Underreporting of AI use: The role of social desirability bias

Alex Imas
University of Chicago
Booth School of Business
alex.imas@chicagobooth.edu

Yier Ling
University of Chicago
Department of Economics
yier.ling@uchicago.edu

Abstract

The integration of artificial intelligence (AI) into work and educational settings is rapidly increasing, yet accurately gauging its adoption remains a challenge. The majority of research uses self-reported surveys. The resulting estimates vary widely, sometimes differing by as much as 40 percentage points in the same setting. This paper studies whether social desirability bias—-the tendency to answer surveys in a way that would be viewed favorably by an outside party—-can potentially explain this discrepancy. We collect data on AI use in a large representative sample of university students. We assess the potential for social desirability bias using a common tool from psychology, *indirect questioning*: all students report both their own AI use and the use of their peers. The data reveals a significant gap, with approximately 60% of students reporting using AI themselves compared to 90% of their peers. In a follow-up study, natural language processing reveals social desirability bias as key driver of the gap between own and others' AI use: students are hesitant to admit AI use due to negative perceptions. This suggests that using self-reports may underestimate the actual prevalence of AI in settings where social desirability bias plays a role, such as education.

Keywords: artificial intelligence, social desirability bias, self-reports

1 Introduction

Artificial intelligence (AI) is no longer a futuristic concept but an increasingly pervasive technology impacting numerous facets of daily life. A non-negligible amount of firms across multiple industries are reported to have an AI adoption rate between 20% to 40%, growing rapidly over time (Crane et al., 2025). In a 2024 in-depth survey of self-reported AI use, almost 40 percent of 18-64 year olds use AI to some capacity, with 9 percent reporting to use it every day at work(Bick et al., 2025). Such AI adoption has impacted numerous industries and labor markets (Tyson and Zysman, 2022). Many have specifically highlighted the potentially transformative role of AI in education. The integration of AI tools in education is becoming more pronounced, with a significant proportion of both teachers and students acknowleding the essential role of AI tools in achieving success in college and future careers (Li and Towne, 2025). While some highlight the potential benefits such as personalized learning experiences and enhanced teaching methodologies (Thomson et al., 2024), others note the threats of AI use. This includes the off-loading tasks that are critical for learning, such as the direct query of facts as well as the process of solving the problems themselves. This underscores the need to accurately assess the extent to which people are adopting and utilizing these technologies.

A majority of studies of AI adoption rely on self-reported surveys. But estimates of AI use vary widely, sometimes varying by as much as 40 percentage points within the same time frame and setting. For example, in the case of education, a Wiley survey conducted in July 2024 reports that 45% of students reporting having used AI in their classes in the past year (John Wiley & Sons, 2024). This contrasts with the results from the Digital Education Council Global AI Student Survey 2024, which reports 86% of students indicating regular use of AI (Digital Education Council, 2024). This discrepancy may be driven by bias in self-reports. Given the often charged conversation around AI, one particularly important source of such bias is *social desirability bias*, which is defined as the tendency to answer in a way that is perceived as positive or acceptable by others instead of truthfully (Krumpal, 2013).

Misreporting on sensitive topics due to social desirability bias is a common occurrence in surveys, often driven by a desire to avoid embarrassment in the presence of an interviewer or to prevent potential repercussions from third parties (Tourangeau and Yan, 2007). This bias can manifest as an overreporting of socially favorable attitudes and behaviors, such as compliance with medical advice or adherence to safety measures, and an underreporting of less desirable ones (Miller, 2020). Prior work has shown that social desirability bias is particularly relevant in educational settings (Bowman and Hill, 2011; Nauta, 2007).

Why would social desirability bias impact reports on AI usage? "AI shaming", involving beliefs that AI-assisted content is deceitful and lacks creativity, has increasingly been documented across several occupations such as researchers and academic writers (Giray, 2024). Research has shown that social image concerns discourage actual AI use in workplace settings, even when such concerns are not instrumental (Almog, 2025). In the context of education, the use of AI tools is also linked to concerns about ethics, academic honesty, plagiarism, and trustworthiness (Kostopolus, 2025; Cotton et al., 2024; Rodrigues et al., 2025). The act of declaring AI use may also be seen as an admission of not being able to complete the work independently, potentially leading to a perception of lower academic ability (Gonsalves, 2024). As a result, in this quickly-evolving landscape, it may be the case that such beliefs are generating norms around (not) using AI, which implies that social desirability bias would discourage self-reporting of AI usage.

We present two studies to explore whether this is indeed the case. To address the challenges posed by social desirability bias in gauging AI use, we follow the methodology in psychology and sociology that uses indirect questioning to elicit more truthful responses on sensitive topics (Fisher, 1993). One such technique involves asking respondents about the behavior of others, such as their peers or friends, rather than directly inquiring about their own actions. The rationale behind this approach is that individuals might feel more comfortable attributing potentially undesirable behaviors to others, thus reducing the pressure to present themselves in a socially desirable way (Fisher, 1993). For example, instead of asking "Have you ever used AI for an assignment?", a researcher might ask "How many of your close friends have used AI for an assignment?".

We adopt the indirect questioning methodology in a large representative survey of students at a medium-sized Midwestern university. In the survey, students are asked detailed questions about their own AI use and the same questions about the AI use of peers within their social circle. A dramatic gap in usage emerges: while approximately 60% report using AI tools, this number increases to nearly 90% when asked about the use of others. Specifically, around 40% of students report their AI use as "Not at all" compared to 10% in the case of their peers. The majority of students report using AI either just "a little" or "not at all"; this is in contrast to others' use, where the two largest categories are using AI "a moderate amount" or "a lot". The gap between own and other AI use appears in nearly every dimension of AI use. For example, looking at frequency of AI use, the largest reported category of own use is 0-1 days a week while the largest category of others' use is 4-5 days a week. The gap in own versus others' AI use is largely stable across gender, year in college, major, and other demographic variables.

We collect data from a separate sample of students to explore the source of the gap. After being presented with

information about the gap, students are asked for the reason behind it both using direct questions and free-form responses. In direct questions, the vast majority of students say that the gap is due to people underreporting their own use relative to others. By far the most selected reason for the gap corresponds to social desirability bias. Specifically, "Students are embarrassed to admit they use LLMs, but are okay saying that their friends do." A similar picture emerges in free responses. Using unsupervised generative AI allows us to classify responses without imposing specific categories. This method generates five broad categories. The vast majority of responses—three times as many as the next largest category—can be classified as the underreporting of one's own AI use due to social desirability bias.

Our results have two main implications. First, they suggest that, depending on the setting, norms against AI use have already been established. These norms may act as an important barrier for organizations that want to encourage AI adoption. Second, our findings suggest that current statistics relying on direct self-reports may underestimate the actual prevalence of AI in settings where social desirability bias plays a role, such as education. This highlights the necessity for methodologies that account for these inherent biases to obtain a more accurate understanding of AI's transformative role in society.

2 Results

2.1 Representative student survey on AI Use

All students in the sample answered a series of questions about their own and others' reliance on AI, as well as questions on demographics and major. We focused on AI in the form of Large Language Models (LLMs), which are by far the most prevalent forms of AI in the setting we studied. Questions thus referred to the use of LLMs, and also included queries on the specific models used (e.g., ChatGPT 4o, Claude, etc.). The full set of questions is listed in the Methods section. Our main questions of interest are (*i*) Do you/peer ever use LLMs (not at all, a little,...,a great deal) and (*ii*) the frequency of LLM use per week (0-1 days, 2-3 days,...,6-7 days).

Figure 1 reports results on the first question, comparing own versus other reliance on AI. Looking at self-reported own reliance, most students suggest that they do not rely on AI at all (38.5%), followed by "A little" (23.7%). This implies that slightly more than 60% of students report that they rely on AI to any extent; combining across the first two categories, the majority of students report using AI either a little or not at all. The picture of AI use changes dramatically when students are asked about the use of their peers. Students report that only 11.2% of their peers do not rely on AIs at all, the lowest across all levels. Rather, students report that most other students rely on LLMs "A moderate amount" (28.4%), followed by "A lot" (24.1%) (figure 1). This implies that nearly 90% of their peers use AI to some extent, with the majority using AI either a moderate amount or a lot. The difference in the distributions of own versus others' AI use is highly significant (Table A1).

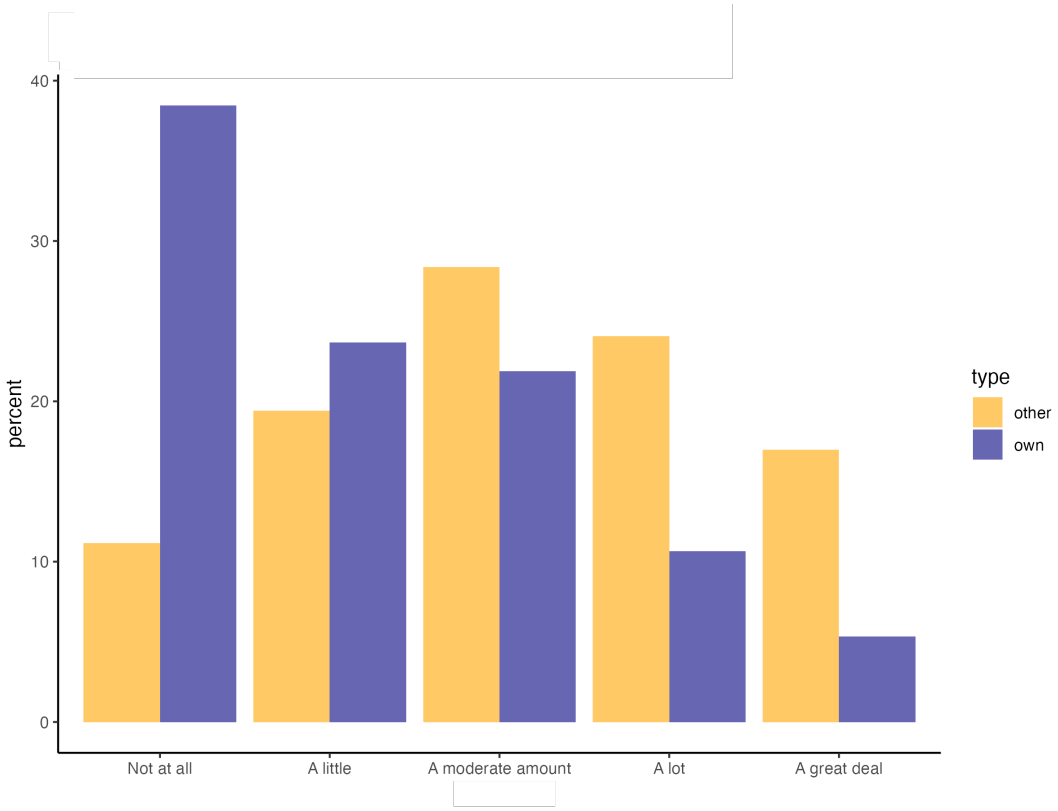


Figure 1: **Self-reporting vs. other-reporting on AI reliance**

Figure 2 reports results on the second question regarding frequency of AI use. Here, we see a similar gap emerge as on the reliance question. Most students self-report to only use AI 0-1 days a week (42%)—by far the largest category. The other categories ("2-3 days", "4-5 days", and "6-7 days" per week) each receive about 20% of responses. This picture of AI use changes dramatically when students are asked about the use of their peers. Now by far the largest category is "4-5 days" a week (47.3%), followed by nearly every day ("6-7 days"; 26.3%). Fewer than 1% of students report that their peers use AI only "0-1 days" a week. The difference in the frequency distributions of AI use between own versus others' is highly significant (Table A1). Notably, neither the gap in overall use nor the frequency of use was reliably predicted by the demographic variables collected nor the major choice.

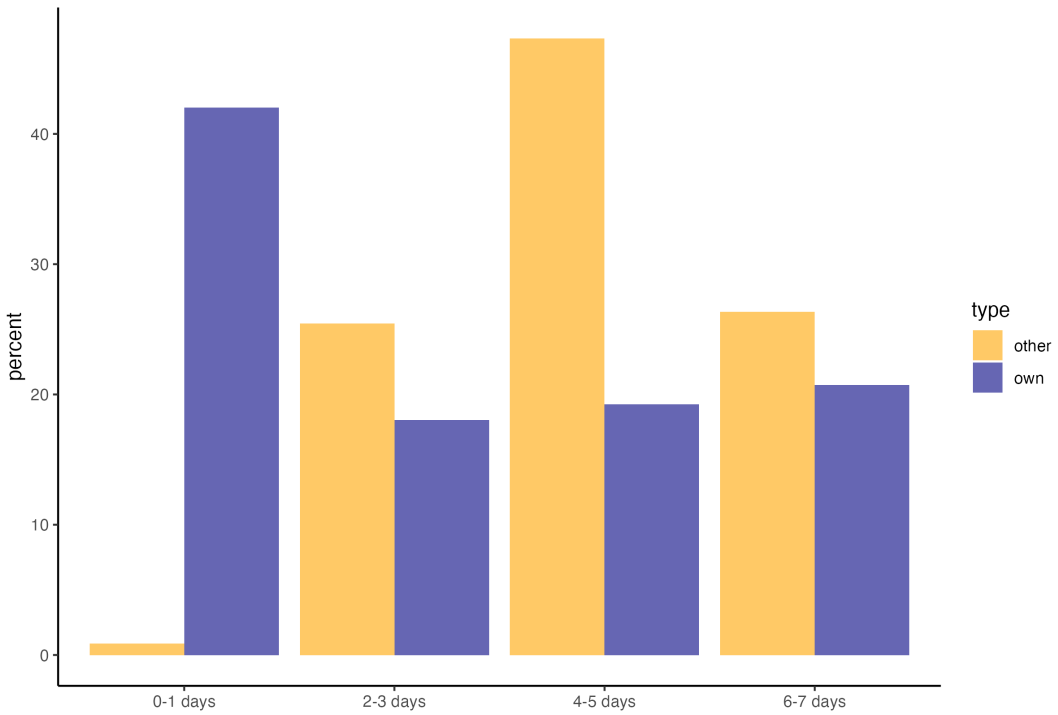


Figure 2: **Self-reporting vs. other-reporting on AI use frequency**

2.1.1 Heterogeneity Analysis

Here we explore whether the own versus other AI reliance gap differs across demographics or area of study. Prior research has shown a substantial gender-based difference in AI use, with women reporting a significantly lower propensity to use AI (Carvajal et al., 2025). While we see a smaller gap in own versus other AI use among male than female students (6%), the difference is not significant (Table A2). The own versus other gap in the *extent* of AI use is similar across genders as well (Table A2). For area of study, we see a nearly 10% higher self-reported reliance on AI among STEM majors compared to non-STEM majors (Table A3). However, the gap in own versus others' AI reliance is similar across both STEM and non-STEM majors (approximately 30%; Table A3).

2.2 Exploring the mechanism

The results from the representative survey paint a clear gap between self-reports of own and others' AI use. We now proceed to explore the potential reasons for this gap. In a follow-up survey of students, we presented information about this gap and elicited reasons for it both through direct questions and free response.

2.2.1 Direct questions

Table 1 lists the specific questions asked and the percent of people who select each answer. The first set of questions asks whether the own vs. other AI use gap is due to students underreporting their own use or overreporting the use of others. The vast majority (79%) report that the gap is due to underreporting of own use. When asked to provide a reason for the gap, by far the largest category of answers is consistent with social desirability bias (70%); the next largest category ("Students don't know how much their friends are using LLMs") received less than one third as many responses.

Table 1: **Direct questions about own vs. other AI use gap**

Question	Percent	Response
Reason	70.8	Students are embarrassed to admit they use LLMs, but are okay saying that their friends do
	20.8	Students have no idea how much their friends are using LLMs
	6.3	Students are inflating how much others are using LLMs
	2.1	None of the above
Under- vs. Over-reporting	79.2	Students under-report how much they use LLMs
	11.5	Students over-report how much others use LLMs
	9.4	Neither

2.2.2 Free response

Results from the free responses corroborated the findings from the direct questions. We use unsupervised generative AI to recover the reasons given for the own vs. other AI use gap. This method allows us to classify responses without prompting the model to use any pre-specified categories. This method generates five broad categories that are reproduced in Figure 3. The vast majority of responses can be classified as the underreporting of one's own use due to social desirability bias (58.1%), followed by overestimating others' AI usage (18.3%).

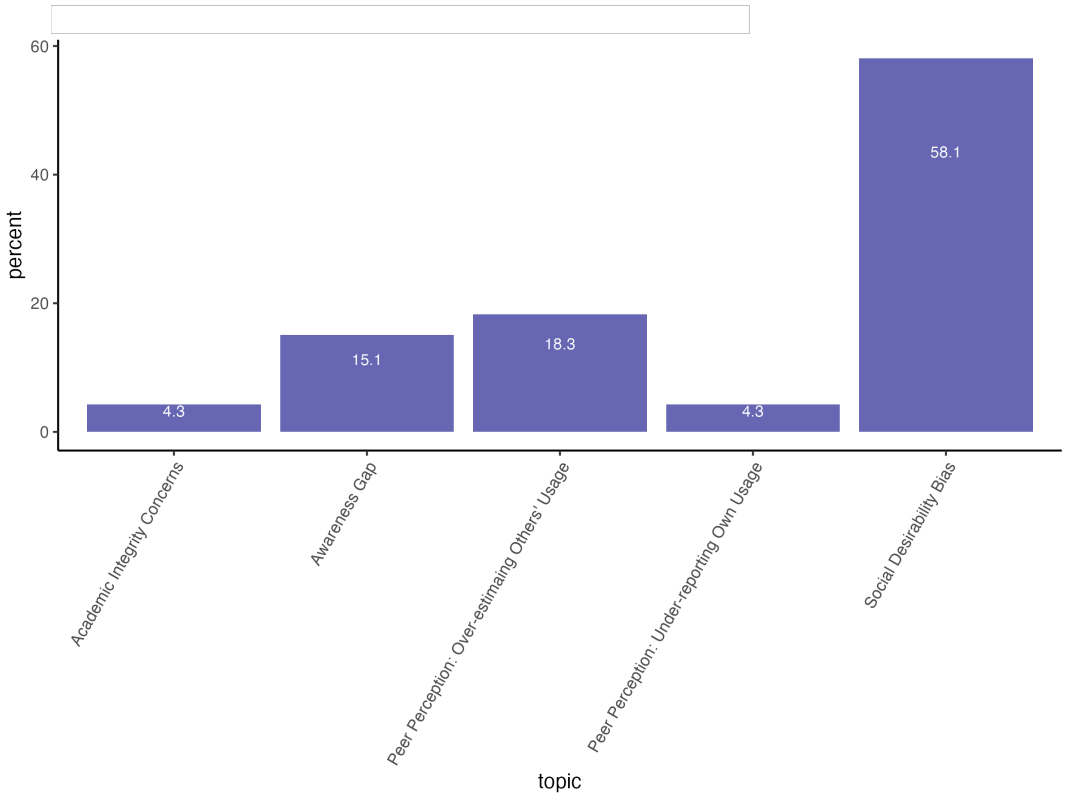


Figure 3: **Reasons for the own vs. other AI use gap**

3 Methods

In this section we illustrate the methodology behind the two surveys: the representative survey examining LLM use; and follow-up experiment examining the mechanism for the own versus other AI use gap.

3.1 Representative student survey on AI Use

3.1.1 Participants

We recruit undergraduate students at medium-sized Midwestern university in collaboration with its behavioral research center, who reached out to students via multiple methods on campus. Student participants were given a gift card for completing the survey, and a university email was required for the gift card. After survey data was collected, participants indicating that they are not undergrads (either reporting that they are graduate students or choosing "Other" in terms of their year in college without providing more details) or that they are studying majors not offered by the university were removed. The final remaining sample size was 338.

Among the 338 participants, around 76.3% of them reported at least one STEM major; 77.5% of the sample has at least one STEM major or minor. Participants were about evenly distributed across years in college, ranging from 20.7% in the third year to 28.1% in the first year. Among participants who reported their gender identity, slightly over half of the sample identify as female, while around 4.2% of the participants identify as Non-binary or other gender. Among participants reporting their racial identity, Asian students make up the largest racial group in this sample (36.1%), followed by White students (33.5%), Other and Multi-racial students (20.7%) and African American

students (8.8%), with less than 1% of Native American or Pacific Islander students. A non-trivial minority of the sample was composed of international students (26.6%). Among participants reporting their family income, 43.5% of participants report an annual household income over $150,000, followed by $25,000 to $50,000 (12.5%) and $75,000 to $100,000 (12.2%).

3.1.2 Survey Design

This survey consists of two major components examining the participant's own use of AI (in the form of LLMs) and their belief of others' usage of LLMs. After the welcome page, all participants were asked if they ever used an LLM. If they answered "yes", they are then directed to answer a series of additional questions about LLMs: how reliant they are on LLMs; which LLM models they use; whether they pay for LLM models, main reasons for using LLMs; how many days a week they use LLMs; what other AI tools they use (if any); and whether they use the built-in features of LLMs. These questions constitute the self-reported own use part of the survey. Each participant also answered the same set of questions about average peers in their social group (not any specific person).

Finally, participants complete a set of demographic questions asking about: their majors (and minors if applicable); year in college; their study devices; whether they are international students; whether English is their first language; their ideal prospects after college; how much they work in major-related or major-unrelated fields part-time during a week (if any; in two separate questions); gender; family income; and race.

3.1.3 Analysis Method

A list of variables is outlined in Appendix Table A4. We used regressions to test for the difference in distributions between own versus others' AI use, and explore heterogeneity in this gap.

3.2 Exploring the Mechanism

3.2.1 Participants

For the follow up study on exploring the mechanism, we recruited a separate group of participants from *Prolific*. We restricted the sample to those who reported being an undergraduate student—all others were screened out. The final sample was 96 participants.

3.2.2 Survey Design

In this survey, we first described the findings from the representative student survey, showing both the percentage of students' own self-reported reliance and frequency of AI use as well as their answers about peer use. All participants were then asked three questions about this gap. The first question asked whether they believed this gap was due to the overreporting of others' use or the underreporting own students' own use. The second question asked participants to choose the most likely explanation: students being embarrassed to admit LLM use; students inflate others' use; students are unaware of others' use, and none of the above. The last question was a free-text entry directly asking for their explanation of the gap.

3.2.3 Analysis Method

We used unsupervised generative AI to classify free responses into interpretable categories without imposing categories on the model. To do so, we set up a *Claude (3.7 Sonnet)* API and performed topic modeling for each of the entries.

We first asked the model to go through all responses and create a list of topic categories. This returned the following list of topics: "Social Desirability Bias", "Peer Perception", "Academic Integrity Concerns", "Self-Report Bias", "Awareness Gap", "Privacy Concerns". To further disambiguate "Peer Perception", we asked the model to split it into "Over-estima[t]ing Others' Usage (peer perception)" and "Under-reporting Own Usage (peer perception)".

Next, for each entry, we prompted the model to choose one of topics from the list above that is the most representative of the response. We provide a prompt fully specifying the context and providing definitions of each topic. We repeat this procedure 30 times for each entry, and choose the most frequent topic assigned by *Claude* as the response's topic. If more than one topic are assigned equally frequently, each response-topic pair is treated as a separate data point (e.g., if a response is assigned "Awareness Group" 15 times and "Privacy Concerns" 15 times, we include both topics and the response appears in the eventual dataset twice). One response that is not assigned any topic throughout the 30 times is removed.

4 Discussion

The findings of this study reveal a substantial disparity between self-reported AI usage and their perception of peer AI usage. We provide evidence that this gap is primarily driven by social desirability bias, where norms against AI use lead people to underreport their own usage relative to others. The significant discrepancy observed in this study has important implications for the accuracy of current statistics on AI usage, particularly in settings where social desirability bias is likely to play a role. Specifically, in the context of our study, the indirect questioning method suggests that the prevalence of AI use in educational settings is magnitudes bigger than what is reported in some surveys. The underestimation from self-reports can have significant consequences for educational policy, curriculum design, and the provision of institutional support for AI literacy. If policymakers and educators rely on inaccurate data, they may fail to adequately address the real needs and challenges associated with the widespread integration of AI in higher education. For instance, institutions might underestimate the demand for AI literacy training or fail to develop comprehensive guidelines that reflect the actual ways students are utilizing these tools.

One important question is why did the norm against AI use develop? In theory, the adoption of new technology could be viewed as an advantage, a sign of being current with the pace of progress. The fact that—at least in the educational setting—AI use appears to be stigmatized suggests a potential barrier to effective adoption. If people are not willing to admit that they use the tool, then this will dampen potential network effects and efforts to improve human-AI synergies. It would therefore be useful for both technology firms and policy makers to better understand the source of stigma to both measure prevalence of AI use and to make the tool more effective. The indirect questioning methodology described here can be used to identify where stigma is likely to be present so that it could be targeted accordingly.

References

Almog, D. (2025). AI recommendations and non-instrumental image concerns. *Available at SSRN 5226740*.

Bick, A., A. Blandin, and D. J. Deming (2025). The rapid adoption of generative AI.

Bowman, N. A. and P. L. Hill (2011). Measuring how college affects students: Social desirability and other potential biases in college student self-reported gains. *New Directions for Institutional Research 2011*(150), 73–85.

Carvajal, D., C. Franco, and S. Isaksson (2025). Will artificial intelligence get in the way of achieving gender equality?

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

Crane, L., M. Green, and P. Soto (2025). Measuring AI uptake in the workplace.

Digital Education Council (2024). What students want: Key results from dec global AI student survey 2024.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of consumer research 20*(2), 303–315.

Giray, L. (2024). AI shaming: the silent stigma among academic writers and researchers. *Annals of Biomedical Engineering 52*(9), 2319–2324.

Gonsalves, C. (2024). Addressing student non-compliance in AI use declarations: implications for academic integrity and assessment in higher education. *Assessment & Evaluation in Higher Education*, 1–15.

John Wiley & Sons, I. (2024). Ai has hurt academic integrity in college courses but can also enhance learning, say instructors, students.

Kostopolus, E. (2025). Student use of generative AI as a composing process supplement: Concerns for intellectual property and academic honesty. *Computers and Composition 75*, 102894.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity 47*(4), 2025–2047.

Li, D. H. and J. Towne (2025, Jan). How AI and human teachers can collaborate to transform education.

Miller, P. R. (2020). Tipsheet – sensitive questions.

Nauta, M. M. (2007). Assessing college students' satisfaction with their academic majors. *Journal of career assessment 15*(4), 446–462.

Rodrigues, M., R. Silva, A. P. Borges, M. Franco, and C. Oliveira (2025). Artificial intelligence: Threat or asset to academic integrity? a bibliometric analysis. *Kybernetes 54*(5), 2939–2970.

Thomson, S. R., B. A. Pickard-Jones, S. Baines, and P. C. Otermans (2024). The impact of AI on education and careers: What do students think? *Frontiers in Artificial Intelligence 7*, 1457299.

Tourangeau, R. and T. Yan (2007). Sensitive questions in surveys. *Psychological bulletin 133*(5), 859.

Tyson, L. D. and J. Zysman (2022). Automation, AI & work. *Daedalus 151*(2), 256–271.

Appendix - Tables

Table A1: Reporting Gap - Regression

	Dependent variable:	
	Reliance	Frequency
	(1)	(2)
Own	0.2731***	-0.8047***
	(10.7001)	(-13.1320)
Constant	0.1115***	1.9911***
	(15.0284)	(49.1089)
Observations	599	676
R^2	0.116	0.142
Adjusted R^2	0.114	0.141

*p<0.1; **p<0.05; ***p<0.01

Note: Own reliance is binary with value 1 ("None/Not at all") or 0 (any other extent); Other reliance is the probability assigned to level "None/Not at all". Both own and other frequency are scored 0 to 3 for using LLMs "0-1 day", "2-3 days", "4-5 days", and "6-7 days". t-statistics reported in the parentheses. Observations indicate the number of responses.

Table A2: Heterogeneity Regressions - Gender

	Dependent variable:	
	Reliance	Frequency
	(1)	(2)
Own	0.2277***	-0.5902***
	(5.5574)	(-5.6430)
Female	-0.0183	0.0771
	(-1.2252)	(0.8920)
Own × Female	0.0658	-0.3067**
	(1.2363)	(-2.3351)
Constant	0.1165***	1.9590***
	(9.7192)	(28.7009)
Observations	558	632
R^2	0.115	0.140
Adjusted R^2	0.111	0.136

*p<0.1; **p<0.05; ***p<0.01

Note: Own reliance is binary with value 1 ("None/Not at all") or 0 (any other extent); Other reliance is the probability assigned to level "None/Not at all". Both own and other frequency are scored 0 to 3 for using LLMs "0-1 day", "2-3 days", "4-5 days", and "6-7 days". SEs are clustered at participant level. t-statistics reported in the parentheses. Observations indicate the number of responses.

Table A3: Heterogeneity Regressions - Major

	Dependent variable:	
	Reliance	Frequency
	(1)	(2)
Own	0.3397***	-0.9375***
	(6.4791)	(-9.0160)
STEM	-0.0144	0.0866
	(-0.6779)	(0.9424)
Own × STEM	-0.0876	0.1739
	(-1.4605)	(1.3662)
Constant	0.1228***	1.9250***
	(6.1927)	(24.3970)
Observations	599	676
R^2	0.123	0.148
Adjusted R^2	0.118	0.144

*p<0.1; **p<0.05; ***p<0.01

Note: Own reliance is binary with value 1 ("None/Not at all") or 0 (any other extent); Other reliance is the probability assigned to level "None/Not at all". Both own and other frequency are scored 0 to 3 for using LLMs "0-1 day", "2-3 days", "4-5 days", and "6-7 days". SEs are clustered at participant level. t-statistics reported in the parentheses. STEM is defined as studying at least one of the following fields as a major: mathematics, engineer, CS, fields with "bio", fields with "stat", fields with "astro", fields with "physic", fields with "neuro", fields with "chem", data science, fields with "cognitive", fields with "psych", variations of "pre-med", economics (excluding business economics), fields with "enviro", and a combined major of environmental science and urban studies. Observations indicate the number of responses.

Table A4: Variable Table

Variable	Explanation
Representative student survey	
own_rely	Self-reported reliance on LLMs
own_frequency	Self-reported frequency using LLMs
other_rely_dist_level_1	Peer-reported reliance on LLMs: *None at all* (fraction in a distribution)
other_rely_dist_level_2	Peer-reported reliance on LLMs: *A little* (fraction in a distribution)
other_rely_dist_level_3	Peer-reported reliance on LLMs: *A moderate amount* (fraction in a distribution)
other_rely_dist_level_4	Peer-reported reliance on LLMs: *A lot* (fraction in a distribution)
other_rely_dist_level_5	Peer-reported reliance on LLMs: *A great deal* (fraction in a distribution)
others_frequency	Peer-reported frequency using LLMs
year	Year in college
first_language_english	Whether English is the first language
international	Whether international student
stem_major	Whether any of major(s) is STEM
stem_any	Whether any of major(s) or minor(s) is STEM
gender	Gender
race	Race
family_income	Family income
Prolific survey	
free_response	Free-form text response on reasons behind the gap
under_over_reporting	Direct question: whether the gap is due to underreporting of own use of LLMs or overreporting of others' use of LLMs
reason	Direct question: which reason is behind the gap