

DOI: [https://doi.org/10.48009/1\\_iis\\_2026\\_109](https://doi.org/10.48009/1_iis_2026_109)

## Generative AI availability and metacognitive calibration: Evidence from higher education

Adam Patterson, *Amazon*<sup>1</sup>, [adapattk@amazon.com](mailto:adapattk@amazon.com)  
Oskar Harmon, *University of Connecticut*, [oskar.harmon@uconn.edu](mailto:oskar.harmon@uconn.edu)

### Abstract

The emergence of generative artificial intelligence (GAI) tools such as ChatGPT has transformed higher education, yet their influence on students' metacognitive calibration, the alignment between self-assessed confidence and actual competence, remains underexplored. Understanding whether GAI availability affects confidence–performance dynamics is critical for instructional design in technology-enhanced learning environments. This study analyzes 1,157 student responses from Excel literacy surveys administered at a large R1 public university between September 2021 and December 2024, spanning periods before and after ChatGPT's public release in November 2022. Using a quasi-experimental design with OLS regression and robust standard errors, we compare self-reported confidence, objective performance, and calibration (overconfidence) across pre- and post-release cohorts, controlling for student demographics and academic characteristics. We document a statistically significant increase in student confidence (0.19 SD,  $p < 0.001$ ) with no corresponding change in actual performance, producing a meaningful widening of the overconfidence gap. Heterogeneity analyses reveal consistent effects across student subgroups. However, placebo tests at false treatment dates yield effect sizes comparable to the main specification, suggesting that broader temporal trends may partially explain the observed patterns. Rather than establishing causal effects of GAI availability, this study documents a descriptive pattern of confidence–performance decoupling that coincides with the GAI era. The findings highlight important considerations for educators designing technology-enhanced learning environments: students may develop inflated self-assessments of independent competence as AI tools become ubiquitous, necessitating deliberate instructional strategies to support metacognitive calibration.

**Keywords:** generative AI, higher education, metacognitive calibration, technology-enhanced learning, Dunning-Kruger effect

### Introduction

The emergence of sophisticated generative artificial intelligence (GAI) tools has fundamentally transformed the educational landscape. OpenAI released ChatGPT in November 2022, providing students with unprecedented access to AI assistance for academic tasks. While considerable attention has focused on direct AI usage, students employing these tools to complete assignments or answer questions, a more subtle but potentially consequential question remains underexplored: Is the mere availability of GAI tools associated with changes in students' confidence in their own abilities?

---

<sup>1</sup> This work does not relate to the author's position

This question holds profound implications for educational technology in higher education. If students become more confident in their skills simply because GAI tools are available, yet their actual competence remains unchanged, educators face a critical metacognitive challenge. Overconfident students may underinvest in genuine skill development and struggle to accurately self-assess their learning needs. Understanding this relationship is particularly important given the rapid adoption of AI tools in educational contexts (Frydenberg et al., 2026) and the growing emphasis on technology-enhanced learning environments across higher education institutions.

While ChatGPT's sudden public release creates a natural experimental opportunity, concurrent pandemic-era changes, such as a return to in-person learning and other temporal shifts, complicate causal interpretation. Our analysis, therefore, proceeds as a descriptive investigation of temporal patterns in confidence–performance calibration, with results interpreted cautiously given these identification challenges.

This study leverages ChatGPT's release to examine whether students' self-assessed confidence shifts when GAI assistance becomes available for academic tasks. Our investigation assumes a similar GAI adoption pattern found in prior research at higher learning institutions (Frydenberg et al., 2026), creating an exogenous event to study variations in student confidence and performance. Using administrative data on student performance combined with pre- and post-release survey-based confidence measures, we employ a quasi-experimental design to identify temporal patterns. While this design approximates a natural experiment, potential confounding factors and unobserved heterogeneity limit our ability to identify causal effects. Accordingly, our findings should be interpreted as associations between GAI availability and changes in student confidence.

Specifically, this study examines whether the widespread availability of generative AI tools is associated with changes in students' confidence in their own technical skills, independent of actual performance. Using objective Excel task data paired with self-reported confidence measures collected before and after ChatGPT's public release, we document a meaningful increase in student confidence without corresponding improvements in skill mastery. These findings highlight a metacognitive challenge for educators in technology-enhanced learning environments: AI availability may inflate students' perceived competence in foundational quantitative skills central to workplace readiness.

## Literature Review

### Technology Adoption and Generative AI in Higher Education

The rapid integration of artificial intelligence tools into educational settings has prompted scholars to examine the factors influencing technology adoption among students and faculty. The Unified Theory of Acceptance and Use of Technology (UTAUT), developed by Venkatesh et al. (2003), provides a comprehensive theoretical framework for understanding how individuals adopt new technologies. In the context of higher education, a systematic review of 162 UTAUT studies found that performance expectancy consistently demonstrated the strongest influence on students' intention to adopt educational technologies, highlighting that learners prioritize tools they believe will enhance their academic performance (Xue et al., 2024). Research applying UTAUT specifically to GAI tools in higher education settings revealed that productivity gains, mentor perspective, peer usage, and breadth of tasks performed drive generative AI adoption in academic settings (Patterson et al., 2024). This study contributes to this prior work by studying adoption outcomes, such as calibration, rather than drivers of adoption.

ChatGPT's release created a natural experiment for studying technology adoption in education. A survey of 399 undergraduate and postgraduate students across multiple disciplines found generally positive attitudes toward GenAI for teaching and learning, with students recognizing the potential for personalized learning support, writing assistance, and research capabilities, yet simultaneously expressing concerns about accuracy, privacy, and the impact on personal development and independent skill-building (Chan & Hu, 2023). Research by Strzelecki (2024) applied the UTAUT framework to examine ChatGPT adoption among Polish university students, finding that performance expectancy, effort expectancy, and social influence significantly predicted adoption intention. Cross-cultural research by Budhathoki et al. (2024) extended this analysis to compare ChatGPT adoption in the United Kingdom and Nepal, revealing that while these core UTAUT factors remained significant across both contexts, the role of anxiety varied substantially between countries. These findings suggest that cultural and institutional factors moderate technology adoption patterns, underscoring the importance of studying technology adoption through natural experiments that capture real-world adoption conditions rather than relying on hypothetical scenarios.

### **ChatGPT's Mixed Effects on Learning Performance and Student Engagement**

Meta-analytic evidence on ChatGPT's impact on learning outcomes presents a nuanced picture. Wang and Fan (2025) conducted a meta-analysis of experimental studies, finding that ChatGPT interventions produced large positive effects on academic performance ( $g = 0.867$ ) and moderate improvements in learning perception ( $g = 0.456$ ) and higher-order thinking ( $g = 0.457$ ). However, these overall effects masked substantial heterogeneity: ChatGPT's effectiveness was significantly moderated by course type, instructional model, and intervention duration, suggesting that benefits are not uniform across educational contexts. Additionally, recent meta-analyses of 51 studies published between November 2022 and February 2025 emphasize that methodological limitations in existing research, including insufficient power analyses, inadequate long-term follow-up, and reliance on post-intervention assessments, warrant cautious interpretation of reported effects.

Not all studies document positive associations between ChatGPT use and academic outcomes. Yang et al. (2025) conducted a quasi-experimental study of high school students learning programming and found that students using ChatGPT reported significantly lower self-efficacy and learning achievement compared to those receiving conventional instruction methods. Post-hoc interviews revealed that students perceived ChatGPT's effectiveness fell short of expectations, suggesting a divergence between perceived ease of use, perceived usefulness, and actual learning outcomes.

### **Calibration, Overconfidence, and the Dunning–Kruger Effect**

A fundamental question in educational psychology concerns the relationship between students' confidence in their abilities and their actual performance, a relationship often termed calibration. Accurate calibration is essential for effective self-regulated learning, as miscalibrated students often employ ineffective study strategies. A longitudinal study of introductory biology students demonstrated that calibration improves over time with repeated feedback and self-reflection, yet students exhibited significant initial miscalibration, with predicted exam scores exceeding actual performance (Osterhage et al., 2019). The lowest-performing students demonstrated the greatest overconfidence, a phenomenon extensively documented in the literature and formalized as the Dunning–Kruger effect.

The Dunning–Kruger effect, replicated across numerous educational contexts, describes the tendency of individuals with limited competence to systematically overestimate their abilities (Kruger & Dunning, 1999). Furthermore, research by Miller and Geraci (2011) found that miscalibration and confidence remained stable across tests, suggesting that low-performing students continued to believe that they would perform well on upcoming exams despite prior evidence to the contrary.

These findings highlight a critical metacognitive gap: students who most need to engage in remedial study efforts are often the least likely to recognize their knowledge gaps. Influential meta-analyses rank student expectations and self-reported grades among the most impactful factors influencing learning, as documented across over 800 meta-analyses (Hattie, 2008) and later expanded to over 900 (Hattie, 2012). This centrality of calibration to learning outcomes underscores the importance of accurately assessing one's knowledge and abilities for educational success.

Despite this, relatively little research has examined how exposure to GAI tools, which provide immediate and often accurate responses to student queries, might influence calibration dynamics in task-specific contexts, particularly for foundational skills like spreadsheet literacy, where performance is objectively measurable, and feedback mechanisms are well defined. According to Mentzer et al. (2024), spreadsheet knowledge is among the worst-performing digital skills among current first-year college students, making the Excel literacy survey deployed in this study highly relevant to institutions focused on student career development.

Two foundational theoretical frameworks inform the metacognitive dynamics this study investigates. Flavell's (1979) model of metacognitive monitoring distinguishes between metacognitive knowledge (what individuals know about their own cognition) and metacognitive experiences (real-time appraisals of cognitive performance), suggesting that external tools altering cognitive effort may disrupt the feedback loop through which learners update their self-assessments. Self-regulated learning (SRL) theory, as elaborated by Zimmerman (2000), further posits that effective learners cycle through forethought, performance, and self-reflection phases; a cycle that accurate calibration is essential for sustaining. When GAI tools provide ready-made responses, learners may short-circuit the self-reflection phase, arriving at inflated confidence without the corrective mechanism of encountering difficulty. Together, these frameworks suggest that environments characterized by readily available GAI assistance may systematically undermine the metacognitive monitoring processes on which accurate calibration depends.

### **The Confidence–Competence Paradox: Risk of Skill Depletion with AI Assistance**

A critical concern emerging from recent literature is that GAI tools, while enhancing perceived competence and engagement, may paradoxically undermine the development of genuine expertise. Recent experimental evidence suggests that GAI promotes learner dependence on technology and triggers metacognitive laziness, a reduced tendency for users to self-regulate and engage critically in the learning process (Fan et al., 2025). This distinction is crucial because students may develop inflated confidence in their abilities while their foundational skills atrophy from reduced practice.

Fan et al. (2025) conducted a randomized experiment comparing learning with ChatGPT, human experts, and conventional tools, finding that while ChatGPT improved immediate task performance, learners in the AI condition exhibited significantly fewer metacognitive processes such as orientation, planning, and evaluation. Their findings suggest that AI assistance suppresses the self-regulatory behaviors essential for accurate calibration. Empirical evidence from Abbas et al. (2024) further supports this concern, demonstrating that ChatGPT usage among university students is associated with increased procrastination, memory loss, and diminished academic performance, outcomes the authors attribute to reduced cognitive effort and overreliance on automated assistance. These findings underscore the “double-edged sword” nature of GAI adoption: while these tools offer genuine utility, their availability may erode the self-regulatory capacities essential for independent learning. This raises an important empirical question: how does access to GAI tools influence the calibration between students' self-assessed confidence in domain-specific skills and their actual task performance?

Two theoretical perspectives help illuminate the mechanisms through which GAI availability may inflate confidence without improving competence. First, distributed cognition theory posits that cognitive processes are not confined to the individual mind but are shared across people, tools, and artifacts within a system (Hutchins, 1995). When students perceive GAI as a readily available cognitive partner, their functional self-assessment may implicitly incorporate the tool's capabilities, for example, a student who knows ChatGPT can generate a VLOOKUP formula may feel confident in their spreadsheet abilities, even though that confidence reflects the capacity of the human–AI system rather than their independent skill.

Second, recent experimental work introduces the concept of metacognitive laziness, defined as learners' tendency to offload self-regulation and critical reflection onto AI tools rather than actively monitoring their own understanding (Fan et al., 2025). Together, these frameworks predict that GAI availability would increase confidence (by expanding the perceived cognitive system) while leaving independent competence unchanged (by reducing metacognitive engagement), producing exactly the confidence–performance decoupling this study investigates.

### Research Gap and Contribution

While active learning approaches employing application-based Excel instruction have demonstrated improvements in self-reported student engagement (Patterson et al., 2023), little research has examined how the availability of GAI systems might disrupt these benefits. GAI tools might alter calibration dynamics in task-specific contexts.

This gap is particularly important for foundational technical skills like spreadsheet literacy, where objective performance measurement and clear feedback mechanisms exist. According to Mentzer et al. (2024), spreadsheet knowledge represents one of the weakest digital competencies among first-year college students, yet it remains essential for successful labor market outcomes in quantitative disciplines and professional contexts (Church et al., 2022).

ChatGPT's release provides a unique opportunity to examine whether students' task-specific confidence in Excel literacy became miscalibrated with their actual performance in the post-ChatGPT era. If the availability of GAI tools increases students' confidence in their ability to perform spreadsheet tasks, due to the psychological reassurance of having an accessible problem-solving resource, without corresponding increases in independent practice and skill development, we would expect a widening gap between confidence and competence.

Research by Urban et al. (2024) identifies a potential mechanism showing that students' perceived usefulness of ChatGPT is associated with higher overestimation of the quality and originality of answers. Such confidence gains from tool availability, when unaccompanied by genuine skill mastery, align with the overconfidence gap that prior literature links to poor self-regulated learning and academic underachievement (de Bruin et al., 2017; Dunlosky & Rawson, 2012; Pintrich et al., 2000).

Taken together, prior research suggests that while generative AI tools are rapidly adopted and can influence learning outcomes, their effects on students' confidence and confidence–performance calibration, particularly in skill-based, business-relevant contexts, remain insufficiently understood. In response to this gap, the present study empirically examines whether the widespread availability of GAI tools coincides with changes in students' confidence relative to their objectively measured performance in spreadsheet literacy. The following section describes the data, measures, and empirical approach used to investigate this question.

## Methodology

### Source Data

Our sample derives from Excel literacy surveys administered to undergraduate students, predominantly freshmen and sophomores enrolled in business and economics courses, at a large R1 public university between September 2021 and December 2024. The nine-question survey (Table A1) assessed students' Excel skills, controlling for confidence, prior spreadsheet skills, prior courses covering spreadsheets, and a set of demographic and academic covariates. Our treatment variable is a binary indicator equal to 1 if the survey was completed after November 30, 2022 (ChatGPT's public release date), and 0 otherwise. This creates a sharp temporal division in our sample. In total, 1,157 observations were collected (736 pre-ChatGPT) over seven academic semesters. Survey administration procedures were held consistent across all semesters: students completed the Qualtrics instrument during class time on a laptop or desktop computer, with identical instructions and task material. No substantive changes were made to survey wording or structure over the observation window, supporting comparability across cohorts.

### Metric Construction

We convert the raw survey results into three primary metrics for analysis: correct, confidence, and calibration. Original confidence responses are transformed from categorical to numeric, with 1 = don't know, 2 = unsure, 3 = pretty confident, and 4 = very confident. The correct variable is an indicator for each survey question, taking the value of 1 for a correct response and 0 for an incorrect response. Calibration (overconfidence) for each student  $i$  is calculated as:

$$Calibration_i = \sum_{j=1}^J Confidence_j - \sum_{j=1}^J Correct_j$$

where  $J$  is the number of total questions asked in the survey, Confidence is the self-reported confidence for question  $j$ , and Correct is an indicator denoting whether the student answered question  $j$  correctly. For this research, calibration and overconfidence refer to the same measure. Calibration captures the gap between total confidence and demonstrated ability. Positive values indicate overconfidence (confidence exceeds demonstrated ability), negative values indicate underconfidence (demonstrated ability exceeds confidence), and zero represents perfect calibration. Student responses with less than 240 seconds (4 minutes) duration are omitted from the analysis to avoid straight lining. To illustrate: a student who reports being "pretty confident" (3) on all seven survey questions but answers only three correctly would have a calibration score of  $21 - 3 = 18$ , indicating substantial overconfidence. Conversely, a student who answers five questions correctly while reporting low confidence throughout would produce a lower score, reflecting lower levels of overconfidence.

### Empirical Strategy

We estimate treatment effects using OLS regression, employing a difference-in-means framework with robust standard errors. Our baseline specification compares outcomes before and after ChatGPT's release. The full controls specification adjusts for student characteristics, including Excel skills, GPA, gender, ethnicity, and class standing. We exclude survey duration from controls due to reverse causality concerns, as students' time spent may be influenced by confidence levels. The coefficient  $\beta_1$  captures the average difference in calibration between post-ChatGPT and pre-ChatGPT periods. We validate results by conducting extensive robustness checks, including placebo tests at false treatment dates. OLS regression in quasi-experimental educational research follows established conventions (Angrist & Pischke, 2009). Placebo or falsification tests at false treatment dates are a standard diagnostic for temporal confounding in difference-in-means designs (Bertrand et al., 2004; Callaway & Sant'Anna, 2021) and have been applied

in prior educational technology research to assess whether observed patterns are specific to the proposed treatment event. We specify a direct estimate as:

$$Calibration_i = \beta_0 + \beta_1 Treat_i + \varepsilon_i$$

While defining the full controls specification as:

$$Calibration_i = \beta_0 + \beta_1 Treat_i + \theta X_i + \varepsilon_i$$

where  $X_i$  is a matrix of controls detailed above.

## Results

### Descriptive Statistics and Balance Check

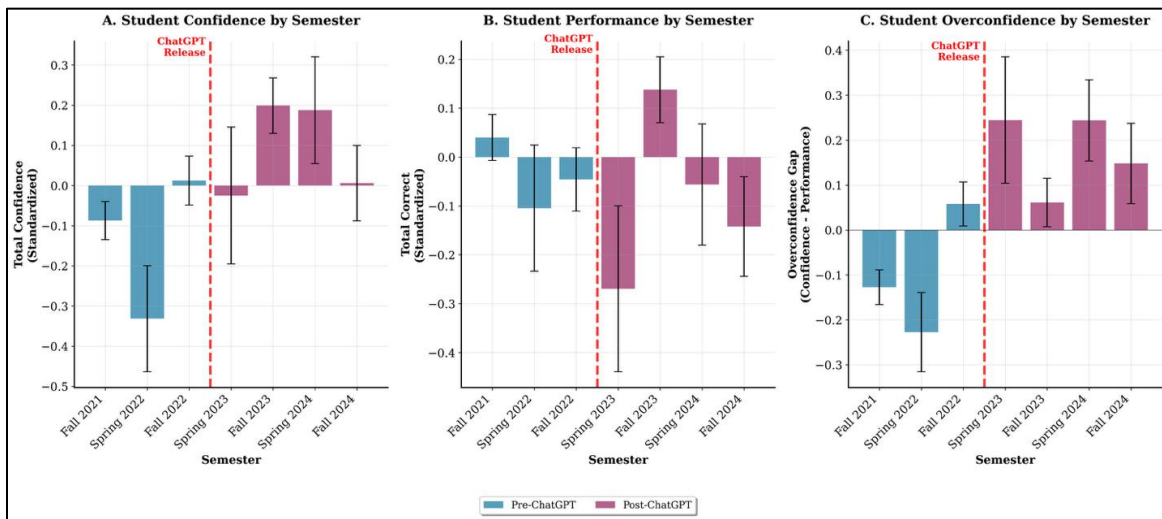
Descriptive statistics are shown in Table 1. All non-indicator metrics are standardized before analysis. Balance tests reveal a significant imbalance in class standing with respect to juniors. This imbalance suggests potential confounding requiring regression adjustment and provides a rationale for fitting a full set of control variables in the regression analysis.

**Table 1. Descriptive Statistics and Balance Check**

| Variable              | Full Mean | Full SD | Pre Mean | Pre SD | Post Mean | Post SD | Diff           | P-val |
|-----------------------|-----------|---------|----------|--------|-----------|---------|----------------|-------|
| <b>Outcomes</b>       |           |         |          |        |           |         |                |       |
| Total Confidence      | 15.86     | 4.06    | 15.56    | 4.01   | 16.40     | 4.12    | <b>0.84***</b> | 0.00  |
| Total Correct         | 3.16      | 1.82    | 3.16     | 1.82   | 3.16      | 1.84    | 0.01           | 0.96  |
| Overconfidence        | 12.71     | 3.13    | 12.41    | 3.07   | 13.24     | 3.17    | <b>0.83***</b> | 0.00  |
| <b>Student Chars.</b> |           |         |          |        |           |         |                |       |
| Excel Skills          | 2.26      | 0.94    | 2.28     | 0.91   | 2.23      | 0.98    | -0.05          | 0.42  |
| GPA                   | 3.48      | 0.43    | 3.49     | 0.42   | 3.46      | 0.45    | -0.03          | 0.28  |
| Female (%)            | 46.40%    | 49.90%  | 47.60%   | 50.00% | 44.20%    | 49.70%  | -3.4pp         | 0.27  |
| Credits Completed     | 16.17     | 25.78   | 16.06    | 25.21  | 16.37     | 26.82   | 0.30           | 0.85  |
| <b>Class Rank</b>     |           |         |          |        |           |         |                |       |
| Freshman (%)          | 74.20%    | 43.80%  | 74.20%   | 43.80% | 74.20%    | 43.80%  | -0.0pp         | 0.99  |
| Sophomore (%)         | 18.30%    | 38.70%  | 19.50%   | 39.70% | 16.10%    | 36.80%  | -3.4pp         | 0.16  |
| Junior (%)            | 5.50%     | 22.90%  | 4.10%    | 19.80% | 8.20%     | 27.50%  | <b>4.1pp**</b> | 0.00  |
| Senior (%)            | 1.90%     | 13.80%  | 2.20%    | 14.60% | 1.50%     | 12.10%  | -0.7pp         | 0.42  |
| <b>Ethnicity</b>      |           |         |          |        |           |         |                |       |
| White (%)             | 57.30%    | 49.50%  | 57.00%   | 49.50% | 58.90%    | 49.40%  | 0.9pp          | 0.76  |
| Asian (%)             | 18.10%    | 38.50%  | 17.60%   | 38.10% | 19.10%    | 39.40%  | 1.5pp          | 0.53  |
| Black (%)             | 5.20%     | 22.20%  | 5.30%    | 22.50% | 5.00%     | 21.70%  | -0.4pp         | 0.80  |
| Hispanic (%)          | 13.70%    | 34.40%  | 14.50%   | 35.20% | 12.40%    | 33.00%  | -2.1pp         | 0.34  |
| Multiracial (%)       | 4.30%     | 20.30%  | 4.80%    | 21.30% | 3.50%     | 18.30%  | -1.3pp         | 0.30  |
| Unknown (%)           | 1.10%     | 10.60%  | 0.80%    | 9.00%  | 1.70%     | 13.10%  | 0.9pp          | 0.16  |

Note. \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.10$ . All non-indicator metrics are standardized before analysis.

Figure 1 illustrates trends in student confidence and calibration by semester. After standardization, both confidence and calibration exhibit an inflection from negative to positive near the treatment date, indicating a clear increasing trend following ChatGPT's release. In contrast, the total correct metric (a summation of correct indicators) shows much less of a discernible pattern.



**Figure 1. Trends in standardized outcomes by semester**

## Regression Results

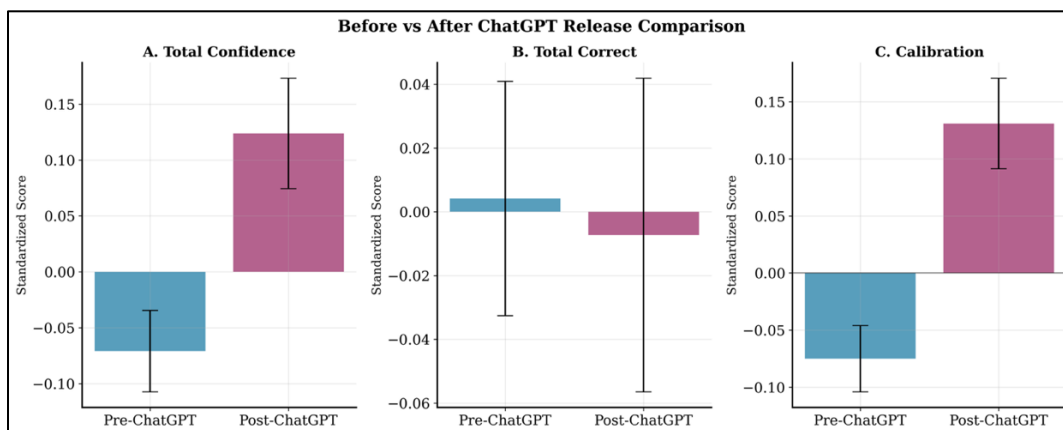
Regression results are reported in Table 2. Calibration rises substantially in the post-ChatGPT period, coinciding with increased confidence and performance stability. The simple model estimates a 0.21 standard deviation increase ( $SE = 0.049$ ,  $p < 0.001$ ), while the full controls model yields a more conservative 0.189 standard deviation increase ( $SE = 0.049$ ,  $p < 0.001$ ).

This difference represents a practically significant widening of the confidence–performance gap by approximately one additional confidence level on one question per student. Cohen’s  $d$  calculations reveal a small-to-medium effect ( $d = 0.27$ ) as shown in Figure 3. These effect sizes are educationally meaningful, representing approximately one-quarter standard deviation shifts in metacognitive calibration.

**Table 2. OLS Regression Results for Confidence, Correct, and Calibration Outcomes**

|              | Confidence           |                       | Correct             |                   | Calibration           |                       |
|--------------|----------------------|-----------------------|---------------------|-------------------|-----------------------|-----------------------|
|              | Direct               | Full                  | Direct              | Full              | Direct                | Full                  |
| Post-ChatGPT | 0.1944**<br>(0.0614) | 0.2064***<br>(0.0554) | -0.0114<br>(0.0614) | 0.0179<br>(0.059) | 0.2059***<br>(0.0491) | 0.1885***<br>(0.0485) |
| Obs          | 1157                 | 1136                  | 1157                | 1136              | 1157                  | 1136                  |
| R-squared    | 0.009                | 0.244                 | 0.000               | 0.133             | 0.015                 | 0.100                 |

*Note.* Robust standard errors in parentheses. \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ . Full specification includes controls for Excel skills, GPA, sex, ethnicity, and class rank.



**Figure 2. Mean standardized outcomes pre- and post-ChatGPT release**

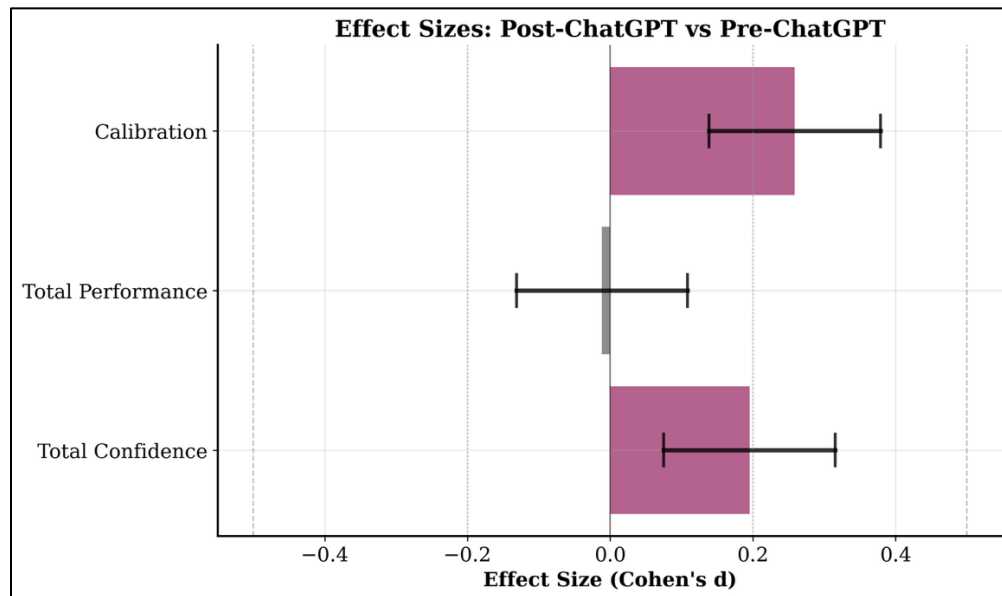


Figure 3. Effect sizes for treatment effects on study outcomes

### Heterogeneity Analysis

We examine whether effects vary across student subgroups by testing interactions between treatment and student characteristics using the following specification:

$$Overconfidence_i = \beta_0 + \beta_1 Treat_i + \beta_2 Subgroup_i + \beta_3 (Treat_i \times Subgroup_i) + \theta X_i + \varepsilon_i$$

where Subgroup is a characteristic indicator (for each: high Excel skills, female, high GPA, senior) and  $X_i$  is a matrix of controls including self-reported Excel skills, GPA, sex, ethnicity, and class rank that are not part of the Subgroup. As shown in Table 3, the aggregate pattern appears consistent across different types of students, suggesting a broadly applicable phenomenon rather than subgroup-specific effects.

Table 3. Heterogeneity Analysis: Treatment Effects by Student Subgroup

| Subgroup          | Treatment Effect | Interaction P-value |
|-------------------|------------------|---------------------|
| Low Excel Skills  | 0.208            | 0.616               |
| High Excel Skills | 0.160            |                     |
| Male Students     | 0.149            | 0.363               |
| Female Students   | 0.237            |                     |
| Low GPA           | 0.129            | 0.208               |
| High GPA          | 0.253            |                     |
| Non-Seniors       | 0.191            | 0.675               |
| Seniors           | 0.046            |                     |

*Note.* Treatment effects from full controls specification with interaction terms. P-values test the null hypothesis of no differential effect across subgroups.

### Discussion

Our results document a clear temporal pattern: during the period following ChatGPT's release, student confidence increased substantially while performance remained unchanged, coinciding with a widening of the overconfidence gap. These associations survive demographic adjustment and appear consistent across student subgroups.

The most significant methodological constraint of this study is the failed placebo test. As shown in Table A2, applying the same specification to a false treatment date six months prior yields a comparable coefficient, indicating that the post-ChatGPT pattern cannot be cleanly attributed to ChatGPT's release specifically. Importantly, however, failed placebo tests (Table A2) suggest that observed patterns may reflect broader temporal changes unrelated to GAI tools such as post-pandemic recovery effects, evolving student populations, or gradual shifts in educational culture. Under this interpretation, attributing effects specifically to ChatGPT would constitute spurious correlation. This finding underscores the methodological challenges inherent in studying the educational effects associated with technologies that achieve near-universal awareness simultaneously.

It is noteworthy that we find no heterogeneity among student demographics and covariates, as prior research has shown that while male and female respondents had comparably correct Excel survey answers, females exhibited lower self-reported confidence on average (Grealis et al., 2024). The results of this study suggest that GAI tool availability may be associated with a narrowing of the gender gap in confidence, as performance remained constant across the study period. If the post-ChatGPT era is characterized by a uniform confidence increase across genders, this could represent a complex and potentially positive social outcome of AI availability, one in which the psychological reassurance of accessible AI assistance disproportionately benefits groups that traditionally underreport confidence. However, this narrowing may also mask a deeper concern: if the gender gap in confidence closes not through genuine skill development but through uniformly inflated self-assessments, the underlying disparities in metacognitive calibration remain unaddressed. Future research should examine whether AI-driven confidence equalization reflects authentic empowerment or merely a more broadly distributed overconfidence.

According to Rruplli et al. (2024), current students have apprehensions regarding technical challenges when using GAI for learning, but still understand the downstream benefit of adoption. Rruplli et al. applied the Value-Based Adoption Model (VAM) to GAI adoption and found that perceived technicality had no significant negative effect on adoption intention, suggesting that the effort barrier to AI tool use has essentially vanished. When a technology is perceived as highly useful and requiring minimal effort, students may develop a sense of technological self-efficacy that is divorced from their actual independent ability. This VAM-based mechanism offers a compelling explanation for the confidence inflation documented in this study: students weigh the perceived benefits of AI access against negligible perceived costs, arriving at an inflated assessment of their own capabilities that implicitly incorporates the tool's capacity.

Viewed through the lens of distributed cognition (Hutchins, 1995), the observed confidence–performance decoupling may reflect students' implicit incorporation of AI capabilities into their self-assessed competence. If students perceive GAI as a persistent cognitive resource, one that can generate Excel formulas, troubleshoot errors, and explain procedures on demand, their functional confidence naturally expands to encompass the human–AI system rather than their independent skills alone. When subsequently asked to perform tasks without AI assistance, as in the Excel literacy survey deployed in this study, the gap between system-level confidence and individual-level competence becomes apparent. Furthermore, the concept of metacognitive laziness (Fan et al., 2025) suggests that habitual access to AI may suppress the self-regulatory processes of orientation, planning, monitoring, and evaluation that are essential for accurate calibration. Students who routinely offload cognitive effort onto AI tools may lose practice in the metacognitive skills needed to recognize their own knowledge gaps, producing the persistent overconfidence documented in our results.

These findings speak directly to ongoing debates in educational technology about the design of technology-enhanced learning environments. Prior survey research has documented that students simultaneously

perceive GenAI as beneficial for learning while expressing concern about its impact on their personal development and independent capabilities (Chan & Hu, 2023). If AI availability shapes student metacognition independently of actual tool use, then institutional policies focused solely on restricting or monitoring AI usage may be insufficient. Educators may need to develop pedagogical strategies that explicitly address the metacognitive consequences associated with living in an AI-saturated environment, helping students distinguish between AI-assisted capability and independent competence. A critical unresolved question is whether the documented confidence inflation reflects mere awareness of AI's availability, a psychological reassurance effect requiring no direct tool engagement, or active usage of AI tools that recalibrates students' functional self-assessments. Disentangling these mechanisms would require individual-level AI usage data, which the present study lacks and which future research with stronger instrumentation should prioritize.

## Limitations and Future Directions

This study faces several identification limitations. First, our data derives from a single institution, limiting generalizability across different higher education contexts. The sample consists predominantly of freshmen and sophomores enrolled in business and economics courses at a large R1 public university, which may differ systematically from community college students, STEM-focused institutions, or institutions in different national contexts. Second, the simultaneous global release of ChatGPT prevents meaningful control group comparisons, as awareness was effectively universal. Third, we lack direct measures of individual AI usage, preventing us from distinguishing between awareness effects and active usage effects. Fourth, our relatively short post-treatment observation window may not capture longer-term calibration dynamics. Finally, and most critically, the failed placebo test, in which a false treatment date six months prior yields a comparable coefficient, prevents attribution of the observed confidence–performance patterns specifically to ChatGPT's release, and represents the study's primary identification constraint.

Future research should employ stronger identification strategies such as multi-institutional datasets, experimental designs, event-study methodologies, or instrumental variable approaches. Collecting direct measures of AI use would enable clearer separation of exposure and adoption effects, while qualitative and longitudinal designs could illuminate underlying mechanisms. Furthermore, semester-level AI adoption estimates from prior research may serve as potential instruments for further investigation (Frydenberg et al., 2026).

## Conclusion

This study provides novel empirical evidence on a timely question in educational technology: how is AI tool availability associated with students' confidence–performance calibration? Analyzing 1,157 Excel literacy surveys collected between 2021 and 2024, we document substantial confidence increases without corresponding performance improvements, a meaningful overconfidence gap that coincides with, but cannot be causally attributed to, the emergence of ChatGPT. While identification challenges prevent definitive causal claims, the descriptive pattern itself warrants educational attention. Whether reflecting ChatGPT's specific impact or broader AI awareness, the confidence–performance decoupling documented here suggests students may be developing inflated self-assessments of their independent capabilities.

These findings carry important implications for the design of technology-enhanced learning environments. Instructors should provide frequent, realistic performance feedback, explicitly address metacognitive calibration, and help students distinguish between AI-assisted task completion and independent competence. Drawing on the distributed cognition framework, educators can implement process-based assessments that require students to explain the logic underlying their solutions, for example, articulating

why a particular Excel function is appropriate, rather than merely producing correct outputs. Such approaches force students to bridge the gap between system-level capability and internalized understanding, counteracting the metacognitive laziness that may accompany habitual AI use (Fan et al., 2025). Institutions should develop thoughtful AI policies that balance access with academic integrity while monitoring student metacognitive development, including the accuracy of student self-assessment and confidence.

As AI tools become increasingly sophisticated and widely available, educators must understand how these technologies influence not only student performance but also students' confidence and self-perceptions. This study contributes early empirical evidence on an understudied dimension of AI's educational impact: how the availability of AI tools shapes students' self-assessment and confidence, even when underlying skills remain unchanged. Further research with stronger identification strategies will be essential for guiding instructional design and institutional policy in the AI era.

## References

- Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, 21, 10. <https://doi.org/10.1186/s41239-024-00444-7>
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.
- Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119(1), 249–275. <https://doi.org/10.1162/003355304772839588>
- Budhathoki, T., Zirar, A., Tchouamou Njoya, E., & Timsina, A. (2024). ChatGPT adoption and anxiety: A cross-country analysis utilising the unified theory of acceptance and use of technology (UTAUT). *Studies in Higher Education*, 49(5), 831–846. <https://doi.org/10.1080/03075079.2024.2333937>
- Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2), 200–230. <https://doi.org/10.1016/j.jeconom.2020.12.001>
- Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20, 43. <https://doi.org/10.1186/s41239-023-00411-8>
- Church, K. S., Riley, J., & Schmidt, P. J. (2022). Has Excel become a “Golden Hammer”: The paradox of data analytics in SME clusters. *Journal of Emerging Technologies in Accounting*, 19(2), 211–234. <https://doi.org/10.2308/JETA-2020-051>
- de Bruin, A. B. H., Kok, E. M., Lobbestael, J., & de Grip, A. (2017). The impact of an online tool for monitoring and regulating learning at university: Overconfidence, learning strategy, and personality. *Metacognition and Learning*, 12(1), 21–43. <https://doi.org/10.1007/s11409-016-9159-5>

- Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, 22(4), 271–280. <https://doi.org/10.1016/j.learninstruc.2011.08.003>
- Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2), 489–530. <https://doi.org/10.1111/bjet.13544>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Frydenberg, M., Mentzer, K., & Patterson, A. (2026). The rapid rise of generative AI adoption among first-year college students. *Information Systems Education Journal*, 24(1), 4–18.
- Grealis, T., Harmon, O., Patterson, A., & Tomolonis, P. (2024). Excel literacy in the classroom. *Business Education Innovation Journal*, 16(2), 173–182.
- Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. <https://doi.org/10.4324/9780203887332>
- Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge. <https://doi.org/10.4324/9780203181522>
- Hutchins, E. (1995). *Cognition in the wild*. MIT Press. <https://doi.org/10.7551/mitpress/1881.001.0001>
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134. <https://doi.org/10.1037/0022-3514.77.6.1121>
- Mentzer, K., Frydenberg, M., & Patterson, A. (2024). Are tech savvy students tech literate? Digital and data literacy skills of first-year college students. *Information Systems Education Journal*, 22(3), 4–24. <https://doi.org/10.62273/MFSD3135>
- Miller, T. M., & Geraci, L. (2011). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(2), 502–506. <https://doi.org/10.1037/a0021802>
- Osterhage, J. L., Usher, E. L., Douin, T. A., & Bailey, W. M. (2019). Opportunities for self-evaluation increase student calibration in an introductory biology course. *CBE—Life Sciences Education*, 18(2), ar16. <https://doi.org/10.1187/cbe.18-10-0202>
- Patterson, A., Frydenberg, M., & Basma, L. (2024). Examining generative artificial intelligence adoption in academia: A UTAUT perspective. *Issues in Information Systems*, 25(3), 238–251. [https://iacis.org/iis/2024/3\\_iis\\_2024\\_238-251.pdf](https://iacis.org/iis/2024/3_iis_2024_238-251.pdf)
- Patterson, A., Mocarsky, M., Cho, J., Harmon, O., & Calvert, C. (2023). Teaching data literacy and sports economic fundamentals using fantasy sports. *Business Education Innovation Journal*, 15(1), 51–61.

- Pintrich, P. R., Wolters, C. A., & Baxter, G. P. (2000). Assessing metacognition and self-regulated learning. In G. Schraw & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 43–97). Buros Institute of Mental Measurements.
- Rruplli, E., Frydenberg, M., Patterson, A., & Mentzer, K. (2024). Examining factors of student AI adoption through the value-based adoption model. *Issues in Information Systems*, 25(3), 218–230. [https://doi.org/10.48009/3\\_iis\\_2024\\_117](https://doi.org/10.48009/3_iis_2024_117)
- Strzelecki, A. (2024). Students' acceptance of ChatGPT in higher education: An extended unified theory of acceptance and use of technology. *Innovative Higher Education*, 49(2), 223–245. <https://doi.org/10.1007/s10755-023-09686-1>
- Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT improves creative problem-solving performance in university students: An experimental study. *Computers & Education*, 215, 105031. <https://doi.org/10.1016/j.compedu.2024.105031>
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478. <https://doi.org/10.2307/30036540>
- Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12(1), 1–21. <https://doi.org/10.1057/s41599-025-04787-y>
- Xue, L., Rashid, A. M., & Ouyang, S. (2024). The unified theory of acceptance and use of technology (UTAUT) in higher education: A systematic review. *SAGE Open*, 14(1), 21582440241229570. <https://doi.org/10.1177/21582440241229570>
- Yang, T.-C., Hsu, Y.-C., & Wu, J.-Y. (2025). The effectiveness of ChatGPT in assisting high school students in programming learning: Evidence from a quasi-experimental research. *Interactive Learning Environments*, 33(6), 3726–3743. <https://doi.org/10.1080/10494820.2025.2450659>
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press. <https://doi.org/10.1016/B978-012109890-2/50031-7>

## Appendix

**Table A1. Qualtrics Survey Questions**

| Domain     | Question                                                                                                                                                          |
|------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Survey     | Who is the 5th player from the top when sorted alphabetically (from a to z) by first name? (Report full name)                                                     |
|            | For the provided data, what is the Home Runs Total (Sum)?                                                                                                         |
|            | For the provided data, what is the Home Run Average?                                                                                                              |
|            | For the provided data, what is the Home Runs Median?                                                                                                              |
|            | In the column labeled Dummy, use the IF function to assign a value of 1 to each row where Home Runs is > 18, and 0 otherwise. Report the sum of the Dummy column. |
|            | Using VLOOKUP, report the value of Previous Year Points for Evan Longoria from the Previous Year Stats sheet.                                                     |
|            | Report the average Fantasy Points by Dummy Variable (=1) from the pivot table.                                                                                    |
| Confidence | For all Survey questions: What is your degree of confidence in answering correct?                                                                                 |
| Controls   | On a scale of 1 to 5, with 5 as the highest, rate your level of proficiency in Excel.                                                                             |

How many classes in or requiring Excel have you had, including high school?

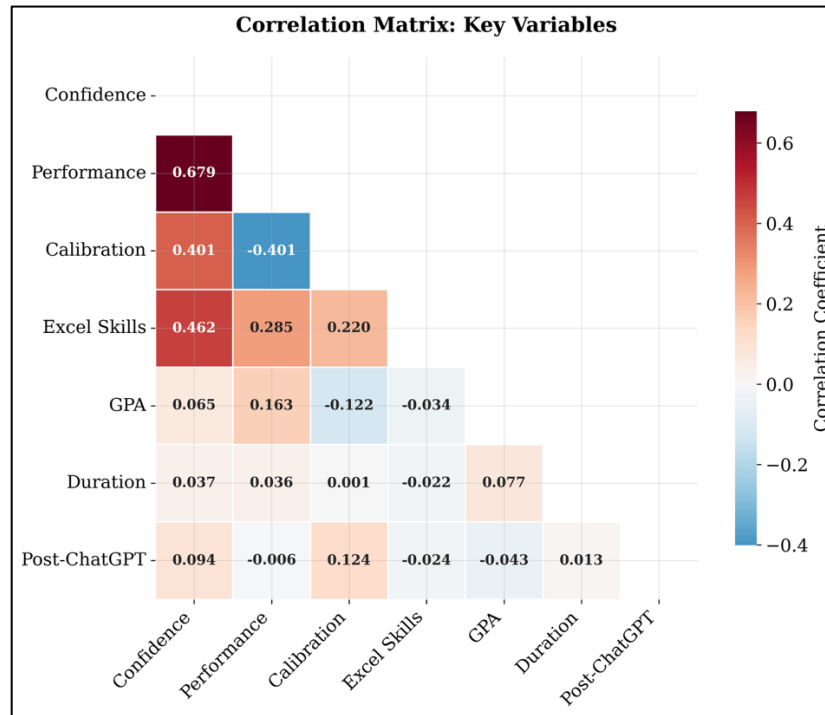


Figure A1. Correlation matrix of key study metrics

Table A2. Placebo Test at False Treatment Date

| Specification    | Baseline              | Placebo 6m Before     |
|------------------|-----------------------|-----------------------|
| Treatment Effect | 0.2064***<br>(0.0554) | 0.2374***<br>(0.0539) |
| Observations     | 1136                  | 1136                  |
| R-squared        | 0.244                 | 0.248                 |

*Note.* We tested whether similar confidence shifts appear at a false treatment date (May 2022) to assess whether effects are specific to ChatGPT's release. Results show non-negligible effects at a placebo date six months before treatment ( $R^2$  values comparable to the main specification), suggesting temporal patterns unrelated to ChatGPT availability may partially explain observed confidence increases. \*\*\*  $p < 0.01$ .