levels of confidence that the prompt provided will give the correct answer. Panel (b) further shows that success rates are positively correlated with confidence levels, and there are no gender differences in success rates within a stated level of confidence.

Confidence explaining away the gender gap in success rates can have two different interpretations. One is that men are better at prompting and they know it. The other is that men are more overconfident about their prompting skills, suggesting that their high level of confidence is at least partially unfounded.[20] In the latter interpretation, people with overconfident beliefs would exert more effort since exaggerated beliefs have a motivational value (Chen and Schildberg-Hörisch, 2019). We assess levels of under- and overconfidence in our sample by constructing two indicator variables. Underconfidence is defined as having a success rate in the prompt of at least 50%, but stating being only slightly confident or not confident at all in the prompt. Overconfidence is the opposite, that is, having a success rate below 50% and stating to be very or extremely confident in the prompt. About 16% of both men

[20]Male overconfidence has long been documented in academically-related domains such as relative performance in adding tasks (Niederle and Vesterlund, 2007, 2011) and cognitive tests (Buser et al., 2018; Möbius et al., 2022), as well as in domains related to the job search of recent graduates (Cortés et al., 2022). For other references see Croson and Gneezy (2009).

and women are underconfident, while 9.5% of women but 19.5% of men are overconfident. Even though the gender difference in overconfidence is large, almost 65% of the men have an accurate idea of their skill, which suggests that overconfidence is not the full story behind men having more successful prompts. Actual proficiency and being aware of it is a main part of the story.

3.3.3 Gender Differences in Experience or Exposure

Finally, a gender gap in the use and skills of ChatGPT might be driven by male and female students having different levels of experience or exposure to the technology, through peers or previous experience. To measure exposure through peers, we asked participants to *"indicate the percentage of people you believe use ChatGPT"* for three different groups: their group of friends, students in their course, and professors at NHH.[21] Figure 7c shows the average percentage indicated by the students for each of the groups. Note that there is a significant difference in the percentage of friends that they believe use ChatGPT, with women stating that it is around 63%, and men indicating that it is 70% of their friends. However, their beliefs about other students in the course and professors at NHH are not different, where both believe around 72% of other students and only around 40% of professors use ChatGPT.[22] In a related proxy measure of experience, we asked students whether they have *"ever received inaccurate or misleading information from ChatGPT?"*, with possible answers being "No, never", "Yes, few times" and "Yes, many times", as well as an option for those who have not used it. In Figure 7c, the percentage of students who have experienced inaccurate information is 16 pp higher for men than for women, the latter being only 27%. Altogether, this evidence shows that not only do men have higher exposure to ChatGPT from their surroundings, but they also have more previous experience.

However, when relating the differences in exposure to the differences in use and skills regarding ChatGPT, Table 1 shows that controlling for our exposure measures does not explain

[21] To avoid concerns of men and women having different anchors when estimating this percentage, we provided the following statement before the question: "A survey conducted among university students in the US in the Spring of 2023 reports that 30% of students use ChatGPT for their schoolwork.".

[22] The differences in the percentage of friends may be driven by women having more female friends than men, and not necessarily from inaccurate assessments of the fractions of friends using ChatGPT by either gender.

a significant part of the gap generated in either of our main outcome variables (Column 5), where the gender gaps are still significant at the 5% level. This evidence suggests a limited role of exposure in explaining the gender gap.

3.4 Hypothetical Policy Experiment: Forbidding ChatGPT Would Widen the Gender Gap in Use

Given the current policy discussion around the world, we included in the survey a policy experiment to assess student responses to policies allowing or banning the use of ChatGPT. We rely on a hypothetical vignette experiment in which we randomize, at the student level, whether the professor in the hypothetical course they are taking allows or forbids ChatGPT use during the course, as described in Section 2.3.[23]

Figure 3 plots the raw gender gaps in intended use when ChatGPT is allowed or forbidden. Intended use equals one if students state that they are likely or very likely to use ChatGPT during the course described in their randomly assigned scenario. When ChatGPT is allowed, over 80% of both men and women intend to use it. However, forbidding ChatGPT opens a large and statistically significant gap in intended use. While men respond to the ban with a decrease of 17.6 pp, from 87% intending to use when allowed to 70% when forbidden, the response of women is much larger at 37.9 pp, from 81% when allowed to 43% when forbidden.

The point estimate for the gender gap in intended use following specification 2 is in Table 1, Panel C, Column 1. When ChatGPT is allowed, the gap is 6.4 pp and not statistically significant. A gender gap in intended use equal to 20.3 pp opens up as a result of the forbidding policy (see interaction coefficient).

We note that intended use is higher for both men and women under the hypothetical scenario when the professor explicitly mentions that ChatGPT is allowed in the course than the baseline use that we documented in Section 3.1. Our take on this difference is that, to the best of our knowledge, the professors in the courses we recruited participants from did

[23]Randomizing this type of policy in real institutions would not be attainable due to the importance that the issue of ChatGPT has for educators that will make the policy difficult to randomize, and the required sample sizes using randomization at the institution level.

not make any explicit statements on whether or not ChatGPT should be used in the course.[24] When not explicitly stated, the default behavior is up to students' interpretation, and some of them may interpret no rule as not encouraged.

As with our previous results, we add different sets of control variables to identify whether the hypothesized factors influencing use and skills can also be behind the differential policy responses. Unlike the gender gaps in high baseline use and prompt success rates, the responses to policies forbidding ChatGPT are not explained by any of the hypothesized factors influencing adoption. Columns 2-6 in Table 1, Panel C show that the coefficients remain similar in magnitude and statistically significant when adding different sets of controls independently or all controls at once.

Given the wide set of controls that we collected, our interpretation of the prevalence of the gender gap after adding the controls is that inclinations towards rule-following, obedience to authority, and trust in the professor's recommendations, play crucial roles in shaping the divergence in intended use. For instance, if female students are more predisposed to follow established rules and trust the guidance of authority figures, they may be more cautious or reserved in adopting new technologies, even if those technologies are intended to enhance learning experiences.

The crucial implication of these findings is the potential unintended consequences of banning ChatGPT in the classroom. Such a prohibition, intended to maintain a level playing field or address concerns by educators, might inadvertently contribute to a gender gap in AI adoption. By restricting access to this technology, female students could be placed at a disadvantage compared to their male peers, hindering their exposure to and familiarity with AI tools, as well as their prospects of success in a rapidly evolving labor market.

4 Discussion and Conclusion

We conducted a student survey at the Norwegian School of Economics to understand the current use of and proficiency in AI tools such as ChatGPT. We find large gender disparities

[24]The professors in the master's course encouraged the use of ChatGPT but the sample coming from that course is very small.

in both dimensions, with male students being more likely to have already adopted and being more proficient at ChatGPT one year after its initial release. Importantly, policies banning ChatGPT in educational institutions would further widen the gender gap in use.

The implications arising from these findings could have profound significance for the career trajectories of female students. The observed gender disparity in ChatGPT usage raises concerns about potential barriers for women in accessing opportunities in a rapidly evolving job market that increasingly values AI proficiency. One potential constraint is that women who do not become proficient in AI tools refrain from applying to jobs that ask specifically for AI skills since their job decisions have been found to depend on features of the job or the workplace where women differ from men, i.e., competitiveness (Flory et al., 2015; Samek, 2019). Another constraint is that, even if they apply, women who do not acquire AI skills may find themselves at a disadvantage in the selection process for a growing number of positions that demand competence in this technology. In addition, once on the job, AI will likely drive differences in productivity and efficiency, leaving those that do not know how to use it properly behind. This could mean that women will miss out on promotions and career advancement if they lack AI skills. This discrepancy not only affects individual career prospects but also contributes to perpetuating gender imbalances in industries where AI proficiency is becoming a prerequisite, hindering diversity and inclusion efforts.

Our results also have wider implications regarding whether AI will reduce or exaggerate existing inequalities between high- and low-skill workers. The main idea is that labor demand is prone to decrease in tasks closely substitutable with the new technology, while it is inclined to rise in tasks that complement it (Brynjolfsson and Mitchell, 2017).[25] The results from early work suggest that AI can reduce inequalities between workers. An experiment with customer support agents shows that the low-skill agents using an AI tool that provides conversational guidance are able to increase their number of issues resolved per hour to the level of the high-skill agents (Brynjolfsson et al., 2023). Similarly, software developers with less developing experience benefit most from having access to the AI tool GitHub Copilot

[25] A study on the early look at the potential impact of large language models such as GPTs finds that around 80% of the US workforce will see that at least 10% of their tasks will be affected by LLMs. In addition, the early predictions suggest that 15% of the tasks can be completed faster while keeping the quality level (Eloundou et al., 2023).

Peng et al. (2023). Our results suggest that, at least in the case of students developing general competencies at the undergraduate level, who can all be considered high-skill, those with higher admission grades have more to gain from AI tools because they are more successful at writing prompts. This implies that the potential benefit of AI tools hinges on the ability to interact with the AI, and that the top women in our sample, who have the lowest ChatGPT adoption rates, are those who may have more to lose from not becoming proficient at AI tools. In the future, the significance of prompting skills may diminish, as recent research has found that the newer version ChatGPT-4 can solve complex tasks in multiple domains with performance close to human level and without the need of special prompting (Bubeck et al., 2023), but for now it is a key skill.

In considering our results' implications for student learning and non-AI skill development, it is crucial to address potential interference between ChatGPT use and other essential skills in education and the labor market, such as critical thinking and problem-solving. We still lack evidence on whether AI adoption affects students' learning or grades, but if more traditional skills are easy to assess during exams or recruitment, students relying heavily on AI tools might find themselves at a disadvantage. Interestingly, the gender gaps we have identified could, in this context, offer advantages to women over men. As AI tools become more integral to work and daily life, influencing, for example, career choices (Reeder and Lee, 2022), the balance between traditional and AI skills in education and the labor market remains uncertain. Nevertheless, given the most likely scenario in which AI becomes increasingly important, it is in the hands of institutions to foster the development of both skill sets in a mutually beneficial manner. This is particularly crucial for female students, who, tending to adhere to rules, should be empowered with the confidence that they can adeptly develop and apply both types of skills, ensuring success in their chosen educational paths and careers.

References

Almås, I., O. Attanasio, and P. Jervis (2023). Economics and measurement: New measures to model decision making. Technical report, National Bureau of Economic Research.

Ameriks, J., J. Briggs, A. Caplin, M. D. Shapiro, and C. Tonetti (2020). Long-term-care utility and late-in-life saving. *Journal of Political Economy 128*(6), 2375–2451.

Andre, P., C. Pizzinelli, C. Roth, and J. Wohlfart (2022). Subjective models of the macroeconomy: Evidence from experts and representative samples. *The Review of Economic Studies 89*(6), 2958–2991.

Avery, M., A. Leibbrandt, and J. Vecci (2023). Does artificial intelligence help or hurt gender diversity? evidence from two field experiments on recruitment in tech. *Evidence from Two Field Experiments on Recruitment in Tech (February 14, 2023)*.

Avilova, T. and C. Goldin (2018). What can uwe do for economics? In *AEA Papers and Proceedings*, Volume 108, pp. 186–90.

Awad, E., L. Balafoutas, L. Chen, E. Ip, and J. Vecci (2023). Artificial intelligence and debiasing in hiring: Impact on applicant quality and gender diversity. *Available at SSRN*.

Bernheim, B. D., D. Björkegren, J. Naecker, and M. Pollmann (2022). Causal inference from hypothetical evaluations. Technical report, National Bureau of Economic Research.

Bimber, B. (2000). Measuring the gender gap on the internet. *Social science quarterly*, 868–876.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016). Stereotypes. *The Quarterly Journal of Economics 131*(4), 1753–1794.

Brañas-Garza, P., L. Estepa-Mohedano, D. Jorrat, V. Orozco, and E. Rascón-Ramírez (2021). To pay or not to pay: Measuring risk preferences in lab and field. *Judgment and Decision Making 16*(5), 1290–1313.

Brañas-Garza, P., D. Jorrat, A. M. Espín, and A. Sánchez (2023). Paid and hypothetical time preferences are the same: Lab, field and online evidence. *Experimental Economics 26*(2), 412–434.

Brynjolfsson, E., D. Li, and L. R. Raymond (2023). Generative AI at work. Technical report, National Bureau of Economic Research.

Brynjolfsson, E. and T. Mitchell (2017). What can machine learning do? workforce implications. *Science 358*(6370), 1530–1534.

Brynjolfsson, E. and G. Unger (2023). The Macroeconomics of Artificial Intelligence. https://www.imf.org/en/Publications/fandd/issues/2023/12/Macroeconomics-of-artificial-intelligence-Brynjolfsson-Unger. [Online; accessed 05-December-2023].

Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Buser, T., L. Gerhards, and J. Van Der Weele (2018). Responsiveness to feedback as a personal trait. *Journal of Risk and Uncertainty 56*, 165–192.

Buser, T., M. Niederle, and H. Oosterbeek (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics 129*(3), 1409–1447.

Buser, T., N. Peter, and S. C. Wolter (2017). Gender, competitiveness, and study choices in high school: Evidence from switzerland. *American Economic Review 107*(5), 125–130.

Chen, S. and H. Schildberg-Hörisch (2019). Looking at the bright side: The motivational value of confidence. *European Economic Review 120*, 103302.

Cimpian, J. R., T. H. Kim, and Z. T. McDermott (2020). Understanding persistent gender gaps in STEM. *Science 368*(6497), 1317–1319.

CNBC (2023). Nine in 10 companies want employees with chatgpt skills. https://www.cnbc.com/2023/09/01/more-companies-see-chatgpt-training-as-a-hot-job-perk-for-office-workers.html. de Visé, Daniel and Klar, Rebecca. [Online; accessed 12-December-2023].

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics 129*(4), 1625–1660.

Coffman, K. B., M. R. Collis, and L. Kulkarni (2023). Whether to apply. *Management Science*.

Cortés, P., J. Pan, E. Reuben, L. Pilossoph, and B. Zafar (2022). Gender differences in job search and the earnings gap: Evidence from the field and lab. Technical report, National Bureau of Economic Research.

Croson, R. and U. Gneezy (2009). Gender differences in preferences. *Journal of Economic Literature 47*(2), 448–474.

Direktoratet for høyere utdanning og kompetanse (2023). Hovedopptak til høyere utdanning ved universiteter og høgskoler. https://hkdir.no/rapporter-undersokelser-og-statistikk/hovedopptak-til-hoyere-utdanning-ved-universiteter-og-hogskoler-juli-2023. [Online; accessed 12-December-2023].

Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Enke, B., R. Rodriguez-Padilla, and F. Zimmermann (2022). Moral universalism: Measurement and economic relevance. *Management Science 68*(5), 3590–3603.

Falk, A., A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics 133*(4), 1645–1692.

Falk, A., A. Becker, T. Dohmen, D. Huffman, and U. Sunde (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science 69*(4), 1935–1950.

Flory, J. A., A. Leibbrandt, and J. A. List (2015). Do competitive workplaces deter female workers? a large-scale natural field experiment on job entry decisions. *The Review of Economic Studies 82*(1), 122–155.

Goldin, C. (2015). Gender and the undergraduate economics major: Notes on the undergraduate economics major at a highly selective liberal arts college. *manuscript, April 12*.

Hainmueller, J., D. Hangartner, and T. Yamamoto (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences 112*(8), 2395–2400.

Hirshman, S. D. and A. Willén (2022). Does increasing risk widen gender gaps? *NHH Dept. of Economics Discussion Paper* (20).

Kessler, J. B., C. Low, and C. D. Sullivan (2019). Incentivized resume rating: Eliciting employer preferences without deception. *American Economic Review 109*(11), 3713–3744.

Kugler, A. D., C. H. Tinsley, and O. Ukhaneva (2021). Choice of majors: Are women really different from men? *Economics of Education Review 81*, 102079.

Landaud, F., É. Maurin, B. Willage, and A. Willén ((accepted) 2023). The value of a high school gpa. *Review of Economics and Statistics*.

Lo, C. K. (2023). What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences 13*(4), 410.

Möbius, M. M., M. Niederle, P. Niehaus, and T. S. Rosenblat (2022). Managing self-confidence: Theory and experimental evidence. *Management Science 68*(11), 7793–7817.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics 122*(3), 1067–1101.

Niederle, M. and L. Vesterlund (2011). Gender and competition. *Annu. Rev. Econ. 3*(1), 601–630.

Noy, S. and W. Zhang (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*.

OECD (2018). Bridging the digital gender divide: Include, upskill, innovate. *OECD*.

Ost, B. (2010). The role of peers and grades in determining major persistence in the sciences. *Economics of Education Review 29*(6), 923–934.

Peng, S., E. Kalliamvakou, P. Cihon, and M. Demirer (2023). The impact of ai on developer productivity: Evidence from github copilot. *arXiv preprint arXiv:2302.06590*.

Pisanelli, E. (2022). A new turning point for women: artificial intelligence as a tool for reducing gender discrimination in hiring. *Available at SSRN 4254965*.

Rask, K. and J. Tiefenthaler (2008). The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. *Economics of Education Review 27*(6), 676–687.

Reeder, K. and H. Lee (2022). Impact of artificial intelligence on us medical students' choice of radiology. *Clinical imaging 81*, 67–71.

Samek, A. (2019). Gender differences in job entry decisions: A university-wide field experiment. *Management Science 65*(7), 3272–3281.

Stango, V. and J. Zinman (2023). We are all behavioural, more, or less: A taxonomy of consumer decision-making. *The Review of Economic Studies 90*(3), 1470–1498.

Titchener, E. B. (1901). *Experimental psychology: A manual of laboratory practice, vol. I: Qualitative experiments*. New York: The Macmillan Company.

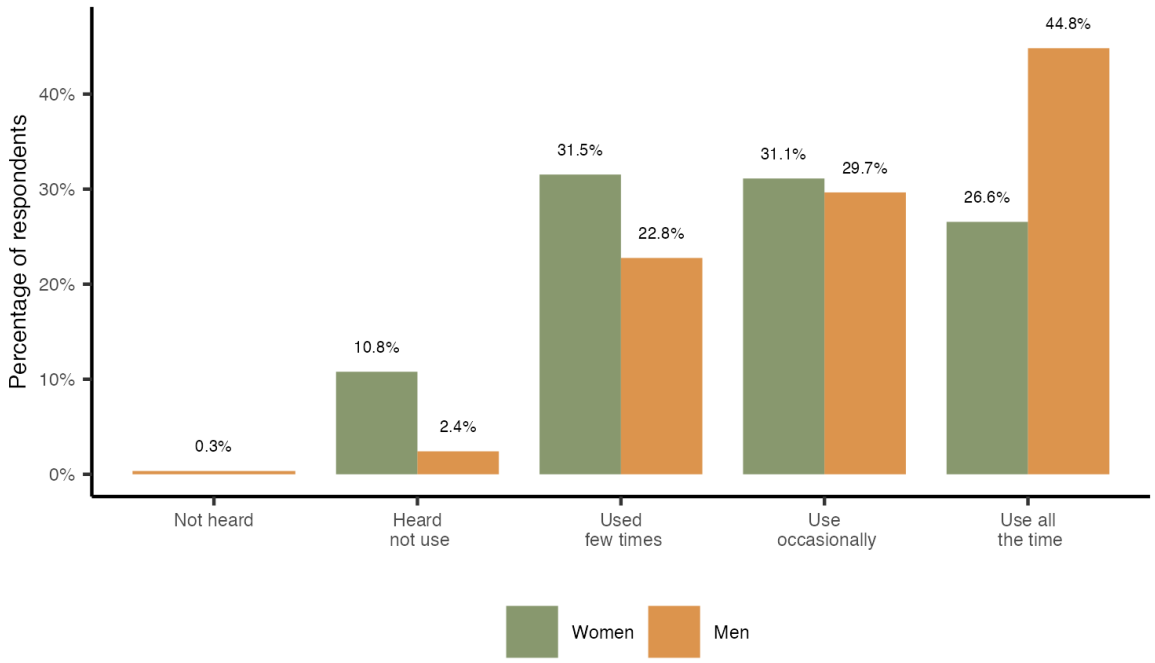
Ugalde, M. P. (2022). Gender, grade sensitivity, and major choice.

Williams, J. C. (2014). Double jeopardy? an empirical study with implications for the debates over implicit bias and intersectionality. *Harvard Journal of Law & Gender 37*, 185.

WSJ (2023). More companies see chatgpt training as a hot job perk for office workers. https://www.wsj.com/tech/ai/talking-to-chatbots-is-now-a-200k-job-so-i-applied-258bd5f0. Stern, Joanna. [Online; accessed 12-December-2023].

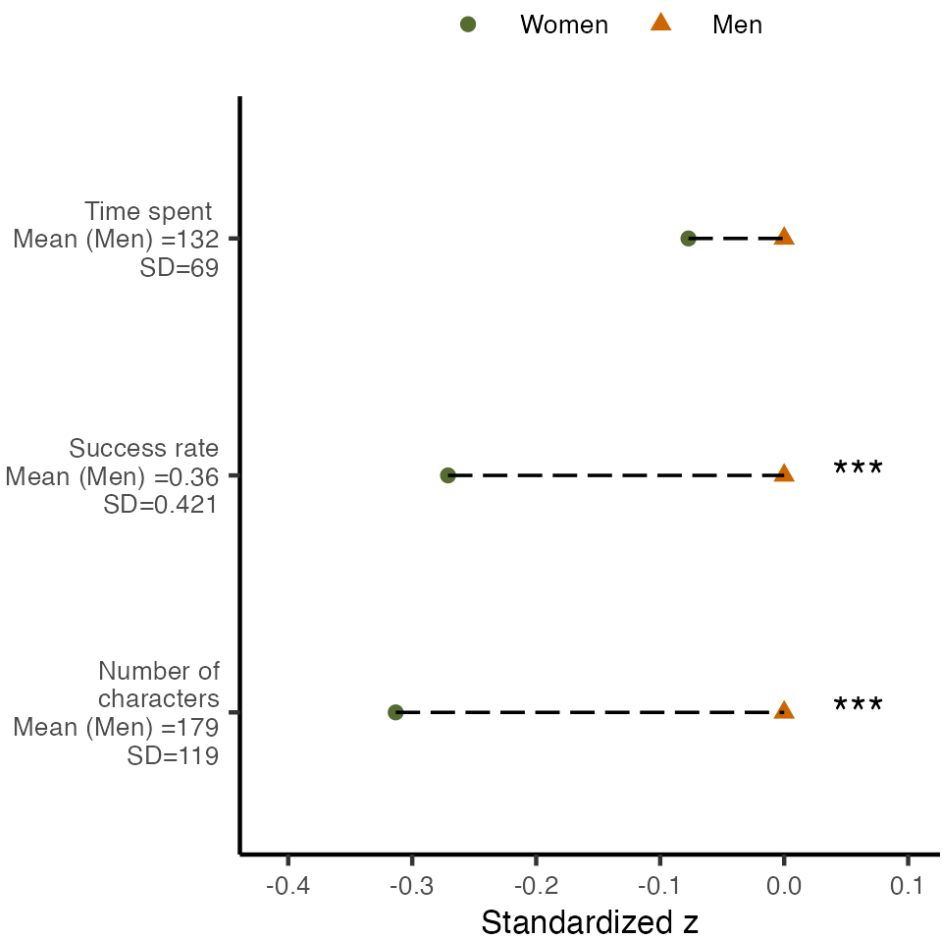
5 Figures

Figure 1: Baseline use



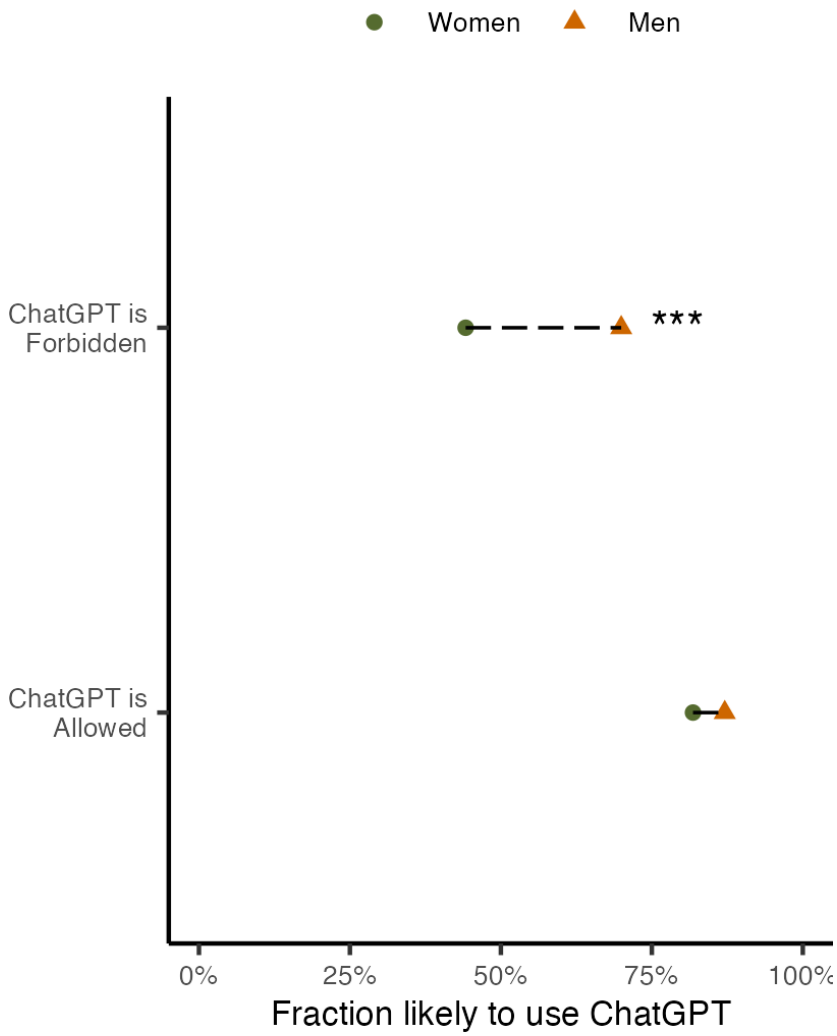
Notes: The figure shows a bar plot with the percentage of women and men indicating each answer to the question "How familiar are you with ChatGPT or similar tools?". Within gender the percentages across categories add up to 100%.

Figure 2: Prompt quality



Notes: The figure plots, by gender, the mean standardized values of three variables: time spent in the prompting task in seconds, success rate, and the number of characters of each prompt. All variables were standardized using the mean of men for each variable.

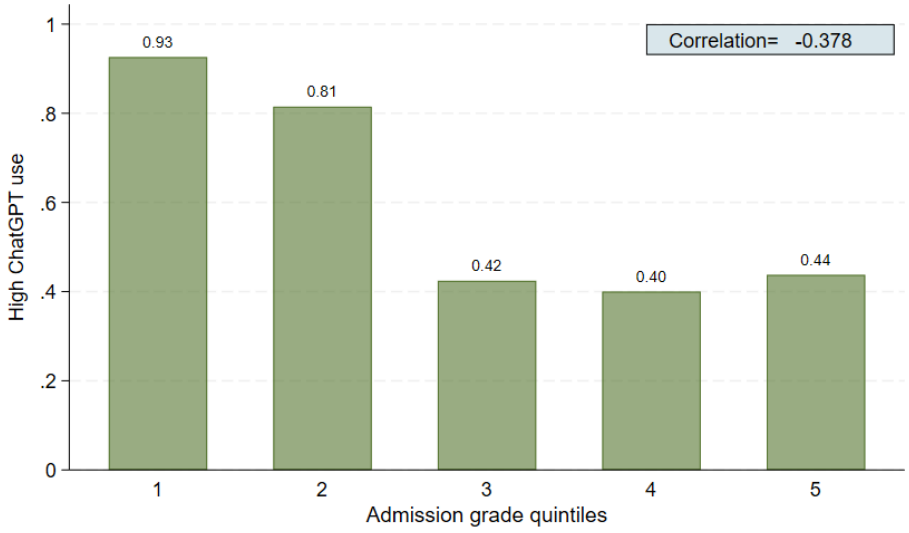
Figure 3: Policy responses



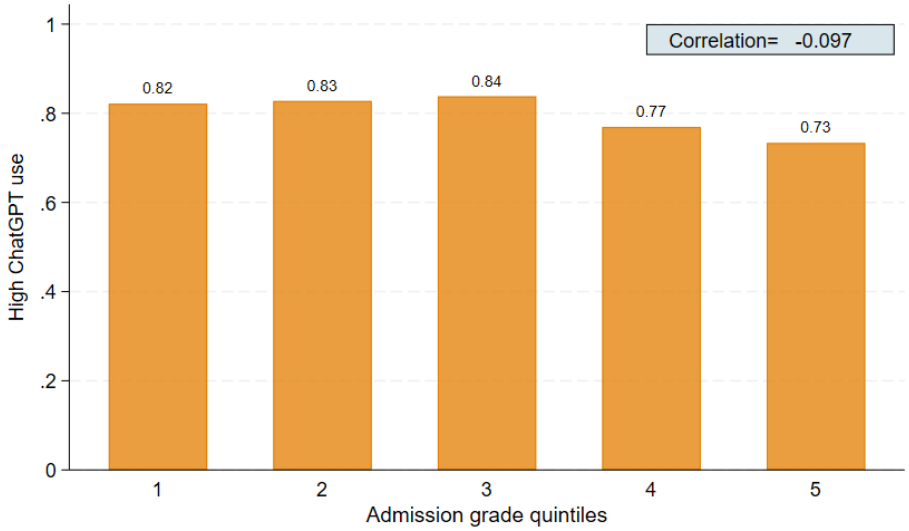
Notes: The figure shows, by gender, the fraction of participants that indicated "Somewhat likely" or "Very likely" to the question of how likely would they use ChatGPT in the hypothetical course presented in the vignette experiment. We show the estimates for the two randomly assigned scenarios: professor "forbids" and "allows" treatment.

Figure 4: Gender differences in baseline use by admission grade quintiles

(a) Women



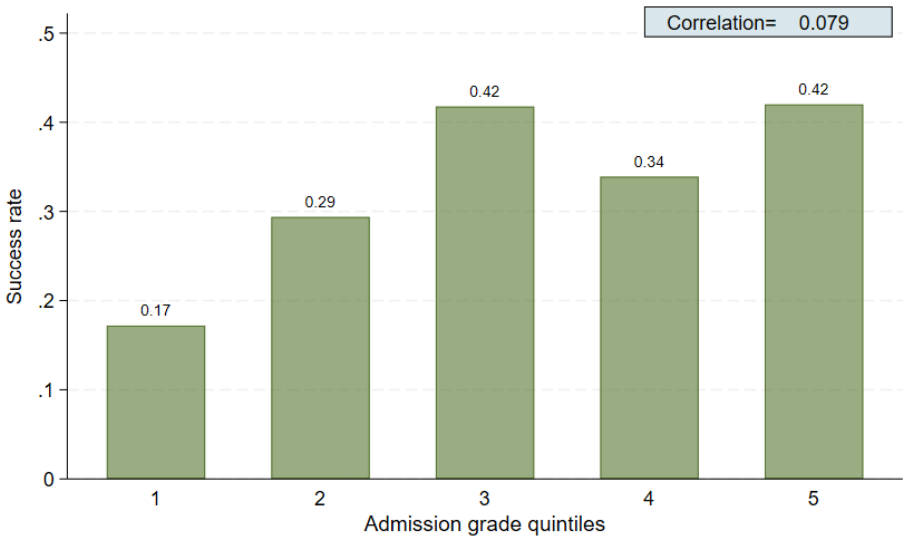
(b) Men



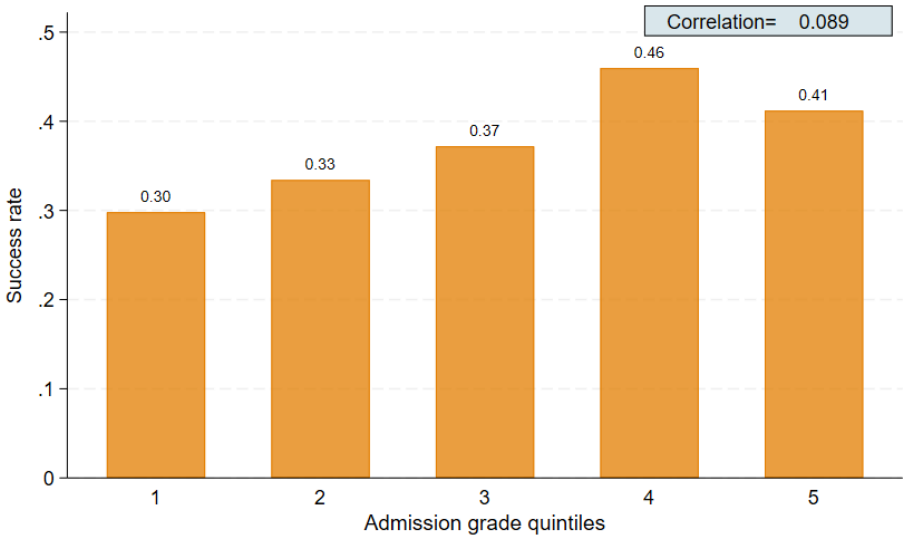
Notes: Panels (a) and (b) show the proportion of women and men, respectively, with high baseline use of ChatGPT across the self-reported admission grade quintiles (273 respondents).

Figure 5: Gender differences in prompt succcess by admission grade quintiles

(a) Women



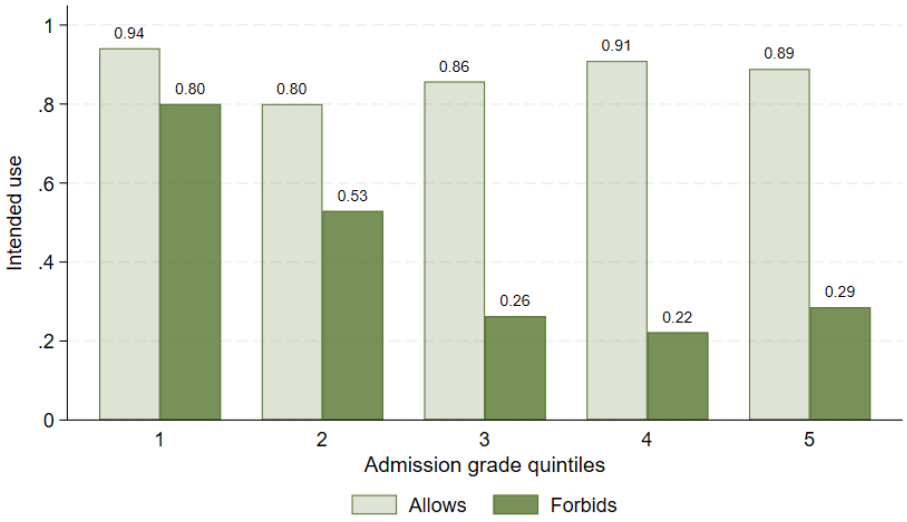
(b) Men



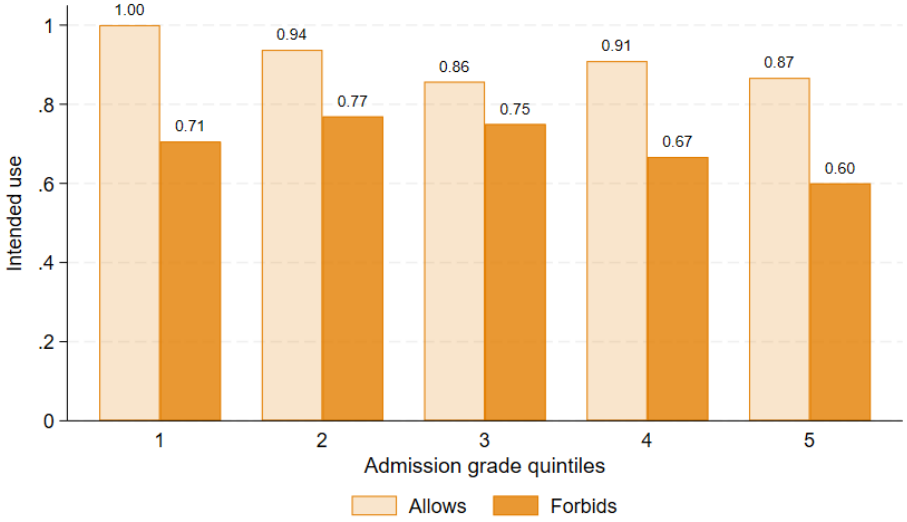
Notes: Panels (a) and (b) show the average success rate in the prompting task for women and men, respectively, across the self-reported admission grade quintiles (273 respondents).

Figure 6: Gender differences in policy response by admission grade quintiles

(a) Women

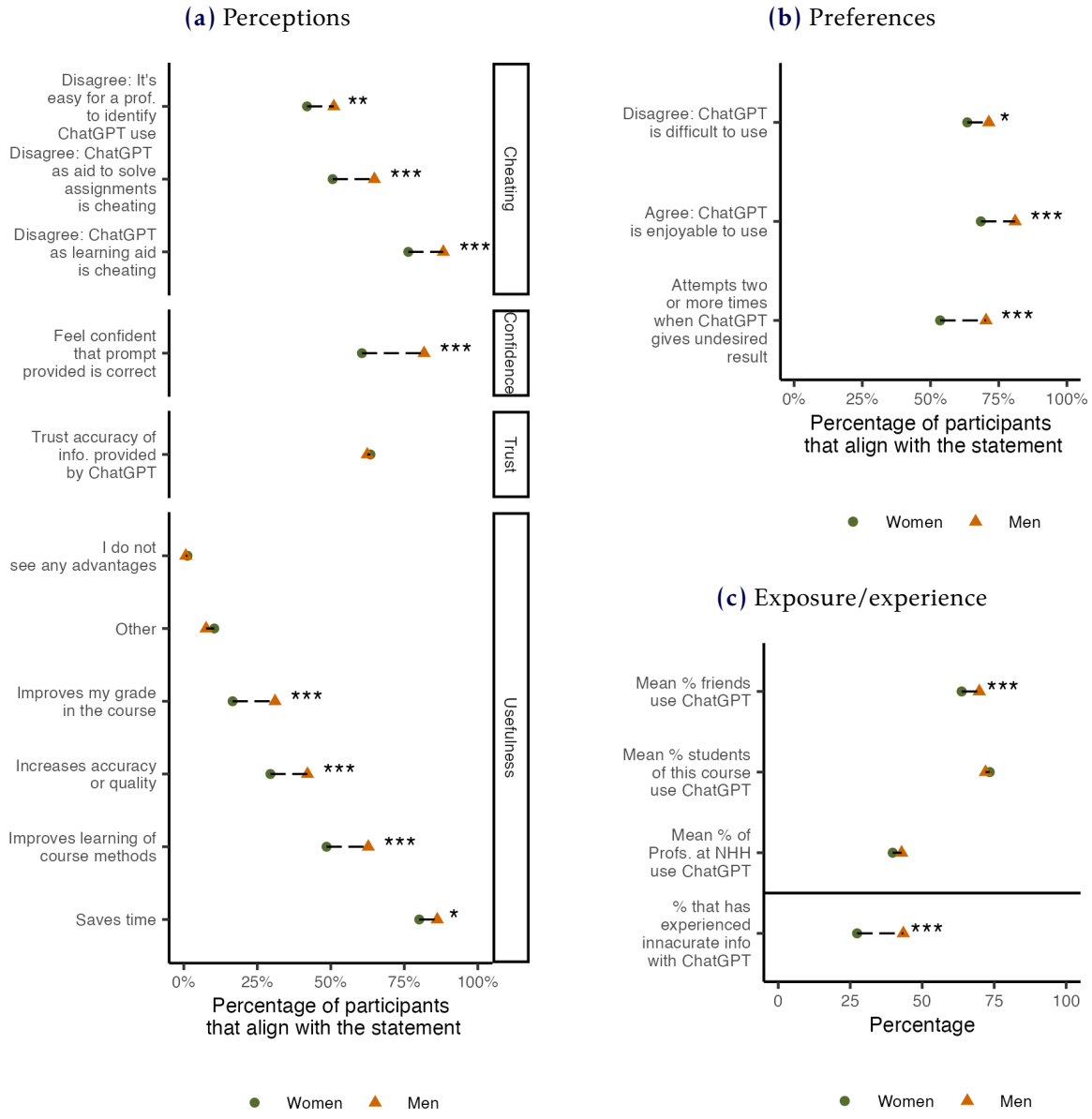


(b) Men



Notes: Panels (a) and (b) show the proportion of individuals who indicated likely intended use of ChatGPT in the vignette experiment for women and men, respectively, across the self-reported admission grade quintiles (273 respondents). In brighter colors is the intended use in the professor "allows" treatment, whereas in darker colors is the intended use in the "forbids" treatment.

Figure 7: Potential factors influencing use and skill: gender differences in attitudes



Notes: Panels (a) and (b) show, by gender, the percentage of participants whose answer aligns with each statement on the left of the corresponding graph. Panel (a) shows the results for the statements related to perceptions, while Panel (b) for the statements related to preferences. Panel (c) shows the variables capturing the exposure/experience channel, where the first three rows indicate, by gender, the mean estimate of the percentage of individuals that the participant believes use ChatGPT within the three indicated groups. The last row shows the percentage of participants that indicated to have experienced inaccurate information from ChatGPT. All gender gaps are raw estimates, without any controls. $* p < 0.10$, $** p < 0.05$, $*** p < 0.01$.

6 Tables

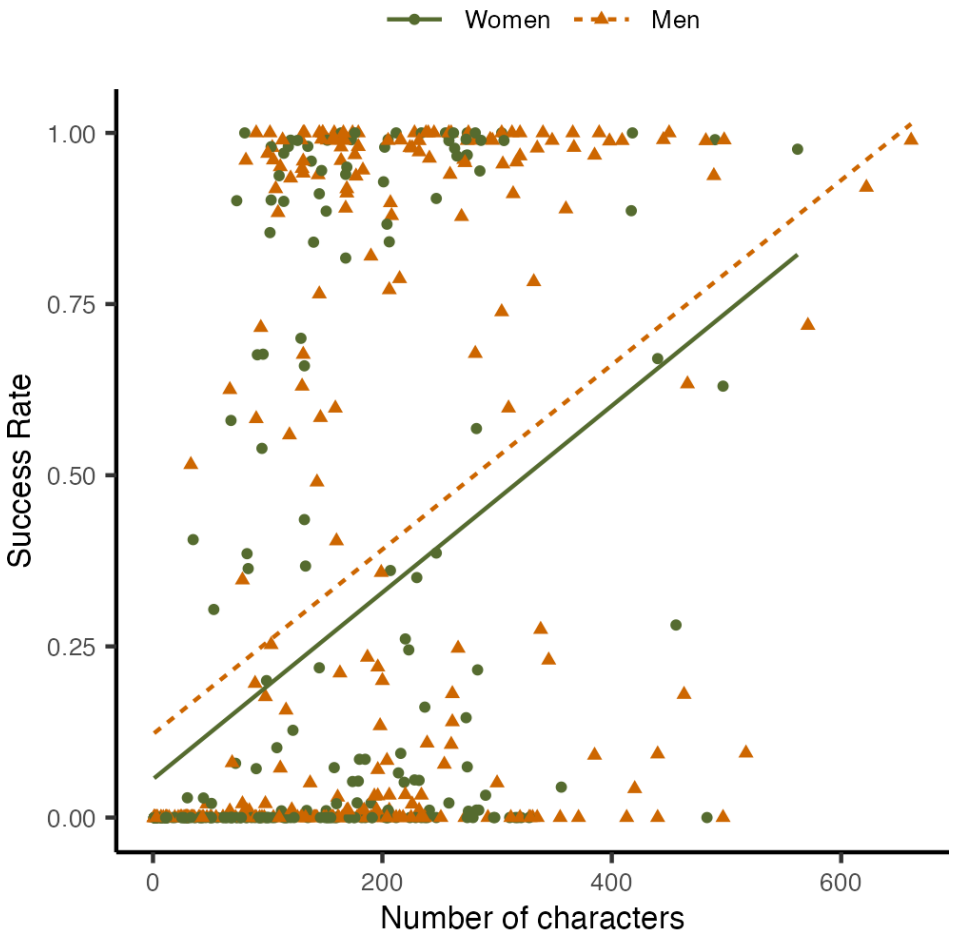
Table 1: Main results

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Use ChatGPT occasionally or all the time (baseline use)						
Male	0.172***	0.098**	0.051	0.023	0.079**	0.012
	(0.042)	(0.043)	(0.037)	(0.043)	(0.037)	(0.040)
Constant	0.569***	1.362***	0.587***	0.127	0.193**	0.537
	(0.033)	(0.227)	(0.145)	(0.210)	(0.079)	(0.352)
Controls	None	Academic, risk & time	Preferences	Perceptions	Exposure/ experience	All
Observations	514	514	514	514	514	514
Panel B: Prompt success rate						
Male	0.111***	0.116***	0.101***	0.021	0.103***	0.026
	(0.037)	(0.043)	(0.038)	(0.038)	(0.039)	(0.044)
Constant	0.249***	-0.663	0.462**	0.061	0.405***	-0.772
	(0.026)	(0.403)	(0.200)	(0.249)	(0.084)	(0.558)
Controls	None	Baseline use, academic, risk & time	Preferences	Perceptions	Exposure/ experience	All
Observations	514	514	514	514	514	514
Panel C: Policy response (likely or very likely to use ChatGPT)						
Male	0.064	-0.047	-0.010	-0.037	0.010	-0.056
	(0.046)	(0.045)	(0.044)	(0.047)	(0.043)	(0.052)
ChatGPT forbidden	-0.379***	-0.384***	-0.396***	-0.397***	-0.355***	-0.391***
	(0.059)	(0.053)	(0.056)	(0.054)	(0.054)	(0.054)
Male × ChatGPT forbidden	0.203***	0.211***	0.221***	0.228***	0.201***	0.218***
	(0.076)	(0.072)	(0.072)	(0.070)	(0.070)	(0.072)
Constant	0.810***	1.160***	0.647***	0.547**	0.521***	1.040**
	(0.037)	(0.332)	(0.205)	(0.257)	(0.086)	(0.469)
Controls	None	Baseline use, academic, risk & time	Preferences	Perceptions	Exposure/ experience	All
Observations	514	514	514	514	514	514

Notes: Panels A and B show point estimates on gender differences in baseline use and success rate of the prompts written by students, respectively. Panel C shows point estimates on intended use from random variation on whether the professor allows or forbids the use of ChatGPT in a hypothetical course presented to the students. Each column title indicates what control variables are included in the regression. Column 1 presents raw estimates and Column 6 includes all controls added one by one in Columns 2-5. Academic controls include year in college, admission grade and an indicator for whether the admission grade is missing. Risk and time preferences are collected using the survey questions from the World Preferences Survey. Preferences include questions on whether students enjoy or find it difficult to use ChatGPT, as well as a measure of persistence in using ChatGPT. Perceptions include views on whether ChatGPT is equivalent to cheating, how useful it is, trust and confidence in own ChatGPT skills. Exposure/experience refers to what fraction of their friends, other students in their class and NHH professors use ChatGPT. $* p < 0.10$, $** p < 0.05$, $*** p < 0.01$.

A Appendix Figures

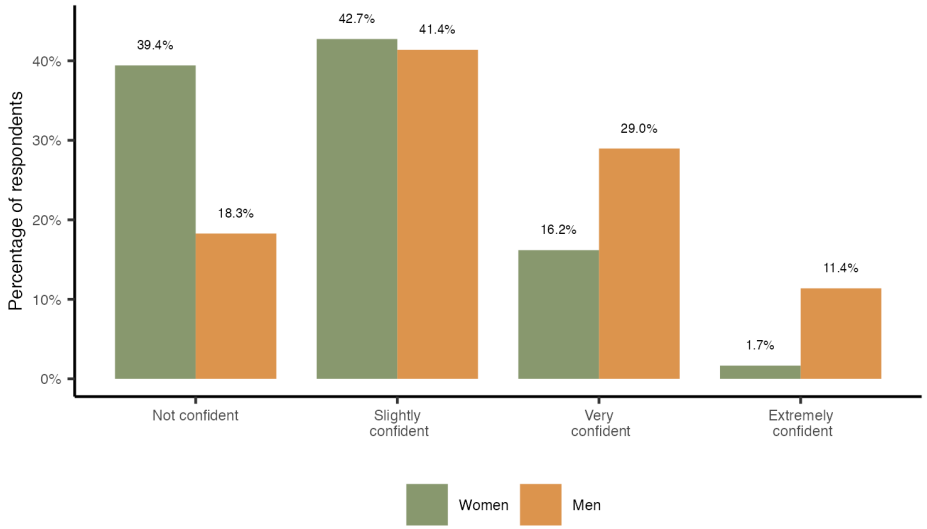
Figure A1: Relationship between success rate and number of characters



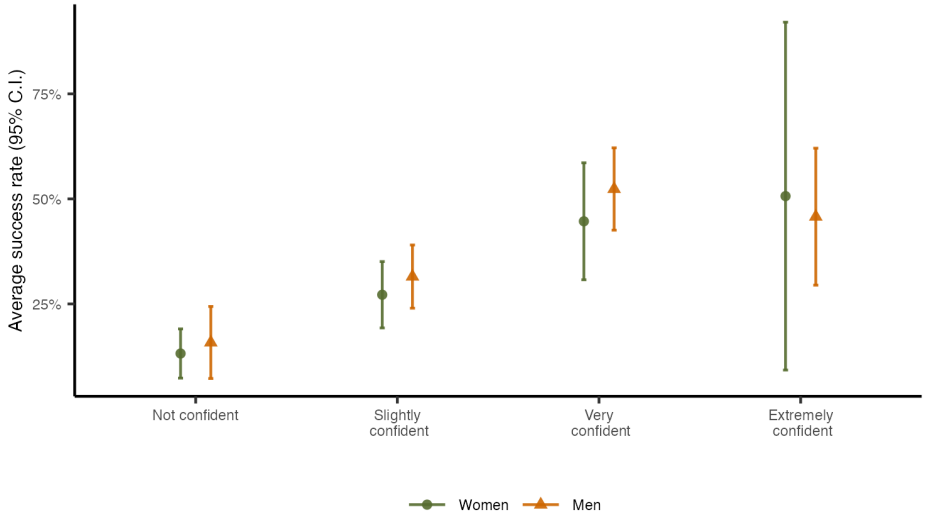
Notes: The scatterplot displays the relationship between the number of characters that students write in their prompt (x-axis) the success rate of the prompt (y-axis), for the full sample. The plot also provides the linear fits for both men (dashed) and women (solid), where the slope is of 0.13 for both.

Figure A2: Confidence in own prompt and success rates by level of confidence

(a) Confidence in own prompt



(b) Success rate by level of confidence



Notes: Panel (a) shows a bar plot with the percentage of women and men indicating each answer to the question "How confident do you feel that the query you just provided will make ChatGPT get the information you need?", which they answered after the prompting skills task. Panel (b) shows the average success rate for each answer option in the confidence question.

B Appendix Tables

Table A1: Role of different perceptions in explaining the main results

	(1)	(2)	(3)	(4)	(5)
Panel A: Use ChatGPT occasionally or all the time (baseline use)					
Male	0.172***	0.100**	0.085**	0.171***	0.106**
	(0.042)	(0.043)	(0.041)	(0.042)	(0.043)
Constant	0.569***	0.527***	0.379***	0.572***	0.440***
	(0.033)	(0.193)	(0.069)	(0.098)	(0.044)
Controls	None	Cheating	Usefulness	Trust	Confidence
Observations	514	514	514	514	514
Panel B: Prompt success rate					
Male	0.111***	0.081**	0.099***	0.106***	0.041
	(0.037)	(0.040)	(0.037)	(0.037)	(0.037)
Constant	0.249***	0.415*	0.217***	0.133*	0.130***
	(0.026)	(0.217)	(0.063)	(0.069)	(0.028)
Controls	None	Cheating	Usefulness	Trust	Confidence
Observations	514	514	514	514	514
Panel C: Policy response (likely or very likely to use ChatGPT)					
Male	0.064	0.007	-0.003	0.063	0.022
	(0.046)	(0.045)	(0.045)	(0.047)	(0.047)
ChatGPT forbidden	-0.379***	-0.409***	-0.373***	-0.378***	-0.391***
	(0.059)	(0.055)	(0.055)	(0.059)	(0.058)
Male × ChatGPT forbidden	0.203***	0.225***	0.211***	0.206***	0.216***
	(0.076)	(0.072)	(0.072)	(0.076)	(0.075)
Constant	0.810***	0.781***	0.647***	0.850***	0.765***
	(0.037)	(0.210)	(0.072)	(0.101)	(0.048)
Controls	None	Cheating	Usefulness	Trust	Confidence
Observations	514	514	514	514	514

Notes: Panels A and B show point estimates on gender differences in baseline use and success rate of the prompts written by students, respectively. Panel C shows point estimates on intended use from random variation on whether the professor allows or forbids the use of ChatGPT in a hypothetical course presented to the students. Each column title indicates what control variables are included in the regression. Column 1 presents raw estimates and Columns 2-5 add a different set of perceptions variables as indicated at the bottom of the respective column. $* p < 0.10$, $** p < 0.05$, $*** p < 0.01$.

C Survey questionnaire

Figure A3: Page 1. Consent

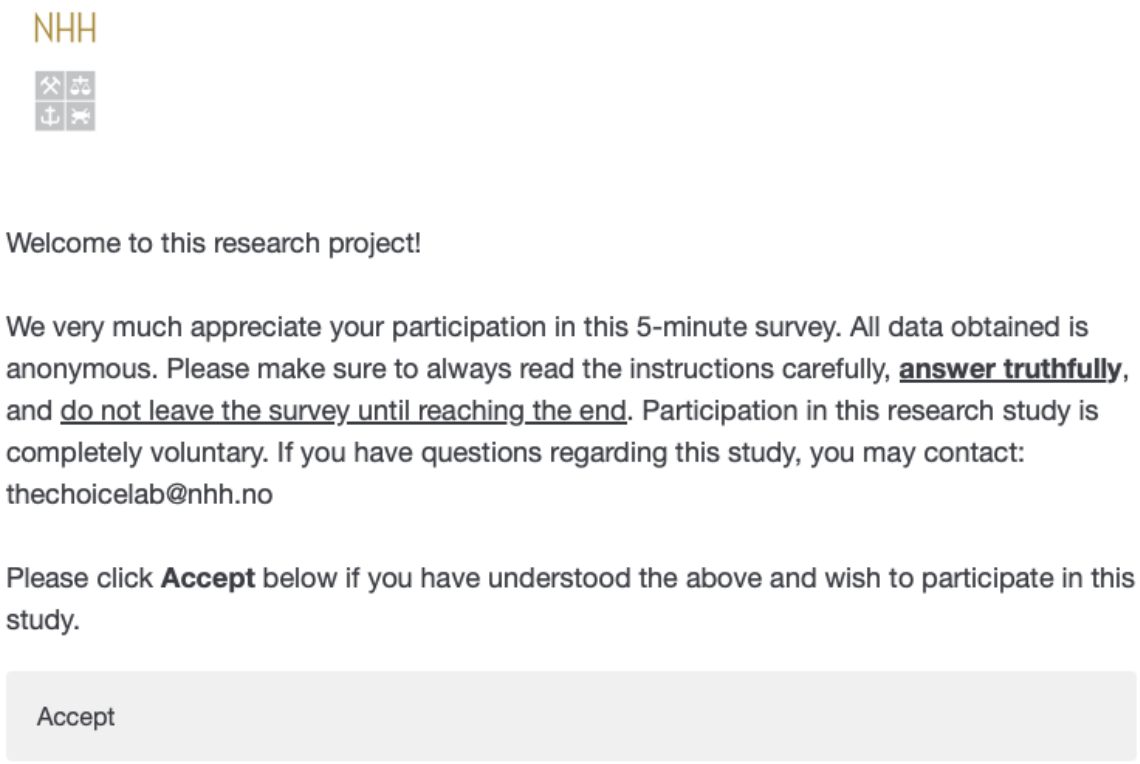


Figure A4: Page 2. Background characteristics

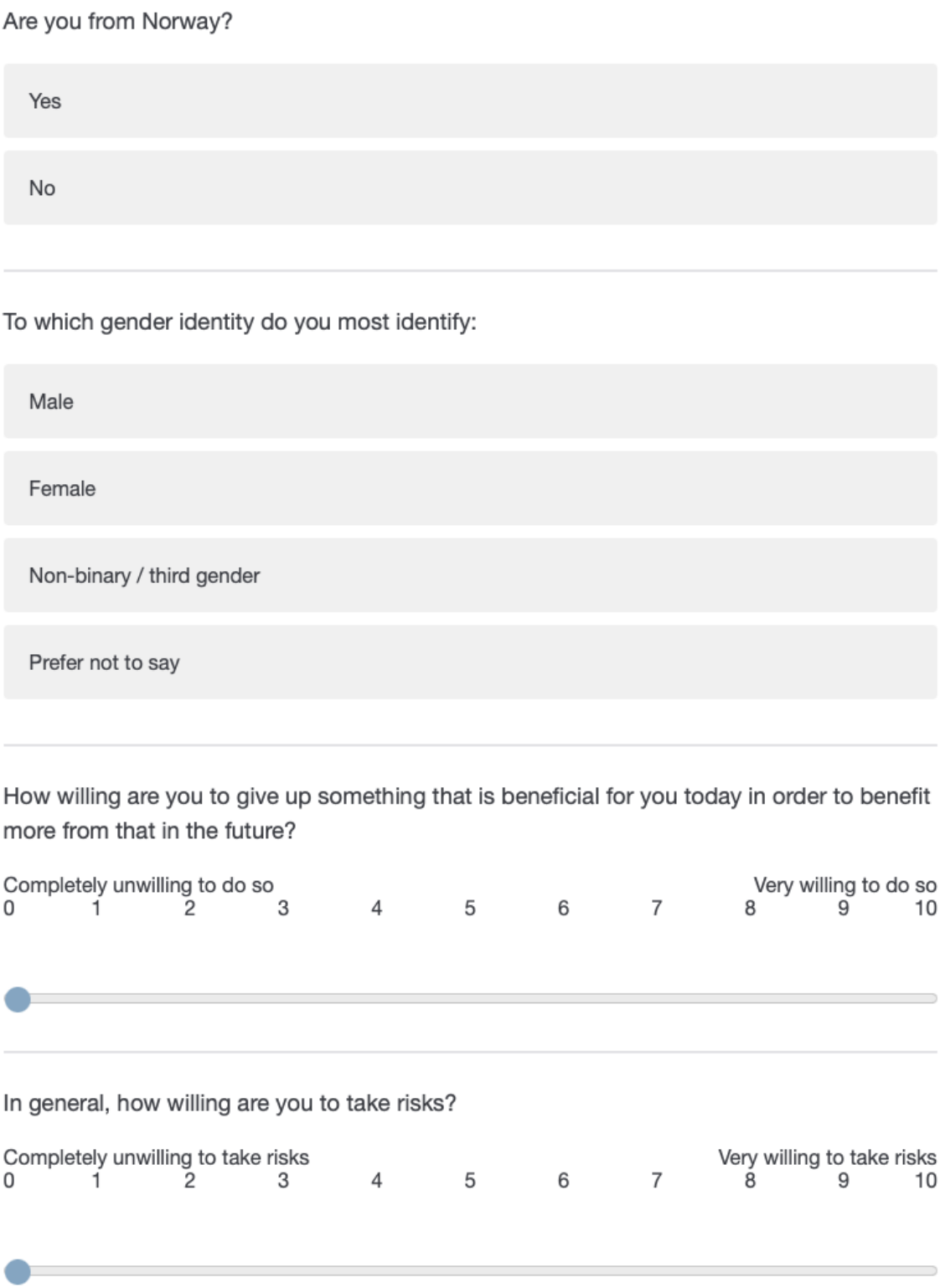


Figure A5: Page 3. "Allows" treatment

Imagine you are enrolled in a course on Environmental Policy and Economic Impact. This course explores the intersection of environmental regulations, economic incentives, and their effects on industry practices and sustainability. The professor explicitly allows the use of ChatGPT during coursework. It is an 8-week course with final evaluation given by a final home exam.

Given this scenario, how likely are you to use ChatGPT throughout the course?

- Very unlikely
- Somewhat unlikely
- Neither likely nor unlikely
- Somewhat likely
- Very likely

Given the scenario, how likely are you to use ChatGPT during the final exam?

- Very unlikely
- Somewhat unlikely
- Neither likely nor unlikely
- Somewhat likely
- Very likely

Figure A6: Page 4. "Forbids" treatment

Imagine you are enrolled in a course on Climate Change Economics. This course delves into the economic analysis of climate change, including the evaluation of mitigation strategies, adaptation costs, and international climate policy agreements. The professor explicitly forbids the use of ChatGPT during coursework. It is an 9-week course with final evaluation given by a final home exam.

Given this scenario, how likely are you to use ChatGPT throughout the course?

- Very unlikely
- Somewhat unlikely
- Neither likely nor unlikely
- Somewhat likely
- Very likely

Given the scenario, how likely are you to use ChatGPT during the final exam?

- Very unlikely
- Somewhat unlikely
- Neither likely nor unlikely
- Somewhat likely
- Very likely

Figure A7: Page 5. Prompting skills task

Do you know how to use ChatGPT?

Please take a moment to carefully check the image presented below.

Using the space provided, please write down the question that **you would ask to ChatGPT** to learn about the official name of this visual phenomenon. Remember ChatGPT cannot observe the image.

Figure A8: Page 6. Confidence question

How confident do you feel that the query you just provided will make ChatGPT get the information you need?

- Not confident at all
- Slightly confident
- Very confident
- Extremely confident

Figure A9: Page 7. ChatGPT use

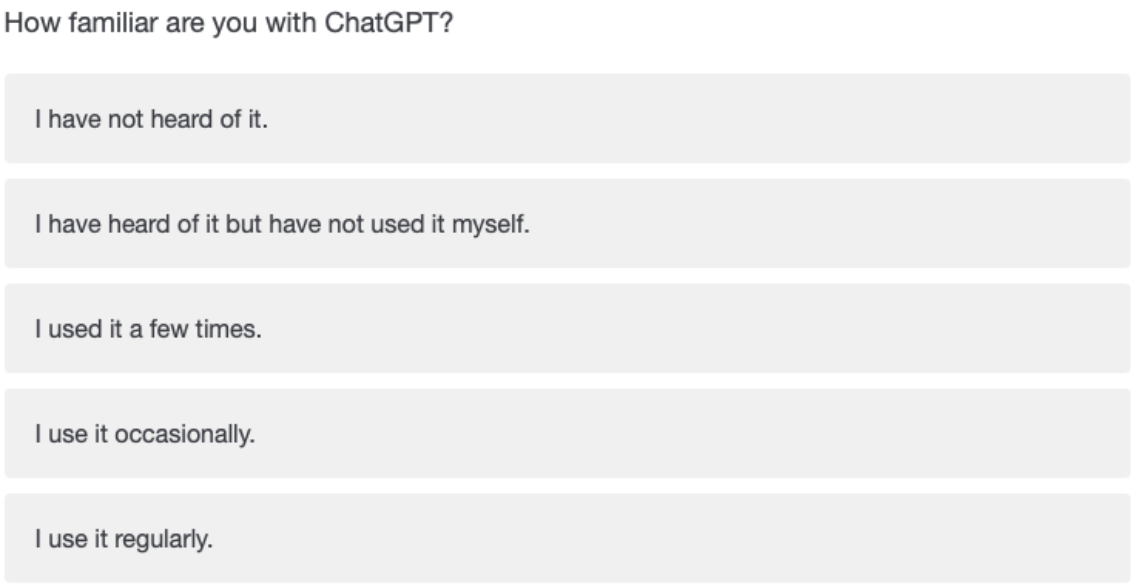


Figure A10: Page 8. Exposure and typical tasks

A survey conducted among university students in the US in the Spring of 2023 reports that 30% of students use ChatGPT for their schoolwork.

Now, for each of the groups below, please indicate the percentage of people you believe use ChatGPT:

0 20 40 60 80 100

Your group of friends

Students in this course

Professors at NHH

What type of tasks do you typically ask ChatGPT to help with? (Please select up to the most common three)

- Coding tasks
- Writing tasks
- Retrieving information
- Solving Math questions
- Other (Please specify)
- I don't use it

Figure A11: Page 9. Frequency by task

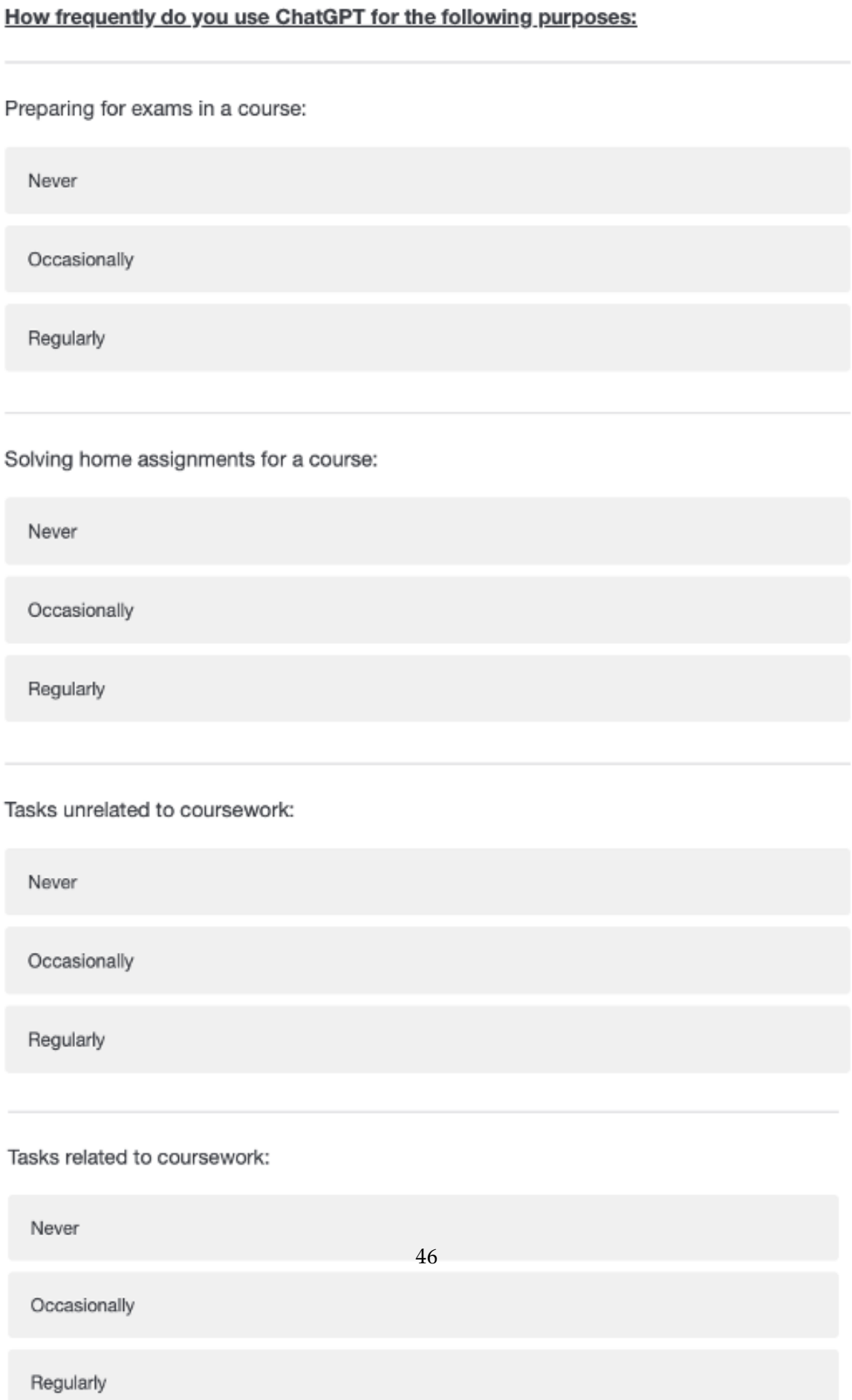


Figure A12: Page 10. Advantages (Usefulness)

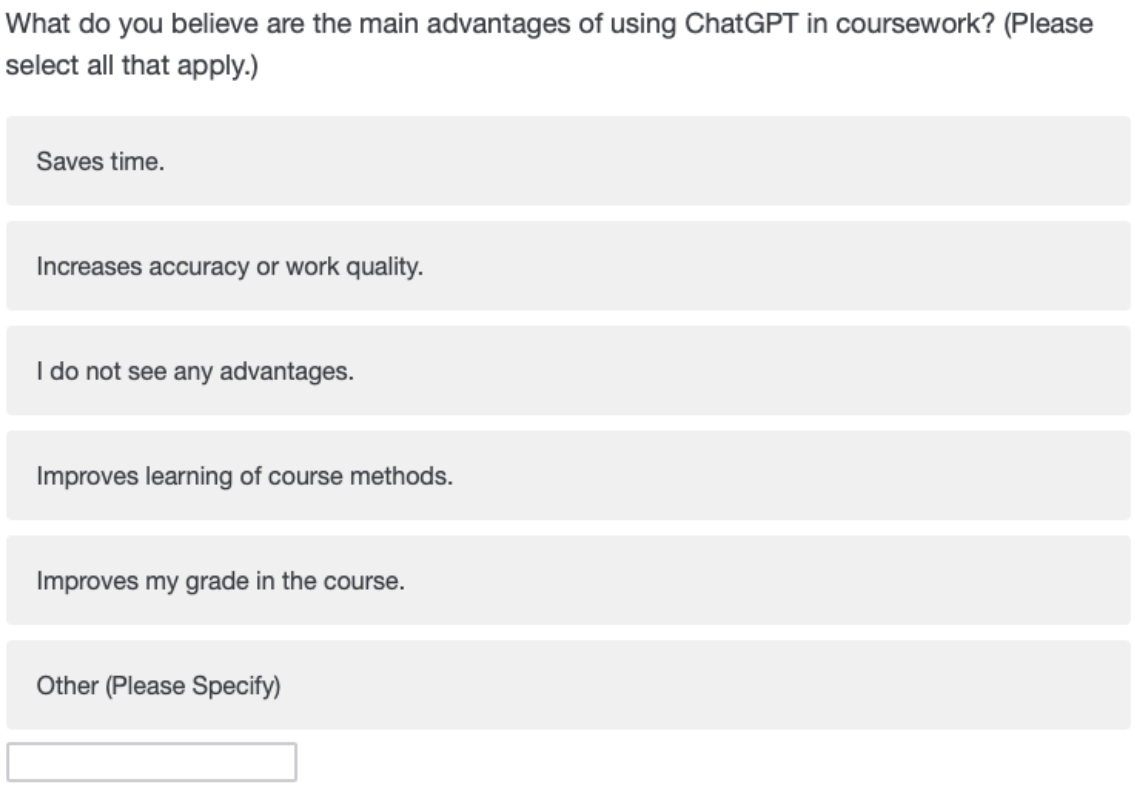


Figure A13: Page 11.1 Agree/Disagree

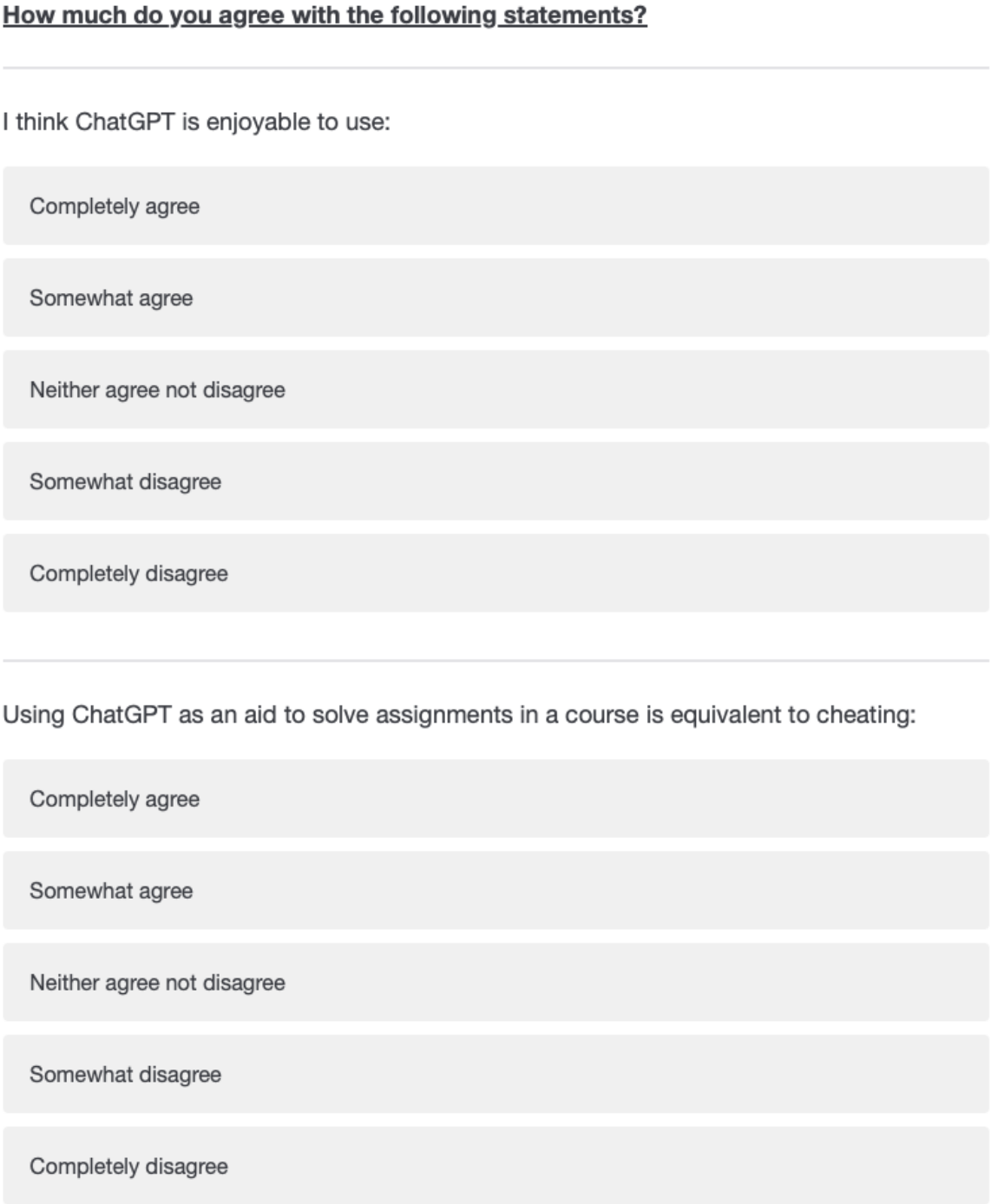


Figure A14: Page 11.2 Agree/Disagree

Using ChatGPT as a learning aid in a course is equivalent to cheating:

- Completely agree
- Somewhat agree
- Neither agree not disagree
- Somewhat disagree
- Completely disagree

I think ChatGPT is difficult to use:

- Completely agree
- Somewhat agree
- Neither agree not disagree
- Somewhat disagree
- Completely disagree

Figure A15: Page 11.3 Agree/Disagree

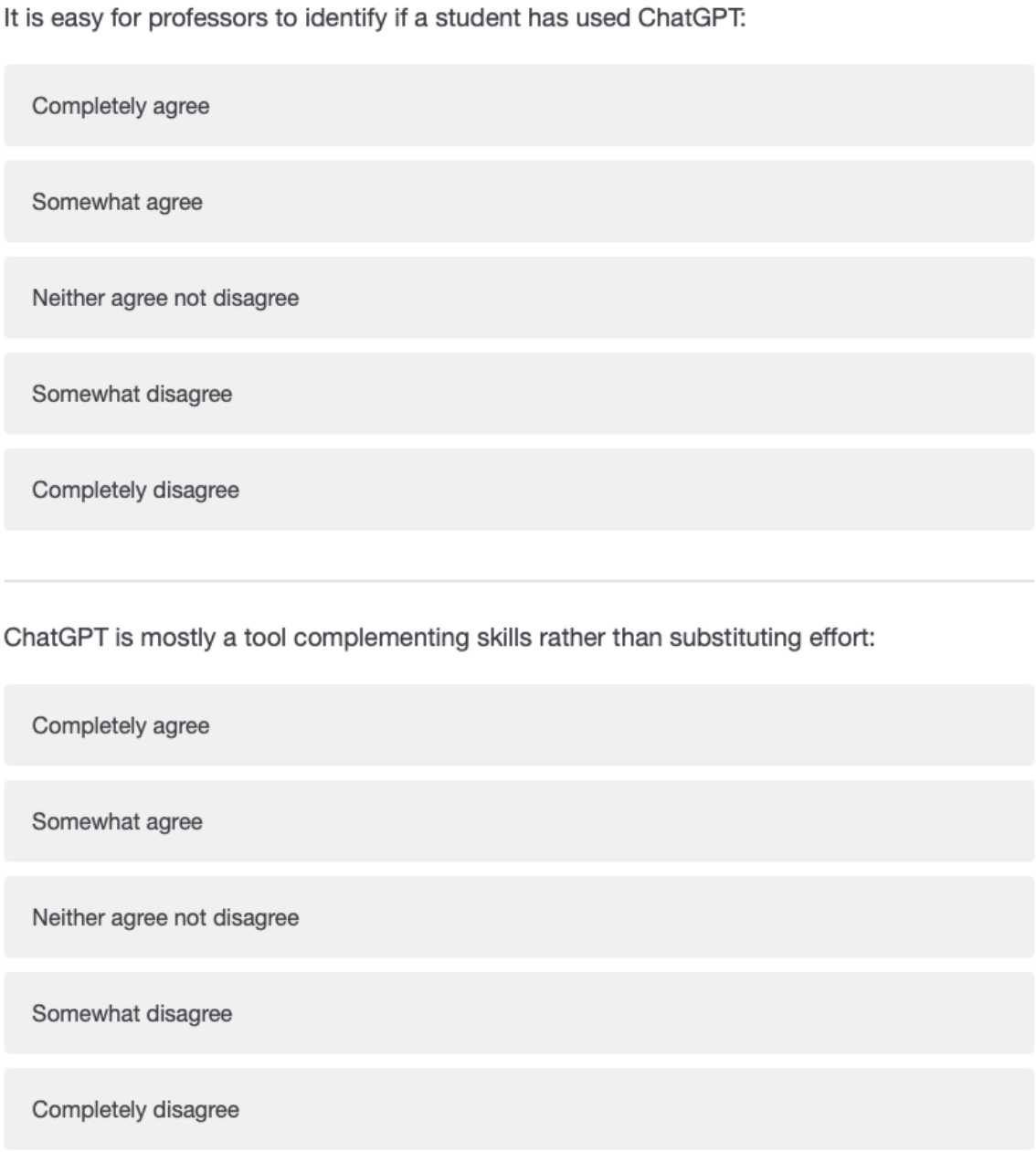
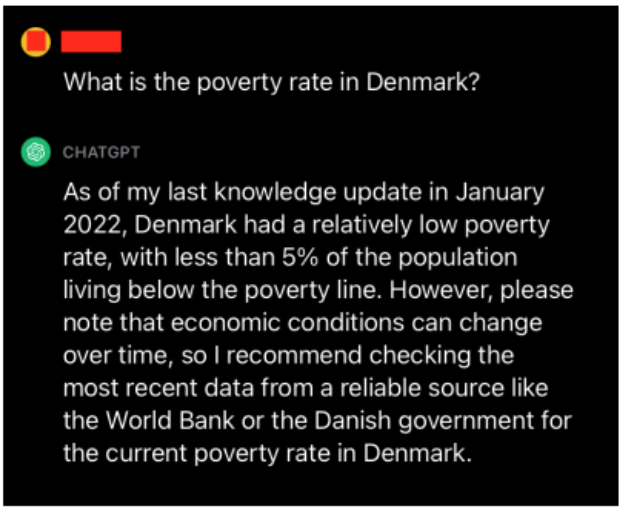


Figure A16: Page 12. Trust accuracy

Below is a screen capture of a query made to ChatGPT, along with the response it provided.



Based on this response from ChatGPT, how much do you trust that the **poverty rate reported** is accurate?

- Completely trust
- Somewhat trust
- Somewhat distrust
- Completely distrust

Figure A17: Page 13. Persistence and inaccuracy

If ChatGPT does not provide the desired answer on your first attempt, how many additional attempts do you typically make?

- None, I move on.
- One more try.
- Two more tries.
- I keep trying until satisfied.
- I don't use it.

Have you ever received inaccurate or misleading information from ChatGPT?

- Yes, many times.
- Yes, few times.
- No, never.
- I don't use it.

Figure A18: Page 14. Subscription and admission grade