Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 22-564

Let's Chat: Leveraging Chatbot Outreach for Improved Course Performance

Katharine Meyer
The Brookings Institution

Lindsay C. Page
Brown University

Catherine Mata
Brown University

Eric Smith
University of Texas, Austin

B. Tyler Walsh
University of Central Florida

C. Lindsey Fifield
Georgia State University

Michelle Tyson
Georgia State University

Amy Eremionkhale
DePauw University

Michael Evans
Georgia State University

Shelby Frost
Georgia State University

Eye Eoun Jung
Georgia State University

This study provides pre-registered, experimental evidence on the use of non-generative artificial intelligence (AI) chatbots to support students in large-enrollment undergraduate courses. We find the chatbot messaging increased students' final grades and engagement with academic supports, such as tutoring. Treatment effects were generally consistent across student demographics, with the exception of treated women in a Microeconomics course, who earned final grades that were seven percentage points higher than women in the control group. This study provides evidence that integrating AI-enabled outreach and communication to students in their college courses can enhance student engagement, learning, and course performance.

VERSION: June 2026

Suggested citation: Meyer, Katharine, Lindsay C. Page, Catherine Mata, Eric N. Smith, B. Tyler Walsh, C. Lindsey Fifield, Michelle Tyson, Amy Eremionkhale, Michael Evans, Shelby Frost, and Eye Eoun Jung. (2026). Let's Chat: Leveraging Chatbot Outreach for Improved Course Performance. (EdWorkingPaper: 22-564). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/es6b-sm82

Let's Chat: Leveraging Chatbot Outreach for Improved Course Performance
Katharine Meyer, Lindsay C. Page, Catherine Mata, Eric N. Smith, B. Tyler Walsh, C. Lindsey Fifield, Michelle Tyson, Amy Eremionkhale, Michael Evans, Shelby Frost, Eye Eoun Jung
JEL No. I21, I23, O33

ABSTRACT

This study provides pre-registered, experimental evidence on the use of non-generative artificial intelligence (AI) chatbots to support students in large-enrollment undergraduate courses. We find the chatbot messaging increased students' final grades and engagement with academic supports, such as tutoring. Treatment effects were generally consistent across student demographics, with the exception of treated women in a Microeconomics course, who earned final grades that were seven percentage points higher than women in the control group. This study provides evidence that integrating AI-enabled outreach and communication to students in their college courses can enhance student engagement, learning, and course performance.

Katharine Meyer
The Brookings Institution
kmeyer@brookings.edu

Lindsay C. Page
Brown University
National Bureau of Economic Research
lindsay_page@brown.edu

Catherine Mata
Brown University
catherine_mata@brown.edu

Eric Smith
University of Texas, Austin
ensmith@utexas.edu

B. Tyler Walsh
University of Central Florida
brendan.walsh@ucf.edu

C. Lindsey Fifield
Georgia State University
cfifield@gsu.edu

Michelle Tyson
Georgia State University
mberg2@gsu.edu

Amy Eremionkhale
DePauw University
amyeremionkhale@depauw.edu

Michael Evans
Georgia State University
mevans33@gsu.edu

Shelby Frost
Georgia State University
sfrost@gsu.edu

Eye Eoun Jung
Georgia State University
ejung11@gsu.edu

I. INTRODUCTION

Despite documented benefits to college completion, more than a third of students who initially enroll in college do not ultimately earn a credential (Snyder & Dillow, 2015). Inequalities in college completion persist along socioeconomic and racial lines, even after accounting for academic achievement and preparation in high school (Belley & Lochner, 2007; Cohen et al., 2024; Holzer & Baum, 2017; Kena et al., 2014; Long & Mabel, 2012; Ma, Pender & Welch, 2019). Institutions have invested in numerous resource-intensive interventions to increase college persistence, including providing students with additional financial aid (Castleman & Long, 2016; Page et al., 2014), enhanced advising (Bettinger & Baker, 2014), or the combination of wraparound advising and financial assistance (Clotfelter, Hemelt, & Ladd, 2018; Scrivener et al., 2015; Weiss et al., 2019; Scuello & Strumos, 2024). While many comprehensive interventions significantly increase persistence and degree attainment, not all programs have positive effects, and some of the most promising interventions can be difficult to scale and sustain at resource constrained institutions (Sommo et al., 2023).

Increasingly, policymakers, educators, and researchers have focused on how informational barriers and administrative hassle factors have stymied students' pathways to graduation. Completing college requires students to navigate both institutional administrative tasks (e.g., applying for financial aid) and course-specific tasks (e.g., completing assignments). In postsecondary education, studies have shown that text-based outreach can be a low-cost, easy-to-implement, and effective strategy for supporting educational attainment by guiding students through complex administrative processes they encounter during college application and initial enrollment (Castleman & Page, 2015, 2017; Castleman et al., 2014; Linkow et al., 2021; Ortagus et al., 2020) and once enrolled (Castleman & Page, 2016; Page et al., 2025). Technological advances, including artificial intelligence (AI), open new avenues through which institutions can reach students and provide resources needed to persist through their degrees. While the most common AI application that comes to mind may be a generative tool (such as ChatGPT), other AI applications, including non-generative AI response and predictive algorithms, have been more widely applied in higher education operations (Demszky et al., 2023; Gandara et al., 2024; Page et al., 2025).

It is an open question whether institutions can leverage AI strategies to improve students' core academic experiences in college. More specifically, can targeted outreach affect academic

inputs, such as study time and assignment completion, in ways that translate into meaningful *outputs*, such as course performance and retention? In this paper, we report on an effort to implement and experimentally test a text-based chatbot with non-generative AI capability to provide course-specific, proactive outreach and support to students in large-enrollment undergraduate courses.[1] Since 2016, a research-practice partnership between Georgia State University (GSU), external researchers, and Mainstay, a technology company, has collaborated to design, build, and investigate the potential of artificially intelligent virtual communication tools (i.e., "chatbots") to support students to and through college.[2] Experimental studies have found that the chatbot communication improved first-year matriculation (Page & Gehlbach, 2017) and the completion of discrete administrative tasks necessary for college persistence, such as handling registration holds and refiling for financial aid (Page et al., 2025).

In this study, we apply the same chatbot technology within the classroom at GSU with the goal of increasing students' course engagement, performance, and completion. It was not ex ante clear that the same text-based reminders that effectively induce students to complete discrete administrative tasks would translate into the college course setting, where success requires completing numerous tasks (e.g., assignments) and engaging with course content consistently over the course of a semester. To implement and evaluate an academic chatbot, we drew on insights from GSU student experiences and prior literature to target courses in which the chatbot had the most potential for impact. We first identified courses with historically high "DFW" rates (DFW refers to a student earning a D, F, or withdrawing from a course). Next, we targeted large-enrollment courses and online courses where students had fewer opportunities to connect with the instructional team or with peers to form academic support systems. GSU also prioritized courses that were consequential to students' progression – courses that fulfilled graduation requirements or were gateway courses for popular academic majors. Finally, faculty buy-in was essential to testing the chatbot – while the tool did not require high engagement from faculty members, having faculty willing to experiment with student support initiatives was essential to a smooth implementation.

[1] We pre-registered the intervention and analysis with the Registry of Efficacy and Effectiveness Studies (REES) for each course under Registry ID 8160 (Government) and Registry ID 13760 (Microeconomics).
[2] For more information on Mainstay, see www.mainstay.com.

Given these target parameters, GSU identified two large enrollment, asynchronous online courses in the Political Science and Economics Departments to empirically evaluate the effectiveness of the academic chatbot via randomized controlled trials. GSU first implemented the academic chatbot in a section of "Introduction to American Government" (hereafter "Government")[3] and then replicated the experiment in multiple sections of "Principles of Microeconomics" (hereafter "Microeconomics") taught by two different professors.[4]

At the beginning of each intervention term, half of the students enrolled in the focal courses were randomly assigned to a treatment group, which received 2-3 scheduled, customized text messages each week. This outreach provided general information on weekly assignment due dates, targeted nudges to complete late/missing assignments, and encouragement and invitations to engage with the bot or the instructional team with any questions. Customization of the outreach included both *personalization* (e.g., "Hi FIRSTNAME") and *targeting* (e.g., messages differentiated for students who had a missing assignment versus students who were up to date on coursework). Students could text message the chatbot at any time of day and receive AI-generated responses drawing on a pre-programmed content knowledge base developed in collaboration with university administrators and the course instructional team. If the chatbot could not find a suitable response in its content knowledge base, those students' messages were flagged for the course TA to review and respond.

To preview our experimental results, across the two courses, the academic chatbot significantly shifted students' final grades, increasing the likelihood that students earned an A or B by four percentage points. Results were similar across the two courses. In subgroup analyses, we find generally similar treatment effects across student demographics, with one notable exception. In Microeconomics, women assigned to treatment earned final grades that were seven points higher than women in the control group; women were also 11 percentage points more likely to earn a final grade of an A or B and 10 percentage points less likely to DFW. There were no treatment effects for men. In Microeconomics, men and women in the control group

[3] Nearly all students at GSU take this course to satisfy a state of Georgia graduation requirement. GSU students are only exempt from taking "Introduction to American Government" through examination (e.g., Advanced Placement exam scores).

[4] At GSU, Microeconomics is a required course for economics and business majors and is one option for students to satisfy a core curriculum social science foundations requirement. Unlike Government, which all students must pass or test out of, the Microeconomics course is one of 15 possible courses students can select to satisfy the social science foundations requirement.

performed similarly, thus the treatment effect for women resulted in treated women significantly outperforming both treatment and control men.

We consider the mechanisms through which the chatbot may have increased final grades. We find that treated students were more likely to attend university tutoring, which may have contributed to higher course performance. We also find suggestive evidence that students assigned to treatment were more likely to complete their homework in Microeconomics (a course that involved multiple weekly assignments), though we find no evidence of differential assignment completion or performance in Government. We additionally test whether the intervention affected academic performance outside of the focal course, examining treatment effects in (a) other courses taken during the intervention semester and (b) course enrollment and performance patterns the semester following the intervention term. We do not find evidence of within-term spillover effects or effects on longer-term outcomes, suggesting that while the intervention improves students' performance on course-specific outcomes, it may not develop underlying habits or study skills to a sufficient extent that subsequent academic endeavors are affected. Ultimately, on end-of-course surveys, students reported enthusiasm for the chatbot, with 82 percent of respondents recommending its continued use in the course and expansion to other courses at GSU.

Our study makes three contributions. First, we build on a growing body of evidence on the positive effects that virtual outreach and communication can have on students' completion of essential tasks needed to successfully progress through college. Prior work has hypothesized that trusted senders and customization are key ingredients of impactful nudge outreach. Our work extends this foundation, contributing a novel case of messaging coming from a course instructional team as the trusted sender and providing more frequent customized updates on course performance, a domain where task completion is an ongoing, dynamic process (in contrast with prior instances of discrete tasks such as FAFSA filing). While there have been promising studies of virtual outreach, not all applications have yielded significant effects. In one related study, Oreopoulos and Petronijevic (2019) found limited effects of a suite of low-touch psychological, peer coaching, and nudge interventions on college students' term grades. While they found improvements to student mental health and increases in reported study time (academic input), these proximal effects did not translate into higher course grades or credits earned (academic output). Pugatch and Wilson (2024) also show that email outreach (ostensibly

sent from a course instructor) can effectively increase students' academic inputs (e.g., completing additional practice problems), though these efforts also did not translate to increases in course grades / academic outputs. In contrast, Carrell and Kurlaender (2023) found that targeted email messages with assignment reminders, encouragement to attend office hours, and notes about current course performance sent directly from a student's professor led underrepresented minority students to earn higher grades in the course and ultimately to graduate from college at higher rates. In other studies, adding current course performance information to student communications from faculty – which provided students with regular reminders about their course standing – increased subsequent homework performance (Smith et al., 2018) and nudging students about course section attendance boosted attendance and subsequent course performance (Cortinhas, 2025).

These varied effects of proactive outreach on course performance suggest that message design features – such as the sender or customization to individual student circumstances – are likely to be important factors in the effectiveness of virtual outreach to meaningfully drive intermediary inputs and their subsequent academic outputs. In this analysis, we study two novel features of virtual outreach. First, in contrast with prior studies of one-way texting, the AI-enabled chatbot responses ensured students received answers to their questions at all hours. Second, embedding a TA in both the course and chatbot administration ensured timely gradebook information that resulted in messages customized to students' performance and engagement in a particular week, and also presented a trusted persona to the students assisting the chatbot and likely engendering more trust in the tool. This paper highlights how a collaborative partnership with faculty in the development and implementation of the chatbot intervention may be a key component to ensuring messaging is well targeted to students' most acute needs and can successfully impact academic achievement.

Second, we highlight the barriers students face in navigating administrative tasks within college classrooms and the role of chatbots to improve students' academic outcomes. Extant research shows that course structures affect student performance, with students typically earning lower grades in larger courses (De Giorgi, Pellizzari, & Woolston, 2012) and in courses taught online (Bettinger, Doss, et al., 2017; Bird, Castleman, & Lohner, 2022; Goodman, 2025; Kofoed et al., 2021). Online course taking has increased substantially since the pandemic, with 54% of undergraduates taking some or all of their courses online in 2022-23, compared to 36% of

undergraduates in 2019-20 (Goulas, 2024). Given the rising prevalence of large and online course structures in higher education, despite the documented negative effects of those modalities on student outcomes, our work offers important insights into how personalized text message outreach can help students navigate structural barriers to college success.

Finally, we advance an understanding of how to incorporate non-generative AI response technologies into educational settings and how the collaboration between student success expertise and technological support to scale responses can offer students more proactive support. The state of AI in education is evolving rapidly, moving from initial skepticism and concern to broader acceptance and innovative applications (Bick et al., 2024). Initially, the introduction of technologies like ChatGPT in 2022 raised alarms in the educational sector. K-12 schools and universities responded initially with bans and restrictions, fearing these tools would enable cheating and diminish critical thinking skills among students. However, the narrative is shifting as educators and institutions recognize the potential of AI to revolutionize teaching and learning methods (e.g., Wang et al, 2024). Increasingly, research has explored how AI-enabled communication can provide students with timely information about academic performance and opportunities. One study found that students in a college section leveraging AI-based chatbots performed better on assessments and reported higher levels of motivation and self-efficacy (Lee, Hwang, & Chen, 2022). Another study designed a chatbot for computer science majors to address their self-reported challenges around time management and found that participants ranked reminders and connections with tutoring resources as the most useful features (Tian et al., 2021). Other chatbots focused on developing academic skills have also increased students' academic performance with high levels of student enjoyment of the tool (Guo et al., 2023). The academic chatbot implemented at GSU provides an opportunity to examine further how AI enabled outreach can lower the cost of implementing text-based outreach and help professors send targeted messages to students at scale and how the effectiveness of that tool may vary in different fields of study and types of courses.

While concerns rightly remain about potential misuses, the time is right for research and development to investigate both the potential and the limitations of AI as a tool for enhancing classroom interactivity, personalizing learning experiences and supports, and aiding students in navigating the administratively complex bureaucracies of educational institutions (Jurenka et al., 2024). Further, beyond the evolution of AI in education, students' relationships with technology

have evolved, and testing a mobile-based communication strategy informs how universities can best meet students on their preferred communication platforms. While interventions have long leveraged text messaging for administrative tasks, there has been hesitancy to use a modality perceived as more "informal" for coursework communication. However, research indicates that today's students increasingly rely on mobile devices for schoolwork; Canvas, a common learning management system, reported that in 2024 39% of student assignments were uploaded via a mobile device (Wells, 2024). How students view AI as a complement to existing educational supports is an open area of study. This study contributes valuable insights to the application of non-generative AI to support students' course management skills.

II. INTERVENTION CONTEXT AND DESIGN

Institutional Context

Georgia State University (GSU) is a public, research university in Atlanta, GA that enrolls more than 52,000 undergraduate students. GSU is a minority serving institution; 63% of students identify as Black, Hispanic, or of two or more races. About 53% of GSU students receive Pell grants. GSU has a pooled college completion rate nearly identical to the national four-year institution average, with about 56% of students earning a bachelor's degree within eight years of initial enrollment (College Scorecard, n.d.). Given positive effects from earlier experimental interventions leveraging chatbot communications to reduce summer melt (Page & Gehlbach, 2017) and improve first-year retention (Page et al., 2025), GSU made chatbot communication regarding pre-matriculation and other required administrative tasks standard practice with all students who have opted into receiving text-based outreach. Currently, about 86% of incoming students opt into text-based communication from the university.

Focal Course Contexts

Both the Government and Microeconomics courses were online, asynchronous courses taught by full-time GSU faculty. Both focal courses have long been offered online by the instructors implementing the intervention (i.e., were not shifted online due to the COVID-19 pandemic). Students' course grades in Government reflected performance on reading quizzes, completion of a visit to a local museum (or alternative assignment), and completion of and performance on 3-4 multiple choice exams (taken asynchronously either at the campus testing center or virtually with a digital proctor). In Microeconomics, students' final grades reflected

participation in discussion boards, completion of and performance on practice and evaluative problem sets, and performance on evaluative quizzes.

Business as usual

Business-as-usual communication in the focal courses was delivered primarily via email. In Government, standard communication from the instructor already included regular, targeted email reminders to students. These emails included reminders about upcoming due dates, encouragement about recent performance, and suggestions for students to meet with the professor when they had failed to turn in an assignment or complete an exam. Some messages went to nearly all students, while others were targeted to a smaller group with specific course flags (e.g., the professor would email about 4-5% of students after they failed to submit exam 1 before the deadline). In Microeconomics, instructors also sent email reminders about upcoming assignments and targeting outreach to students who had not logged into the course website in five days. Those instructors also encouraged students to engage with them via virtual office hours as part of their standard communications.

All students in the analytic sample (treatment and control) had opted in to receive regular text-based communication from GSU's university-wide retention chatbot. This other chatbot sends messages to students about upcoming administrative tasks (e.g., when next semester's registration opens), targeted, data-informed messages about their enrollment and accounts, and messages proactively asking students if they need support or if the bot can connect them with a GSU resource. Prior research found positive effects of the retention chatbot on student completion of important college persistence tasks (Page et al., 2025).

Intervention Description

All students enrolled in the focal courses received standard communications from the course instructional team, as described above. Treatment students were sent 2-3 scheduled text messages each week (about 40 messages throughout the semester).[5] These messages were designed to: (1) provide timely reminders about course requirements; (2) provide customized feedback on each student's individual progress; (3) connect students to course-relevant academic supports; and (4) serve as an additional channel of communication between students and their course instructors. The chatbot messages fell into three broad categories: weekly updates, encouragement messages, and reminder messages. Weekly updates sent every Monday

[5] See online Appendix B for a representative set of messages from the fall 2021 implementation in Government.

previewed course tasks and responsibilities for that week and were customized by whether students had completed the previous week's assignments. Encouragement messages were signed by the course teaching assistant (TA) and were crafted to promote a growth mindset and to invite students to provide feedback on how their semester was going. Reminder messages were sent to students as needed (e.g., outreach to students who had not completed an online exam by a given time).

Students could text back or initiate a conversation with the chatbot at any time. When students texted in with a question, the chatbot used AI to respond with the closest match response in the system's knowledge base. When the system flagged a chatbot response with a low probability of response match, the question was then directed to and answered by the TA to provide personalized follow-up, as needed. These manual responses from the TA were then used to update the system knowledge base. TAs still reviewed each bot response to student queries on a daily basis (including responses with a high probability of response match) for quality control purposes. Figure 1 outlines this review process of student queries and how the chatbot draws on a content knowledge base to craft immediate responses while the TA provides general oversight; throughout the semester, the content knowledge base is updated based on student questions and accuracy of bot replies to ensure future questions on that topic are more accurately answered. [6,7]

One additional novel feature of the Government course chatbot was a function called #quizme through which students could request a quiz on the course material covered in an upcoming exam. Through #quizme, students could receive and answer a set of multiple-choice questions. For each one, the bot would indicate whether the student answered correctly and/or direct the student to where in the textbook they could read more about the topic and find the correct answer. The bot promoted #quizme in several weekly digests and additional promotional messages. Students could activate #quizme during the two weeks prior to each course exam due date.[8] Since the Microeconomics course did not have exams, GSU did not develop and deploy a #quizme tool for that course.

---

[6] The course TA closely monitored the message interface the two hours following a scheduled campaign, given that most students that replied to the academic chatbot did so shortly after receiving a scheduled message. The TA also checked for flagged messages at least once (and often 2-3 times) each day.

[7] For more information on the technical details of this system, please see: https://patents.google.com/patent/US20180131645A1/en.

[8] The course TA updated the bank of #quizme questions throughout the semester to reflect the chapters covered on the next exam.

In spring 2021, we launched a pilot study of the course chatbot in Government and distributed messages to all enrolled students who consented to text message communication from the university.[9] The pilot study enabled us to refine message scripts, receive qualitative feedback from students about the chatbot experience, and examine engagement patterns. Students enthusiastically recommended the bot – 90% of end-of-course survey respondents recommended that GSU continue the bot for Government. The spring pilot also enabled us to develop a more robust bank of academic chatbot responses and better train the bot to the course structure and context.[10] We also adapted messages based on student feedback and engagement patterns. Most notably, we transitioned to sending students weekly digests customized to their course performance to date (e.g., noting whether they had assignments missing or congratulating those who worked ahead) rather than generic notices of upcoming due dates.[11]

Randomization Design

Each semester of the RCT, we randomized all students enrolled in the focal courses who had consented to receive text messages from GSU to either the academic chatbot treatment condition or to the control group (students who had not consented to receive text messages are not included in the study). We separately randomized students enrolled as of the first day of class and a second roster of students who enrolled during the semester add/drop period; we account for randomization wave blocks in our analysis (see analytic model below). We do not remove students from our analysis who dropped the course during the add/drop period since dropping the course occurred after treatment began. Instead, we code all outcomes as zero for students who dropped. Instructors were not notified which students were assigned to the chatbot; while we cannot rule out that instructors may have learned about specific students' receipt of chatbot messages from conversations with students or the TA supporting the chatbot, there was no systematic communication about students' treatment assignment. Given the online, asynchronous nature of the class with no collaborative assignments, we hypothesize that there were few if any opportunities for spillover of treatment information from students assigned to treatment to

[9] Of the 828 students enrolled in the course during spring 2021, 705 had previously opted in for texting from the university and received the pilot messages.

[10] For example, during the pilot the bot would often interpret a question about a course due date as a question about a GSU administrative due date (such as FAFSA or registration). The summer following the pilot, the course TA worked to substantially expand the knowledge base to ensure course-specific questions received a course-specific answer.

[11] Smaller edits included reducing the frequency of emojis and formalizing policies around message length.

students assigned to control, though we cannot observe this directly. There was no observed crossover (no control students attempted to gain access to chatbot messages).

Analytic Sample

Across the fall 2021, spring 2022, and fall 2022 academic semesters we randomized a total of 1,568 students enrolled in Government, and during the 2022-23 academic year, we randomized 915 students enrolled in Microeconomics.[12] In Table 1 we report balance on student characteristics between treatment and control students (pooled across courses in panel A, Government in panel B, Microeconomics in panel C). We observed no significant differences in characteristics between students in the treatment and control conditions.

Student demographics were similar across the two courses. Just over half of the Government sample were women, 45% were Black, 22% were white, 25% were first-generation college students, 57% were eligible for the Pell grant, their average high school grade point average (GPA) was around 3.5, and almost 9% had previously attempted the course and were re-taking it. [13] Overall, about 63% of students were freshmen, though the grade-level composition of the course varied considerably across intervention terms, from 43% in fall 2021 to 78% in fall 2022.[14] In Microeconomics, about half of the students were women, 52% were Black, 18% were white, 21% were first-generation students, 56% were eligible for the Pell grant, and their average high school GPA was about 3.4. Unlike Government, the Microeconomics courses had a higher share of students re-taking the course – 15%. Historic DFW rates were slightly higher in Microeconomics, resulting in a larger pool of students who might retake the course for a higher grade. While overall about 30% of the Microeconomics analytic sample were freshmen, this varied between 12% of the sample in the fall and 47% in the spring semester.[15] Microeconomics is more likely than Government to be taken by students who are not in their first semester of college. The instructional team indicated that this is because the course has a math pre-requisite, which many students complete in their first semester.

[12] We estimated a minimum detectable effect size of approximately 0.157 for our main outcomes of interest (included in pre-registration).

[13] The share of students missing high school GPA values ranged from 7% of students in the fall 2022 semester to 16% in the spring 2022 semester. Most transfer students (~70%) are missing high school GPA. We present high school GPA summary statistics for students with a valid high school GPA value in summary tables. In our impact analyses including covariates we use dummy imputation and code missing high school GPAs as zero and include an indicator for missing GPA in our models.

[14] See Appendix Table 1 for by-term summary statistics for Government.

[15] See Appendix Table 2 for by-term summary statistics for Microeconomics.

III. ANALYTIC STRATEGY

Our primary analytic goal is to estimate the effect of being assigned to receive academic chatbot messaging on course performance, course engagement, and student sense of institutional support. To estimate these effects, we use regression models of the following general form:

$$Y_{irzc} = \beta T_i + \boldsymbol{X}\gamma + \rho_{rzc} + \varepsilon_{irzc}$$

Where Y_{irzc} represents the outcome for study participant *i* randomized in round *r* and enrolled in term *z* in course *c*, T_i is the indicator for treatment assignment and is equal to one if the study participant *i* is randomized to the chatbot group and zero otherwise. $\boldsymbol{X}$ represents a vector of baseline characteristics for individual *i* (included primarily to explain residual variation in outcomes and thus to improve precision of estimation), and ε_{irzc} is an error term. We include fixed effects to account for randomization blocks (ρ_{rcz}), which were nested within course and instructor (c) and within academic term (z).[16,17] In both courses we randomized the roster of students enrolled on the first day of course and then ran a second round of randomization among students who enrolled during the add/drop period. In Microeconomics, we additionally blocked randomization by course section since there were multiple instructors. We estimate effects for the course samples overall and for selected subsamples to examine differences by student characteristics and to test for equality of treatment coefficients across subgroups of students.[18]

We focus on the intent-to-treat (ITT) effects of being assigned to treatment. The ITT effect is the most policy-relevant outcome to universities which might use this tool in the future, as it estimates the effect of being initially opted-in to receive chatbot messages. We did not run treatment-on-treated (TOT) analyses, since there is no clear indicator for treatment receipt. The only observable variation in treatment received by students was driven by student opt-out behavior – students who dropped the course or opted out from messaging did not receive the full set of messages, but otherwise students assigned to treatment were slated to receive all messages. We treat dropping the course at any time as an outcome of interest (coding students who drop as

[16] Across the two terms and four intervention semesters, there were 26 unique randomization blocks. In American Government the only blocking within term was between randomization rounds, and blocks averaged about 350 students. Within each Microeconomics term, students were blocked by instructor and randomization rounds, and blocks averaged about 70 students.

[17] We also calculate randomization inference p-values using the Stata ritest (Heß, 2017) command to account for the blocked nature of treatment assignment and comparatively small sample size (including some small blocks) and report the results in Appendix Table A3. We find consistent results.

[18] When comparing heterogeneous treatment effects, we test for equality of coefficients across different subsamples.

having a zero for other outcomes – for example, earning a 0 when calculating final numeric grade). Since chatbot opt-out rates were low, we do not run separate analyses to estimate effects for students to whom the full set of messages was sent.

Data and Measures

Most outcomes come from deidentified course gradebooks, course learning management system records, and GSU administrative records, provided directly to the research team for analysis. We also studied the effect of chatbot communication on students' class experiences and perceptions of the instructor via an end-of-course survey (online Appendix C reports the specific attitudinal questions we asked students).[19] We also included a set of survey items to ask treatment participants about their experience with the course chatbot, including the extent to which they found the communication helpful, whether they read the text messages, whether they knew about and used the #quizme function (where applicable), and whether they would recommend future use of the chatbot at GSU. As we detail below, two limitations to our survey analysis are low response rates and differential survey participation by student characteristics. Other measures of engagement come from the Mainstay message logs, which we code to identify whether and how frequently students messaged the platform as well as characteristics of their messages (e.g., opt-outs vs. questions). Our pre-registered, confirmatory outcome measures were students' final numeric course grade, a binary measure of passing the course (e.g., the "DFW" rate), submitting assignments and completing readings on time, and sense of academic support from the institution (measured on surveys). GSU was unable to provide data on "on time" assignment completion or readings and instead we examine ever completing assignments. Figure 2 outlines the intervention theory of change, including details on the core intervention features (weekly assignment reminders, connections with instructional team, and recommendations of additional academic supports) and how we hypothesized they would impact student behaviors, proximal/within-course outcomes, final course outcomes, and distal college persistence outcomes. In Figure 2 we visually distinguish between our original confirmatory and exploratory outcomes.

IV. RESULTS

Main effects: Final course performance

[19] We were not able to field an end-of-course survey in Microeconomics.

We first examine the ITT effect of the chatbot on students' course performance. In Table 2, we report the effects of being assigned to receive chatbot communications on students' numeric final course grade and attainment of performance benchmarks (earning an A, earning a B or higher, etc.) and on course completion (whether students withdrew or dropped the course). Earning an A, B, or C ensures students receive college credit for the course, is required for maintaining the Georgia HOPE scholarship, and enables degree progress. Other outcomes have different impacts on students' transcripts and degree progress. Students can *drop* a course during the formal add/drop period at the start of the semester (approximately the first two weeks) without penalty. Dropping a course is therefore not an inherently negative outcome. However, since the focal courses in this study fulfill various graduation requirements, limiting drops was an outcome of interest. Students can *withdraw* from a course between the add/drop period and a mid-semester withdrawal deadline. While withdrawals do not affect students' grade point averages, GSU students have a limited number of withdrawals they can take throughout their college careers (after exceeding that limit, students will receive an A-F grade for all subsequent courses), and future employers or graduate school admissions officers may look unfavorably on excessive withdrawals on a student's transcript. Therefore, reducing withdrawals was a goal of the intervention. Finally, for some students a withdrawal may be preferable to earning a course grade of D or F, which will negatively impact their GPA and degree progress. Therefore, we separately examine the share of students earning a D or F as well as pooled with withdrawals to estimate the impact of the intervention on overall "DFW" rates.

Across both courses, the largest effects of chatbot outreach are on students earning an A or a B in the course. Panel A shows the pooled estimates. Treatment students are four percentage points more likely to earn an A relative to 36% of control-group students (about an 11% increase) and arere four percentage points more likely to earn an A or B relative to 61% of control-group students (about a 7% increase). Point estimates for these outcomes are nearly identical though less precisely estimated for each course. We also run multiple comparison corrections for this suite of final grade outcomes, reporting the Westfall & Young (1993) adjusted p-value for each outcome. In the full sample, based on the adjusted p-values the treatment does not lead to significant increases in earning an A but does significantly increase earning a B or higher.

Overall, treated students are not significantly less likely to DFW or to drop the course relative to control students. However, panel A masks variation in course completion by subject. In Government (panel B), students assigned to treatment are three percentage points less likely to DFW compared to 18% of control students. This effect on the DFW outcome is primarily driven by a reduction in earning a D or F (where we observe a two percentage point reduction, not statistically significant). In Microeconomics (panel C) students assigned to treatment are three percentage points less likely to *drop* the course relative to 8% of control students. As noted above, dropping the course must be done earlier in the semester. These differences in effects may be due to differences in course structure and assignment timing. Microeconomics had more assignments due in the early weeks of the semester than Government did, and the intervention may have helped ease the transition into this heavy Microeconomics workload for treated students.

Main effects: Performance conditional on completion

One mechanism through which the chatbot could have improved course grades is by increasing the likelihood students complete the course and earn a non-zero grade. In Table 3, we compare our main ITT effects for the pooled sample, which are unconditional impacts of treatment assignment, to estimates based on two different conditional samples: the sample of students who did not drop the course during the early add/drop period and the sample of students who completed the course (i.e., did not drop or withdraw). The treatment effect on earning a B or higher remains nearly identical in magnitude across samples – students assigned to treatment are about four percentage points more likely to earn an A or B in the course.[20]

Heterogeneous treatment effects: Final course performance

We then examined *for whom* the course chatbot improves final grade outcomes, testing whether treatment effects differ according to certain student characteristics. To illustrate the variance in treatment response, in Figure 3 we plot the treatment coefficients from separate regressions estimating the effect of being assigned to treatment on final numeric grade and on earning a B or higher for different subgroups of students (pooled across courses, including other

[20] Results are also generally consistent if we look at each course for the conditional sample. Looking only at students who completed the course (i.e., did not drop or withdraw), treated students in Government were five percentage points more likely to earn an A or B (statistically significant at $p<0.05$). We observe a smaller, not statistically three percentage points difference in earning an A or B in Microeconomics.

student covariates as controls).[21] We also graph the full sample results at the top, visualizing the effects reported in Table 2. We find significant treatment effects on the likelihood students earn a B or higher across multiple subgroups.[22] While the full sample treatment effect on final numeric grade is not statistically significant, we observe statistically significant increases for upperclassmen, first-time course takers, and students who were not Pell eligible. Generally, the directional pattern of the numeric course grade effects are similar to the pattern of effects for earning a B or higher. For example, we observe null effects for Hispanic students on both outcomes, and on both outcomes, the differences are larger in magnitude for first-generation students compared to continuing generation students). While the differences are statistically significant for some subgroups but not others, we generally fail to reject the null hypothesis of differential impact (finding, for example, that the treatment effects for Black students and for white students do not differ significantly). We run formal tests of equality across the full set of outcomes reported in Table 2 for each of the subgroups graphed in Figure 3 and find no evidence of substantially different treatment responses by student characteristics.[23]

However, pooling across courses masks heterogenous treatment effects by student demographics that are course specific. In Figures 4a and 4b we replicate the structure of Figure 3 separately by subject. In both subjects we again see instances in which a subgroup has a statistically significant treatment response while their counterpart does not; as with Figure 3, many of the heterogeneous treatment effects reported in Figures 4a and 4b are not significantly different from other subgroup effects.[24] The notable exception is the treatment effect for women in Microeconomics.

[21] For example, the "first generation" row reports the treatment coefficient and confidence interval for a regression limited to only first-generation students who were assigned to treatment or control in either Government or Microeconomics, including all other covariates (*inter alia*, race, sex, prior academic performance, Pell receipt).
[22] Results are similar if we look at earning an A as an outcome instead.
[23] We run formal tests of equality on each of the main final grade outcomes reported in Table 2 comparing treatment effects for subgroups reported in Figure 1: first-generation vs. continuing generation students, Pell eligible vs. Pell ineligible students, women vs. men, course retakers vs. first-time takers, Black vs. white students, and freshmen vs. upperclassmen. We find across those 42 tests of equality, five are statistically significant: for three outcomes in the comparison between Pell eligible and Pell ineligible and for two outcomes in the comparison between course retakers and first-time takers. Full results available upon request.
[24] In Government we find four instances of non-equal treatment effects across 42 tests of equality – two instances for Pell-eligible vs. Pell-ineligible students and two instances for course retakers vs. first-time course takers. In Microeconomics, except for women vs. men, we find only one instance of non-equal treatment effects across the other 35 tests. Full set of tests available upon request.

In Table 4 we report the treatment effect in Microeconomics for each of the final grade benchmarks for women and men with the formal test of equality across the two subgroup regressions. We consistently see large effects for women and no effects for men, with many of these subgroup effects statistically different from each other. In Microeconomics, women assigned to treatment earn final grades that are seven points higher than control group women. They are 11 percentage points more likely to earn a B or higher, 10 percentage points more likely to earn a C or higher, and six percentage points less likely to drop the course than women in the control group. Results are consistent even after multiple comparison corrections, as reported in Table 4. Control group women in this course were not performing substantially lower than men (averaging final grades of 68.9 and 70.5 respectively). The treatment effect more than closes this modest gap and results in treatment-assigned women substantially outperforming both control and treatment-assigned men.[25]

Quantile regression estimates with the sample pooled across courses capture the relationship between treatment and final grades along the distribution of final grades (Koenker & Bassett, 1978). It may be that treatment is differentially impactful at different levels of course performance. In Figure 5, we plot the treatment effect for each quantile from the 10th through the 95th quantile, with the main OLS effect from Table 2 plotted at the top of the graph. We find treatment assignment results in a statistically significant increase in final course grades between the 35th and 75th quantiles of the course distribution (as well as at the 95th quantile). This indicates that treatment effects are non-uniform along the distribution of course performance, and that treatment is most impactful for students in the middle of the academic distribution (and those at the very top of the distribution).

Mechanisms

We next consider potential mechanisms through which the course chatbot increases final grades. We examine whether student behaviors and proximal, within-course outcomes vary by treatment assignment. We also conduct mediation analyses to assess whether, and through which

[25] We ran a similar analysis as reported in Table 3 just for women in Microeconomics, conditioning the sample on only women who completed Microeconomics. On this selected sample, we continue to see a statistically significant seven percentage point increase in earning an A or B (relative to 11 percentage points in the unconditional sample of women in Microeconomics). Despite observing that women were less likely to drop Microeconomics in the full sample (a six percentage point decline reported in Table 4), among women who complete the course, those in the treatment group still outperformed those in the control group, pointing to the chatbot affecting persistence as well as performance.

mechanisms, these intermediary outcomes may have contributed to final course performance. We generally find small and marginally significant effects on proximal course outcomes. While suggestive, we caution over-interpretation and highlight that additional work is needed to fully understand the mechanisms through which chatbots may affect final grades.

We first examine whether the treatment affects student performance on specific course deliverables. There are no treatment differences in assignment completion or performance in Government, as reported in Table 5.[26] In Microeconomics, however, we do find suggestive evidence of treatment effects on assignment completion and performance. As reported in Table 6, treatment-assigned students are five percentage points more likely to complete the available practice quizzes and thus, earn practice quiz grades 3.59 points higher; these quizzes were worth 15% of their final grade). The Microeconomic chatbot outreach does not affect course participation or reading checks (cumulatively worth 40% of the final grade), Nevertheless, we observe consistent positive trends across assignments in both courses. A possible hypothesis is that while the chatbot may not drive treatment students to significantly outperform control group students on any one assignment, the bot results in slightly higher performance across assignments which cumulatively results in treated students earning higher final grades.

Messaging in both courses highlighted the availability of "supplemental instruction" (SI), a form of course-specific tutoring offered at GSU. SI could be one mechanism for higher course performance if students are more likely to attend SI and, in turn, gain a better understanding of course materials through SI sessions. Consistent with this possibility, across courses, students assigned to treatment are about two percentage points more likely to attend SI (Table 7). These effect sizes are large in relative terms compared to low baseline rates of SI attendance – 7% of students in Government and 2.4% of students in Microeconomics. We conduct a formal mediation analysis to see whether treatment assignment increases students' final grades through the mechanism of increased SI attendance. The natural indirect effect of tutoring on final numeric grades is about 0.32 grade points ($p=0.041$; 17.8% of the total effect) and the natural indirect effect of tutoring on earning a grade of B or higher is 0.004 percentage points ($p=0.046$, 9.6% of the total effect). Overall, we take this as suggestive evidence that treatment students are

[26] Sample size for assignments vary because the professor changed assignments across the intervention terms, dropping the fourth exam after the first intervention term, dropping the field trip after the second intervention term, and switching textbooks after the second intervention term to a system that did not track reading time.

more likely to attend SI, a plausible mechanism through which they may have done better in the course, though SI explains a small share of the overall treatment effect.

Mechanisms for heterogeneous treatment effects

Given the large treatment effects for women in Microeconomics, we explore in detail the mechanisms through which the intervention may have increased women's performance in the course. In Table 8, we examine the treatment effects on SI use and assignment completion in Microeconomics separately for men and women. We replicate the impact on final grades in the first row from Table 4, showing treated women earn final grades seven points higher than control women while there are no treatment effect for men. Looking at student behaviors, we find a statistically significant three percentage point increase in SI attendance for treated women and no treatment effect for men, though these estimates are not significantly different from each other (column 5). Using the same formal mediation analysis as with the full analytic sample, we find the natural indirect of SI attendance on final numeric grades for women in Microeconomics was about 0.95 grade points (*p=0.94*, 13.5% of the total treatment effect). We view these together as corroborating evidence that treatment may have worked through increasing SI attendance, but it was not the primary mechanism through which treatment improved final grades.

Turning to assignment completion, we see that for most of the assignment categories (e.g., practice assignments, quizzes) women in the treatment group have statistically higher completion rates than women in the control group, with no significant differences in assignment engagement for men by treatment assignment. For example, treated women are 8-9 percentage points more likely to complete practice assignments and earn assessment grades about seven points higher than women in the control group. Running formal mediation analyses, completion rates for each individual assignment explain large shares of the overall increase in final numeric grades for women in Microeconomics (when running models individually, the natural indirect effect of each assignment's completion explains between 48-73% of the total treatment effect). This suggests that chatbot messages can drive students to seek out academic support (SI tutoring) and shift completion of and performance on course assignments, which together contribute to higher final course grades.

Spillover and medium-term chatbot effects

Finally, we explore whether the chatbot affects students' outcomes outside of the focal intervention courses. Specifically, we test for *spillover* of the chatbot treatment onto students'

performance in the same semester and for *medium-term effects* of the chatbot treatment on students' persistence in the course subject. We did not have a prior hypothesis about spillover direction. On the one hand, if students have finite study time available and chatbot messages direct them to spend more time on the focal course, then we might observe a negative treatment effect on their grades in other courses. However, if messaging helps students develop better time management or study skills, they may leverage those skills to navigate all their courses and show higher performance overall during the intervention term. If these potential mechanisms operate in different ways for different students, then, on average, we may see little impact on students' performance.

In Table 9 we report on students' overall term GPA during the intervention term, their intervention term GPA excluding the focal (Government or Microeconomics) course (with different methods of treating withdrawals), and whether they enroll in another course in the subject the following term. We do not find strong evidence of positive or negative spillover or medium-term effects for the overall sample. We also consider each outcome for women in economics, where we previously found the strongest within-course effects, and did not observe significant effects of treatment on spillovers or subsequent semester enrollment for that subgroup. An open question remains regarding the extent to which an academic intervention in one course might affect students' long-term academic engagement and performance and the components necessary to affect such change.

Student attitudes

We fielded an end-of-course survey in Government; however, demographic selection into survey completion (see Appendix Table C1) tempers the applicability of the results. Only about half the students in Government completed the survey, and response rates were significantly higher for Asian or Hispanic students and lower for Black students, course re-takers, or students with lower high school GPAs. Among students who completed the survey, we find no consistent treatment effects on a suite of attitudinal measures (see Appendix C for a full list of measures and Appendix Table C2 for regression results). We note this as a potential area for future research.

Descriptive Analysis: Student Bot Experience and Engagement

About 90% of treated students who completed the end-of-course survey in Government recalled receiving chatbot communication, with 72% of students reporting that they read most of

the messages and 64% reporting that the weekly messages were helpful. While 65% of students said they knew the #quizme tool was available, only around 38% of students reported using #quizme. About a third (35%) of all respondents, representing 89% of the #quizme users, said that the #quizme tool was helpful. That such a high rate of tool users found it helpful, but only half of students aware of the tool reported using it, indicates opportunities for future work to explore how students decide whether to use academic supports and whether additional messaging can more effectively increase #quizme take-up. When asked whether GSU should continue using the chatbot, 82% indicated that chatbot use should continue in Government and should expand to other courses.

In Table 10 we summarize student engagement based on de-identified logs of messages exchanged between students and the chatbot platform (including pre-scheduled messages, automated bot responses, and supplemental human responses). The treatment was implemented as intended – over 98% of students assigned to treatment were sent at least one message.[27] Dosage was similar across courses – students received about 44-46 messages throughout the semester, inclusive of scheduled bot messages and responses to their inquiries. Relatively few students opted out – 4% in Government and 3% in Microeconomics.[28] About half of the students ever messaged back to the bot, with slightly higher reply rates in Government (54%). The average number of replies was higher in Government – an average of 4.6 replies overall (and 8.5 replies among the students who ever replied). This may be due to the use of the #quizme feature in Government which was designed for multiple back-and-forth messages as students attempted the sample quiz questions, while Microeconomics had fewer interactive components. However, students varied considerably in their active engagement – the highest engagers in Government and Microeconomics sent 76 and 24 messages, respectively, throughout the semester.

Descriptively, students who reply to the chatbot earn higher grades. In Figure 6 we plot the density of final grades for students in the treatment group who remained enrolled in the course. Those that ever replied have an average final grade of about 80 compared to about 71 for never-repliers. This difference is a statistically significant 8.5 percentage points based on the full

[27] This is not 100% because some students may have dropped the course between randomization (typically conducted the Friday prior to the first week of classes) and the first launch message being distributed (typically the Monday or Tuesday of the first week of classes).
[28] We code opt-outs based on student replies including use of formal opt-out language (e.g., "#pause") and informal requests (e.g., "stop txting me").

regression model that accounts for student covariates and randomization blocks. Looking at reply intensity rather than a binary measure of ever engaging with similar regression-based approach, among treated students who complete the course, each additional reply is associated with a 0.62 point difference in final course grade.[29]

Conceptualizing intervention costs

We did not conduct a formal cost-effectiveness analysis for this study.[30] However, following the Ingredients Method (Levin et al., 2018) for economic evaluation in education, we can categorize implementation inputs into four broad areas to inform how other institutions might weigh the cost of implementing such a tool: personnel, facilities, materials, and other costs (e.g., training). Because this intervention is low-touch and technology-based, the primary ingredients are personnel and technology-related materials. Personnel inputs include a staff lead responsible for directing the project and coordinating efforts across departments, a data analyst managing day-to-day data for targeted campaigns, TAs overseeing student questions and chatbot responses, and faculty collaborating with staff to develop course-specific content and tools (e.g., #quizme). The amount of time personnel devote to these tasks depends on factors such as prior experience with the technology and the stage of implementation. While the first semester of implementation may involve a steep learning curve, subsequent iterations tend to become more routine for experienced staff, reducing also the time required to support each additional academic chatbot. Although these implementation costs are not trivial, we hypothesize that GSU's prior investments in technology-enabled student support substantially lowered the marginal costs of implementation relative to institutions starting without an existing chatbot infrastructure or a centralized student-support office.

V. DISCUSSION

This study evaluates the effect of a course-specific academic chatbot to provide students with customized, timely, and regular notifications about course requirements and feedback on their performance in large, online undergraduate courses. Given prior work showing that chatbots can successfully improve students' completion of administrative college tasks, we

[29] Looking instead at the full sample of students assigned to treatment, whether they completed the course or not, ever engaging is associated with a 10.9 percentage point increase in final numeric grade, and each reply a student sends is associated with a 0.7-point increase in final grades.

[30] we are currently conducting such an analysis in related work that compares chatbot implementation across institutions operating under different organizational ecosystems

hypothesized that the course-specific integration of chatbot communication would improve overall course performance as well as timely completion of course tasks, such as assigned readings and exams. We find compelling evidence that the chatbot communication shifted students' final course grades, increasing the likelihood that students earn an A or B in the course. The chatbot was most effective at increasing final grades through the middle of the grade distribution. We also find suggestive evidence that the chatbot improved course completion, with treated students less likely to DFW in Government and treated students less likely to drop in Microeconomics. We find suggestive evidence that the chatbot affected final grades by changing students' engagement with academic tasks throughout the semester – attending tutoring and potentially by increasing assignment completion. Despite significant improvements in the focal course performance, these positive effects did not translate to improved performance in other courses. In short, chatbot messages led students to implement the specific advice offered (e.g., attend tutoring, complete your assignments in this course on time), but students did not appear to generalize these insights into other settings. This aligns with prior work finding that chatbot messaging can be effective at increasing completion of specific tasks but may not observably build underlying skills or knowledge that can be applied to other contexts.

In Microeconomics, subgroup analyses reveal large treatment effects of the chatbot on final grades , tutoring attendance, and assignment completion and performance particularly for women. There is a large, longstanding literature highlighting the underrepresentation of women in the economics profession and the importance of diversifying undergraduate economics departments to attract a more representative student body (Bayer & Rouse, 2016; Bayer et al., 2020; Dynan & Rouse, 1997; Yellen, 2019). Women are underrepresented in college majors such as economics and science, technology, engineering, and math (STEM) in part due to a lower sense of belonging as well as their lower persistence when they receive lower early college course grades (Allen & Robbins, 2008; Good, Rattan, & Dweck, 2012). Further, women frequently underestimate how well they are doing in a course or on tasks that are seen as stereotypically male (Coffman et al., 2024). Our findings suggest course chatbots – or other means of regular and proactive communication – are a potentially promising strategy to increase representation in economics, as students receive both up-to-date information about their course performance and encouragement to connect with campus supports, though more work is needed to understand longer-term effects.

The AI chatbot technology enables instructional teams to provide targeted, clear information to students about their course performance to date and the necessary tasks to complete to ensure success in the course. It is worth underscoring that the piloting and implementation of the technology required substantial upfront investment from the course instructional teams and the university support office, and piloting the tool for a semester prior to implementation was crucial for building the AI content knowledge base and accuracy of responses. In addition to targeting courses where students would likely benefit from the academic chatbot (e.g., high enrollment courses, virtual courses, courses with high "DFW" rates), the chatbot was also easier to launch in long-standing courses with established course syllabi and schedules. Course chatbot implementation was also facilitated by GSU's existing technology contracts and their combination of staff experience deploying chatbots and students' experience receiving chatbot messages regarding administrative tasks. Launching a similar tool at a university without such familiarity would involve additional start-up costs and efforts. Ultimately, we were encouraged by students' positive response to the chatbots, with high awareness of the tool and endorsement of GSU's use of course chatbots to support learning across subjects. Indeed, based on that feedback and this analysis, GSU has adopted the academic chatbot as a status quo tool in these courses.

Successful implementation also requires ongoing human monitoring of incoming messages to ensure students receive timely and accurate responses to their questions. For example, in one message, a student noted they had been dropped from the course due to tuition non-payment. The chatbot replied immediately with a website link to student financial services. In addition, the course's human teaching assistant followed up with a message that the professor could provide the student with access to the online course textbook while the student resolved their account hold so they would not fall behind on the reading. Successful implementation of a course chatbot requires sustained commitment and attention from the instructional team to ensure a high-quality student experience that is well aligned with the course itself. Notably, after the initial pilot period, weekly time spent monitoring and responding to messages declined substantially, with the course TA typically devoting less than two hours per week to system monitoring.

The exchange regarding tuition non-payment also highlights the importance of providing students with multiple communication channels to reach their instructional team and the pros and

cons of AI-enabled messaging. Some students may feel uncomfortable discussing sensitive topics – such as being dropped for account nonpayment – directly with an instructor but may feel more comfortable sharing such information via text message. In this way, the chatbot may support students' sense of psychological safety by offering another channel through which to develop positive relationships and establish trust (Wanless, 2016). Many factors hinder students from seeking help in introductory courses, including a lack of confidence and uncertainty about how to approach an authority figure such as a professor (Stitzel & Raje, 2022). A chatbot provides a low-stakes way of asking questions and can be a source of information that students can access immediately and free of concerns about judgment (de Gennaro et al., 2020). On the other hand, concern exists with individuals, especially youth, over-anthropomorphizing AI tools and developing unhealthy parasocial relationships with technology (Toppo, 2024). As technology advances, more work is needed to understand how college students perceive information they receive from these technologies and the impact of trust on the tools' efficacy.

Our work adds to a burgeoning literature around how to leverage faculty as student-success partners to improve student performance. The closest study to ours is by Carrell and Kurlaender (2023) who tested the effect of emails from course faculty providing students with feedback on their grades and encouraging them to access supplemental supports. Their pilot implementation targeted students who had not submitted the first course assignment (and were therefore starting the semester behind) and found an eight-percentage point increase in students' final course grades. Their scale-up targeted all students enrolled in the participating courses and found precise null overall effects on final course grades, though significant heterogeneous treatment effects. For example, the email communication increased the likelihood freshmen earned an A or B by nine percentage points, and their focal student populations of underrepresented minority students and underclass students were significantly more likely to persist in college. The authors note that context matters for anticipating the potential efficacy of outreach.

Despite similar effects across subjects and faculty in these analyses, we hypothesize the marginal benefit of the academic chatbot may vary by course context. We targeted large, online, asynchronous courses precisely because they were settings where many students struggle to complete the course. In contrast, effects of an academic chatbot may be smaller to null in courses with higher average performance. The tool may also be less effective in courses where professors

already engage in high-touch reminders and communication with students, though the professors teaching in our focal courses did engage in some personalized email communications with students. While many of the key components of the intervention – breaking down large assignments into manageable tasks, providing customized information about performance to date, and opening a line of communication between the students and instructional team – can translate across college subjects and courses, some features such as #quizme or specific questions about course content may be more difficult to scale. This tool is well suited to helping students through the administrative tasks of college courses (e.g., setting aside time for tutoring and completing assignments) and may be less effective in courses where the primary barriers to success are the difficulty of and students' preparation for handling the course content o. Future work on which we are embarking will explore the implementation and effectiveness of academic chatbots across other subjects and within different course structures to shed more light on these questions.

REFERENCES

Allen, J., & Robbins, S. B. (2008). Prediction of college major persistence based on vocational interests, academic preparation, and first-year academic performance. Research in Higher Education, 49, 62–79.

Bayer, A., Bruich, G., Chetty, R. & Housiaux, A. (2020) Expanding and diversifying the pool of undergraduates who study economics: Insights from a new introductory course at Harvard, The Journal of Economic Education, 51:3-4, 364-379, DOI: 10.1080/00220485.2020.1804511

Bayer, A. & Rouse, C. (2016). Diversity in the Economics Profession: A New Attack on an Old Problem. *Journal of Economic Perspectives*, 30 (4): 221-42**.**

Belley, P., & Lochner, L. (2007). The changing role of family income and ability in determining educational attainment. *Journal of Human Capital, 1*(1)

Bettinger, E., & Baker, R. (2014). The effects of student coaching: An evaluation of a randomized experiment in student advising. *Educational Evaluation and Policy Analysis, 36*(1), 3-19.

Bettinger, E., Doss, C., Loeb, S., Rogers, A., & Taylor, E. (2017). The effects of class size in online college courses: Experimental evidence. *Economics of Education Review, 5*8, 68-85.

Bettinger, E. P., Fox, L., Loeb, S., & Taylor, E. S. (2017). Virtual Classrooms: How Online College Courses Affect Student Success. *American Economic Review*, 107 (9): 2855-75.

Bick, A., Blandin, A., & Deming, D. (2024). The rapid adoption of generative AI. (No. w32966). National Bureau of Economic Research.

Bird, K., Castleman, B. L., & Lohner, G. (2022). Negative impacts from the shift to online learning during the COVID-19 crisis: Evidence from a statewide community college system. *AERA Open, 8*.

Carrell, S. E. & Kurlaender, M. (2020). My professor cares: Experimental evidence on the role of faculty engagement (No. w27312). National Bureau of Economic Research.

Castleman, B. L., & Long, B. T. (2016). Looking beyond enrollment: The causal effect of need-based grants on college access, persistence, and graduation. *Journal of Labor Economics, 34*(4).

Castleman, B. L., & Page, L. C. (2015). Summer nudging: Can personalized text messages and peer mentor outreach increase college going among low-income high school graduates? *Journal of Economic Behavior & Organization*, *115*, 144-160.

Castleman, B. L., & Page, L. C. (2016). Freshman year financial aid nudges: An experiment to increase FAFSA renewal and college persistence. *Journal of Human Resources*, *51*(2), 389-415.

Castleman, B. L., & Page, L. C. (2017). Parental influences on postsecondary decision making: Evidence from a text messaging experiment. *Educational Evaluation and Policy Analysis*, *39*(2), 361-377.

Castleman, B. L., Page, L. C., & Schooley, K. (2014). The forgotten summer: Does the offer of college counseling after high school mitigate summer melt among college-intending, low-income high school graduates? *Journal of Policy Analysis and Management*, *33*(2), 320-344.

Clotfelter, C. T., Hemelt, S. W., & Ladd, H. F. (2018). Multifaceted aid for low-income students and college outcomes: Evidence from North Carolina. *Economic Inquiry*, *56*(1), 278-303.

Coffman, K. B., Collis, M., & Kulkarni, L. (2024). Stereotypes and Belief Updating. *Journal of the European Economic Association 22*(3): 1011–1054.

Cohen, E., Huo, H., Guyot, K., Gaffney, C. & Christopher, E. (2024). A first look at the 2021 postsecondary enrollment, completion ,and financial aid outcomes of fall 2009 ninth-graders. NCES 2024-022. *National Center for Education Statistics*.

College Scorecard (n.d.). Georgia State University Profile. Retrieved from https://collegescorecard.ed.gov/.

Cortinhas, C. (2025). Does nudging higher education student improve attendance and does it matter? A quasi-natural experiment. *International Review of Economics Education, 49*, 100317.

De Giorgi, G., Pellizzari, M, & Woolston, W. G. (2009). Class size and class heterogeneity. IZA Discussion Papers 4443, Institute of Labor Economics (IZA).

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2023). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis, 46*(3), 483-505.

Dynan, K. E., & Rouse, C. E. (1997). The underrepresentation of women in economics: A study of undergraduate economics students. *The Journal of Economic Education*, *28*(4), 350-368.

Gandara, D., Anahideh, H., Ison, M. P., & Picchiarini, L. (2024). Inside the black box: Detecting and mitigating algorithmic bias across racialized groups in college student-success prediction. *AERA Open, 10*.

Giani, M. S., & Martin, A. (2021). Mobilizing developmental education: The causal effect of mobile app courseware on the college outcomes of developmental education students. *Educational Evaluation and Policy Analysis, 43*(4), 668-687.

Good, C., Rattan, A., & Dweck, C. S. (2012). Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology, 102*(4), 700–717. https://doi.org/10.1037/a0026659

Goodman, J. (2025). "Online Post-secondary Education," in *Live Handbook of Education Policy Research*, in Douglas Harris (ed.), Association for Education Finance and Policy, https://livehandbook.org/higher-education/miscellaneous/online-postsecondary-education/.

Goulas, S. (2024). Twelve facts about the economics of education. The Hamilton Project, The Brookings Institution. Retrieved from https://www.brookings.edu/wp-content/uploads/2034/06/20240627_THP_EducationFacts_PDF.pdf

Guo, K., Zhong, Y., Li, D., Chu, S. (2023). Effects of chatbot-assisted in-class debates on students' argumentation skills and task motivation, *Computers & Education, 203*(2023), https://doi.org/10.1016/j.compedu.2023.104862

Heß, S. (2017). Randomization inference with Stata: A guide and software. *Stata Journal, 17*(3), 630-651.

Holzer, H., & Baum, S., (2017). *Making College Work: Pathways to Success for Disadvantaged Students*. Washington, DC: Brookings Institution Press.

Jack, A. (2016). *The Privileged Poor: How Elite Colleges are Failing Disadvantaged Students*. Cambridge, MA: Harvard University Press.

Jurenka, I., …, & Ibrahim, L. (2024). Towards responsible development of generative AI for education: An evaluation-driven approach. Google Working Paper. Retrieved from https://storage.googleapis.com/deepmind-media/LearnLM/LearnLM_paper.pdf

Kena, G., Aud, S., Johnson, F., Wang, X., Zhang, J., Rathbun, A., ... & Kristapovich, P. (2014). The Condition of Education 2014. NCES 2014-083. *National Center for Education Statistics*.

Kofoed, M., Gebhart, L., Gilmore, D., & Moschitto, R. (2021). Zooming to class?: Experimental evidence on college students' online learning during COVID-19. IZA Discussion Papers 14356, Institute of Labor Economics (IZA).

Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica, 46*(1), 33-50.

Lareau, A. (2003). *Unequal childhoods: Race, class, and family life.* (Second, Ed.) Oakland, CA: University of California Press.

Lee, Y., Hwang, G., & Chen, P. (2022) Impacts of an AI-based chat bot on college students' after-class review, academic performance, self-efficacy, learning attitude, and motivation. *Educational technology research and development, 70*(5). 1843-1865.

Levin, H., McEwan, P., Belfield, C., Bowden, A., & Shand, R. (2018). *Economic Evaluation in Education*. SAGE.

Linkow, T., Miller, H., Parsad, A., Price, C., & Martinez, A. (2021). *Study of College Transition Messaging in GEAR UP: Impacts on Enrolling and Staying in College*. Institute of Education Sciences

Long, B. T., & Mabel, Z. (2012). Barriers to college success: Income disparities in progress to completion. Unpublished manuscript. Harvard University.

Ma, J., Pender, M, & Welch, M. (2019). Education Pays: 2019. Retrieved from: https://research.collegeboard.org/media/pdf/education-pays-2019-full-report.pdf

Marsicano, C., Felten, K., Toledo, L., & Buitendorp, M. (2020). Tracking campus responses to the COVID-19 pandemic. *APSA Preprints*. doi: 10.33774/apsa-2020-3wvrl.

NCES (2019). U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS) Fall Enrollment component, Spring 2013 through Spring 2019.

Oreopoulos, P., & Petronijevic, U. (2019). *The remarkable unresponsiveness of college students to nudging and what we can learn from it* (No. w26059). National Bureau of Economic Research.

Ortagus, J., Tanner, M., & McFarlin, I. (2020). Can re-enrollment campaigns help dropouts return to college? Evidence from Florida community colleges. *Educational Evaluation and Policy Analysis, 43*(1), 154-171.

Page, L. C., Castleman, B. L., & Meyer, K. (2020). Customized nudging to improve FAFSA completion and income verification. *Educational Evaluation and Policy Analysis*, *42*(1), 3-21.

Page, L. C., & Gehlbach, H. (2017). How an artificially intelligent virtual assistant helps students navigate the road to college. *AERA Open*, *3*(4).

Page, L. C., Kehoe, S. S., Castleman, B. L., & Sahadewo, G. A. (2019). More than dollars for scholars: The impact of the Dell Scholars Program on college access, persistence and degree attainment. *Journal of Human Resources, 54*(3), 683-725.

Page, L. C., Meyer, K., Lee, J., & Gehlbach, H. (2025). Conditions under which college students can be responsive to nudging. *Journal of Research on Educational Effectiveness.*

Pugatch, T., & Wilson, N. (2024). Nudging Demand for Academic Support Services: Experimental and Structural Evidence from Higher Education. *Journal of Human Resources, 59*(5), 1637-1682

Scrivener, S., Weiss, M. J., Ratledge, A., Rudd, T., Sommo, C., Fresques, H. (2015). Doubling graduation rates: Three-year effects of CUNY's Accelerated Study in Associate Programs (ASAP) for developmental education students. MDRC evaluation report. Retrieved from: https://www.mdrc.org/sites/default/files/doubling_graduation_rates_fr.pdf

Scuello, M., & Strumbos, D. (2024). Evaluation of Accelerate, complete, engage (ACE) at CUNY John Jay College of Criminal Justice: Final Report (ASAP/ACE (CUNY John Jay)). https://www.cuny.edu/wp-content/uploads/sites/4/page-assets/about/administration/offices/student-success-initiatives/asap/about/ace/300414_CUNY_March_2024_ACE_Final_Report_m1-1.pdf

Simon, H. A. (1982). *Models of bounded rationality.* Cambridge, MA: MIT Press

Smith, B. O., White, D. R., Kuzyk, P. C., & Tierney, J. E. (2018) Improved grade outcomes with an e-mailed "grade nudge", *The Journal of Economic Education, 49*:1, 1-7, DOI: 10.1080/00220485.2017.1397570

Snyder, T., & Dillow, S. (2015). Digest of Education Statistics 2013. Retrieved from https://nces.ed.gov/pubs2015/2015011.pdf.

Sommo, C., Slaughter, A., Saunier, C., Scrivener, S., & Warner, K. (2023). *Varying Levels of Success.* MDRC.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness.* London: Penguin Books.

Tian, X.,Risha, Z., Ahmed, I., Narayanan, A., & Biehl, J. (2021). Let's Talk It Out: A Chatbot for Effective Study Habit Behavioral Change. Proc. *ACM Hum.-Comput. Interact. 5, CSCW1, Article 97* (April 2021). https://doi.org/10.1145/3449171

Toppo, G. (2024, August 7). AI 'companions' are patient, funny, upbeat – and probably rewiring kids' brains. *The 74 Million*.

Walton, G., & Cohen, G. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology, 92*(1), 82-96.

Wang, R., Ribeiro, A., Robinson, C., Loeb, S., & Demszky, D. (2024). Tutor CoPilot: A human-AI approach for scaling real-time expertise. EdWorkingPapers Working Paper No. 24-1054. Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/81nh-8262

Wanless, S. B. (2016). The role of psychological safety in human development. *Research in Human Development*, *13*(1), 6-14.

Weiss, M. J., Ratledge, A., Sommo, C., & Gupta, H. (2019). supporting community college students from start to degree completion: Long-term evidence from a randomized trial of CUNY's ASAP. *American Economic Journal: Applied Economics*, *11*(3), 253-97.

Wells, G. (2023, August 24). Gen Z-ers are computer whizzes. Just don't ask them to type. *The Wall Street Journal*.

Westfall, P. H., Young, S.S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Hoboken, NJ: John Wiley & Sons

Yellen, J. (2019, September 24). Former Fed chair Janet Yellen on gender and racial diversity of the federal government's economists. The Brookings Institution forum on "The gender and racial diversity of the federal government's economics."

TABLES

Table 1: Analytic Sample and Randomization Balance

	Panel A: Pooled across courses			Panel B: Government			Panel C: Microeconomics		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Control mean	Treatment difference	N	Control mean	Treatment difference	N	Control mean	Treatment difference	N
Female	0.54	0.02	2483	0.56	0.03	1568	0.51	0.00	915
		(0.020)			(0.025)			(0.033)	
Asian	0.22	0.00	2483	0.23	0.00	1568	0.20	0.01	915
		(0.017)			(0.021)			(0.026)	
Black	0.46	0.03	2483	0.43	0.04	1568	0.51	0.01	915
		(0.020)			(0.025)			(0.033)	
White	0.22	-0.03	2483	0.23	-0.03	1568	0.20	-0.02	915
		(0.016)			(0.021)			(0.025)	
Hispanic	0.14	-0.01	2483	0.15	-0.01	1568	0.12	0.00	915
		(0.014)			(0.018)			(0.021)	
First Generation	0.22	0.02	2483	0.24	0.02	1568	0.20	0.01	915
		(0.017)			(0.022)			(0.027)	
Pell Eligible	0.57	0.00	2483	0.58	-0.02	1568	0.55	0.02	915
		(0.020)			(0.025)			(0.033)	
Course Re-takers	0.11	0.01	2483	0.08	0.01	1568	0.15	0.01	915
		(0.012)			(0.014)			(0.024)	
Freshman	0.52	-0.01	2483	0.63	0.01	1568	0.33	-0.04	915
		(0.017)			(0.022)			(0.028)	
Upperclassman	0.40	0.00	2483	0.30	-0.01	1568	0.55	0.03	915
		(0.018)			(0.021)			(0.032)	
Transfer	0.08	0.01	2483	0.07	0.00	1568	0.12	0.02	915
		(0.011)			(0.013)			(0.021)	
High School GPA	3.47	0.02	2043	3.52	0.01	1393	3.37	0.05	650
		(0.017)			(0.020)			(0.033)	
Joint F-Test		0.5571			0.4746			0.7442	

Notes: Robust standard errors in parentheses. Includes randomization blocks. Models pooling across subjects include subject fixed effects. High school GPA reported here excludes missing cases.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 2: Intent-to-treat effect of course chatbot on final grades

	Panel A: Pooled across courses				Panel B: American Government				Panel C: Microeconomics			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Control mean	Treatment Effect	Treatment Effect	*MCC p-value*	Control mean	Treatment Effect	Treatment Effect	*MCC p-value*	Control mean	Treatment Effect	Treatment Effect	*MCC p-value*
Final Grade	71.07	1.58	1.80	*0.460*	71.88	0.81	1.24	*0.680*	69.66	2.90	3.07	*0.480*
		(1.283)	(1.226)			(1.539)	(1.463)			(2.279)	(2.193)	
Earned A	0.36	0.04 +	0.04 *	*0.170*	0.31	0.04	0.04 +	*0.240*	0.44	0.03	0.04	*0.560*
		(0.019)	(0.018)			(0.024)	(0.022)			(0.032)	(0.031)	
Earned B or higher	0.61	0.04 *	0.04 **	*0.090*	0.61	0.04	0.04 *	*0.240*	0.62	0.05	0.05 +	*0.270*
		(0.019)	(0.018)			(0.024)	(0.022)			(0.031)	(0.030)	
Earned C or higher	0.73	0.02	0.02	*0.460*	0.75	0.01	0.02	*0.660*	0.71	0.04	0.04	*0.480*
		(0.017)	(0.017)			(0.022)	(0.021)			(0.029)	(0.028)	
D or F	0.15	-0.01	-0.02	*0.530*	0.15	-0.02	-0.02	*0.420*	0.15	0.00	0.00	*0.920*
		(0.014)	(0.013)			(0.017)	(0.017)			(0.023)	(0.023)	
Withdrew	0.04	0.00	0.00	*0.820*	0.03	0.00	0.00	*0.680*	0.07	0.00	-0.01	*0.920*
		(0.008)	(0.008)			(0.008)	(0.008)			(0.016)	(0.016)	
DFW	0.19	-0.02	-0.02	*0.460*	0.18	-0.03	-0.03 +	*0.310*	0.22	-0.01	-0.01	*0.920*
		(0.015)	(0.015)			(0.019)	(0.018)			(0.027)	(0.026)	
Dropped	0.07	0.00	0.00	*0.820*	0.07	0.01	0.01	*0.680*	0.08	-0.03 +	-0.03 +	*0.270*
		(0.010)	(0.010)			(0.013)	(0.014)			(0.016)	(0.016)	
Covariates included			X	X			X	X			X	X
N students		2483	2483	2483		1568	1568	1568		915	915	915

Notes: Robust standard errors in parentheses. Includes randomization blocks. Models pooling across subjects include subject fixed effects. "DFW" stands for earning a D or F in the course or withdrawing from the course. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA (imputed as zero for missing cases, with a covariate flag for having a missing GPA). MCC p-value reports on the Westfall & Young p-value after multiple comparison correction.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 3: Conditional outcomes among course completers

	Panel A: Full sample		Panel B: Excluding early drops		Panel C: Excluding all drops and withdrawals	
	(1)	(2)	(3)	(4)	(5)	(6)
	Control mean	Treatment Effect	Control mean	Treatment Effect	Control mean	Treatment Effect
Final Grade	71.07	1.803	76.56	1.789 +	80.26	1.301
		(1.2262)		(1.0213)		(0.8175)
Earned A	0.36	0.036 *	0.39	0.036 +	0.40	0.037 +
		(0.0177)		(0.0186)		(0.0191)
Earned B or higher	0.61	0.044 **	0.66	0.047 **	0.69	0.044 **
		(0.0180)		(0.0180)		(0.0179)
Earned C or higher	0.73	0.025	0.79	0.025	0.83	0.021
		(0.0167)		(0.0157)		(0.0149)
D or F	0.15	-0.016	0.16	-0.018	0.17	-0.020
		(0.0135)		(0.0144)		(0.0148)
Withdrew	0.04	-0.005	0.05	-0.007	0.00	0.000
		(0.0079)		(0.0085)		(0.0000)
DFW	0.19	-0.020	0.21	-0.025	0.17	-0.020
		(0.0148)		(0.0157)		(0.0148)
Dropped	0.07	-0.004	0.00	0.000	0.00	0.000
		(0.0104)		(0.0000)		(0.0000)
Covariates included		X		X		X
N students		2483		2306		2204

Notes: Robust standard errors in parentheses. Includes randomization blocks. Models pooling across subjects include subject fixed effects. "DFW" stands for earning a D or F in the course or withdrawing from the course. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA (imputed as zero for missing cases, with a covariate flag for having a missing GPA). Panel B excludes students flagged as "dropping" the course (during the formal university add/drop period), while Panel C also excludes students who withdrew from the course prior to the university withdrawal deadline.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 4: Treatment effect of chatbot on final grades, by gender, Microeconomics

	Panel A: Women				Panel B: Men			
	(1)	(2)		(3)	(4)	(5)	(6)	(7)
	Control mean	Treatment Effect		*MCC p-value*	Control mean	Treatment Effect	*MCC p-value*	Test of Equality
Final Grade	68.86	7.00	*	*0.080*	70.48	-0.68	*1.000*	0.071
		(3.066)				(3.201)		
Earned A	0.43	0.07		*0.310*	0.46	0.00	*1.000*	0.238
		(0.045)				(0.044)		
Earned B or higher	0.60	0.11	**	*0.030*	0.64	0.00	*1.000*	0.058
		(0.043)				(0.043)		
Earned C or higher	0.71	0.10	**	*0.040*	0.71	-0.02	*0.990*	0.033
		(0.040)				(0.042)		
D or F	0.14	-0.03		*0.610*	0.16	0.02	*0.970*	0.252
		(0.032)				(0.034)		
Withdrew	0.07	-0.01		*0.630*	0.07	0.00	*1.000*	0.814
		(0.021)				(0.024)		
DFW	0.21	-0.04		*0.590*	0.22	0.02	*0.980*	0.246
		(0.036)				(0.038)		
Dropped	0.09	-0.06	**	*0.020*	0.07	0.00	*1.000*	0.065
		(0.023)				(0.024)		
Covariates included		X		X		X	X	
N students		463		463		452	452	915

Notes: Robust standard errors in parentheses. Includes randomization blocks. "DFW" stands for earning a D or F in the course or withdrawing from the course. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA. MCC p-value reports on the Westfall & Young p-value after multiple comparison correction. Test of equality evaluates equality of the treatment effect coefficient from separate regressions.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 5: Completion of and Performance on Government Assignments

	(1) Control Mean	(2) Treatment Effect	(3) Treatment Effect
Final grade	71.88	0.81	1.24
		(1.539)	(1.463)
Reading score	78.31	0.39	0.74
		(1.679)	(1.623)
Completed Exam 1	0.86	0.00	0.00
		(0.018)	(0.018)
Performance on Exam 1	65.75	-0.31	0.23
		(1.505)	(1.425)
Complete Exam 2	0.83	0.01	0.01
		(0.019)	(0.019)
Performance on Exam 2	60.22	1.29	1.80
		(1.525)	(1.446)
Completed Exam 3	0.82	0.00	0.01
		(0.019)	(0.019)
Performance on Exam 3	62.62	0.57	1.09
		(1.591)	(1.516)
N students		1568	1568
Completed Exam 4	0.81	0.02	0.02
		(0.034)	(0.033)
Performance on Exam 4	58.14	1.30	1.75
		(2.617)	(2.483)
N students		509	509
Completed Field Trip	0.79	0.03	0.03
		(0.025)	(0.024)
Grade on Field Trip	83.20	3.08	3.05
		(2.768)	(2.644)
Reading minutes	578.85	2.92	1.60
		(21.018)	(20.780)
N students		990	990
Covariates included			X

Notes: Robust standard errors in parentheses. Includes randomization blocks. Course assignments changed across intervention terms; exam 4 was only administered the first intervention term and the field trip assignment was only required the first two intervention terms. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA. +p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 6: Completion of and Performance on Microeconomics Assignments

		(1)	(2)	(3)	
		Control Mean	Treatment Effect	Treatment Effect	
Final course grade		69.66	2.90	3.07	
			(2.279)	(2.193)	
Participation (10%)	Grade on Pre/Post Quizzes	73.50	1.81	2.02	
			(2.485)	(2.401)	
	Completed Discussion Posts	0.61	0.00	0.00	
			(0.032)	(0.031)	
	Grade in Discussion Posts	65.20	1.16	1.19	
			(2.637)	(2.565)	
Practice Interactive Tools (15%)	Completed	0.67	0.03	0.04	
			(0.030)	(0.030)	
	Grade	77.05	3.37	3.64	
			(2.487)	(2.419)	
Assessment Interactive Tools (25%)	Completed	0.62	0.02	0.02	
			(0.032)	(0.031)	
	Grade	69.97	2.56	2.79	
			(2.440)	(2.351)	
Practice Quizzes (15%)	Completed	0.74	0.04	0.05	+
			(0.028)	(0.027)	
	Grade	70.49	3.45	3.59	+
			(2.234)	(2.149)	
Assessment Quizzes (35%)	Completed	0.60	0.01	0.01	
			(0.032)	(0.031)	
	Grade	64.10	3.08	3.23	
			(2.184)	(2.096)	
N students			915	915	
Covariates included				X	

Notes: Robust standard errors in parentheses. Includes randomization blocks. Percentages listed next to assignment components reference the weight each assignment received in final grade calculations. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 7: Treatment effect on take-up of supplemental instruction

	(1)	(2)		(3)		(4)
	Control Mean	Treatment		Treatment		N
Pooled: Used SI	0.05	0.02	*	0.02	*	2483
		(0.010)		(0.010)		
American Government: Used SI	0.07	0.03	+	0.02		1568
		(0.014)		(0.014)		
Microeconomics: Used SI	0.02	0.02		0.02		915
		(0.012)		(0.012)		
Covariates included				X		

Notes: Robust standard errors in parentheses. Includes randomization blocks. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA. +p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 8: Treatment effect of chatbot on within-course outcomes, by gender, Microeconomics

		Panel A: Women			Panel B: Men		
		(1)	(2)		(3)	(4)	(5)
		Control Average	Treatment Effect		Control Average	Treatment Effect	Test of Equality
Final course grade		68.86	7.00	*	70.48	-0.68	0.071
			(3.066)			(3.201)	
Used SI		0.03	0.03	+	0.02	0.00	0.230
			(0.019)			(0.015)	
Participation (10%)	Grade on Pre/Post Quizzes	72.81	6.58	*	74.22	-2.77	0.044
			(3.275)			(3.564)	
	Completed Discussion Posts	0.61	0.03		0.60	-0.05	0.194
			(0.045)			(0.045)	
	Grade in Discussion Posts	67.11	4.09		63.26	-1.72	0.245
			(3.578)			(3.780)	
Practice Interactive Tools (15%)	Completed	0.67	0.08	*	0.67	-0.01	0.116
			(0.041)			(0.042)	
	Grade	76.96	7.28	*	77.14	0.41	0.141
			(3.334)			(3.546)	
Assessment Interactive Tools (25%)	Completed	0.60	0.09	+	0.63	-0.04	0.036
			(0.044)			(0.044)	
	Grade	69.03	6.93	*	70.93	-0.98	0.082
			(3.272)			(3.440)	
Practice Quizzes (15%)	Completed	0.73	0.09	**	0.74	0.00	0.090
			(0.038)			(0.040)	
	Grade	69.91	7.18	**	71.08	0.19	0.093
			(2.972)			(3.165)	
Assessment Quizzes (35%)	Completed	0.58	0.07		0.62	-0.06	0.035
			(0.044)			(0.044)	
	Grade	62.38	7.62	**	65.85	-1.15	0.031
			(2.951)			(3.054)	
Covariates included			X			X	
N students			463			452	915

Notes: Robust standard errors in parentheses. Includes randomization blocks. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA. Test of equality evaluates equality of the treatment effect coefficient from separate regressions.

+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 9: Treatment effect of chatbot on non-course academic outcomes

		Pooled				American Government				Microeconomics			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
		Control mean	Treatment effect	Treatment effect	N	Control mean	Treatment effect	Treatment effect	N	Control mean	Treatment effect	Treatment effect	N
Overall semester performance	Term GPA including focal course	2.57	0.05	0.06	2483	2.60	0.08	0.11 +	1568	2.52	-0.02	-0.01	915
			(0.053)	(0.048)			(0.066)	(0.059)			(0.088)	(0.083)	
Performance in other courses, excluding focal course	Earned credits	0.90	0.00	0.00	2483	0.91	-0.01	-0.01	1568	0.88	0.02	0.02	915
			(0.012)	(0.012)			(0.015)	(0.015)			(0.021)	(0.021)	
	Term GPA (zero if missing)	2.43	0.01	0.02	2483	2.50	0.05	0.07	1568	2.31	-0.05	-0.04	915
			(0.057)	(0.053)			(0.071)	(0.065)			(0.094)	(0.089)	
	Term credits (zero if missing)	7.65	0.09	0.13	2483	8.04	0.18	0.23	1568	6.98	-0.06	0.01	915
			(0.180)	(0.165)			(0.225)	(0.206)			(0.296)	(0.276)	
	Term GPA (zero only if withdrew from other courses)	2.67	0.01	0.03	2263	2.72	0.08	0.09	1437	2.58	-0.09	-0.06	826
			(0.053)	(0.049)			(0.066)	(0.059)			(0.089)	(0.085)	
	Term credits (zero only if withdrew from other courses)	8.40	0.11	0.17	2263	8.74	0.28	0.32 +	1437	7.79	-0.15	-0.04	826
			(0.167)	(0.151)			(0.204)	(0.183)			(0.280)	(0.265)	
Persistence in subject	Took a course in department next term	0.13	0.00	0.00	2483	0.07	-0.01	-0.01	1568	0.24	0.01	0.01	915
			(0.013)	(0.013)			(0.013)	(0.013)			(0.028)	(0.028)	
Covariates included				X				X				X	

Notes: Robust standard errors in parentheses. Includes randomization blocks. Models pooling across subjects include subject fixed effects. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA. Whether student has a spillover measure is an indicator for whether the student (1) completed the intervention course and (2) completed at least one other course that semester. Term GPA (measured on a 4.0 scale) and term hours (with each GSU course bearing about 3 credit hours) calculated only for students with a spillover measure. The "zero if missing" rows impute a zero value of GPA and credits if the student is missing those measures for any reason. The "zero only if withdrew from other courses" measure imputes a zero value of GPA and credits if the student withdrew from all other courses during the semester (e.g., they attempted other courses but did not complete them) and drops students who did not attempt any other courses in that semester (e.g., the student only attempted the focal course).
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

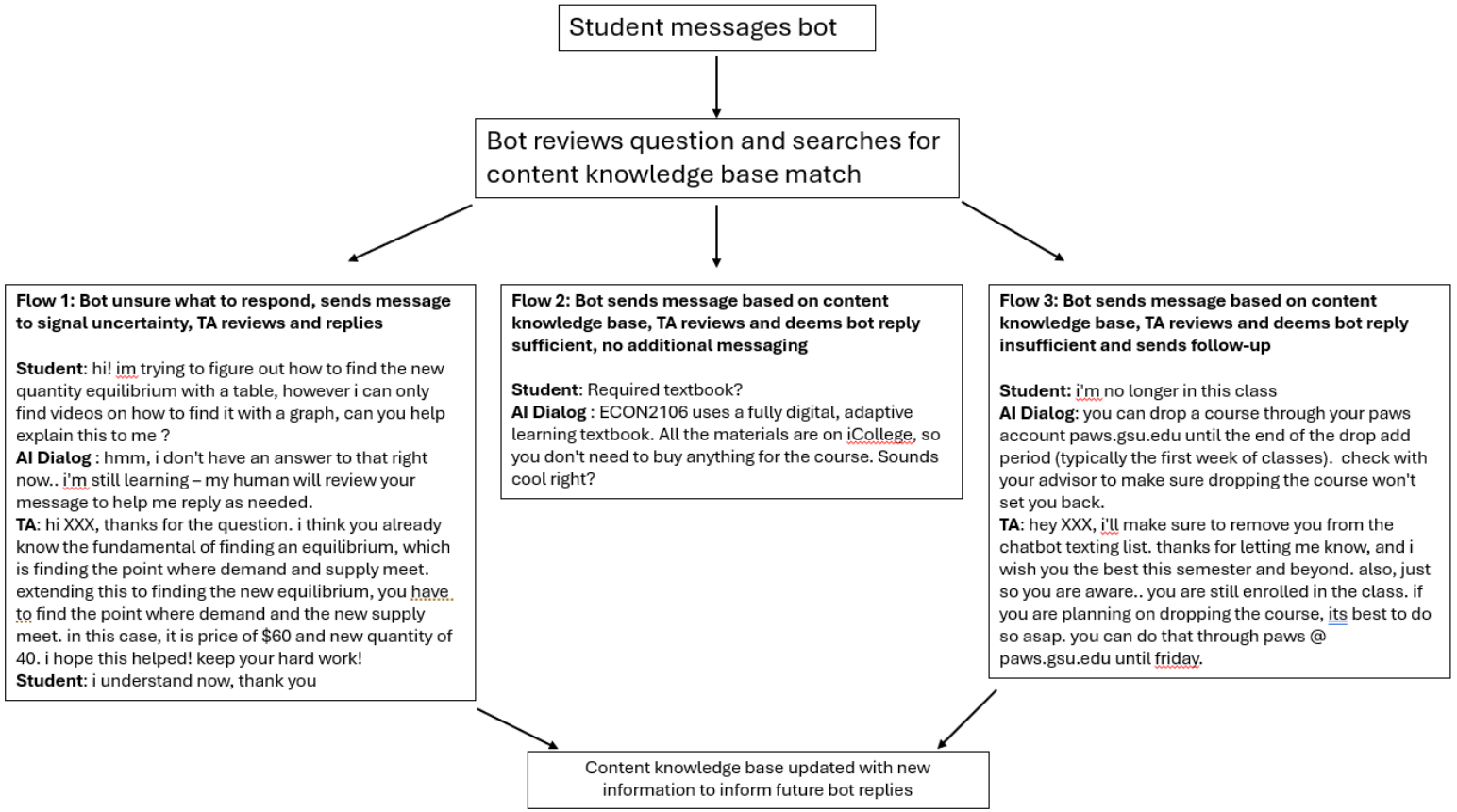
Table 10: Chatbot engagement summary, treated students

	Ever received	Total received	Ever opt out	Ever Reply	Total Replies
American Government	0.99	46.01	0.04	0.54	4.63
	[0.080]	[23.636]	[0.206]	[0.499]	[9.172]
Microeconomics	0.98	44.30	0.03	0.43	1.07
	[0.138]	[14.449]	[0.165]	[0.496]	[2.087]

Note: Summarizes chatbot engagement among treated students. Standard deviations in brackets.

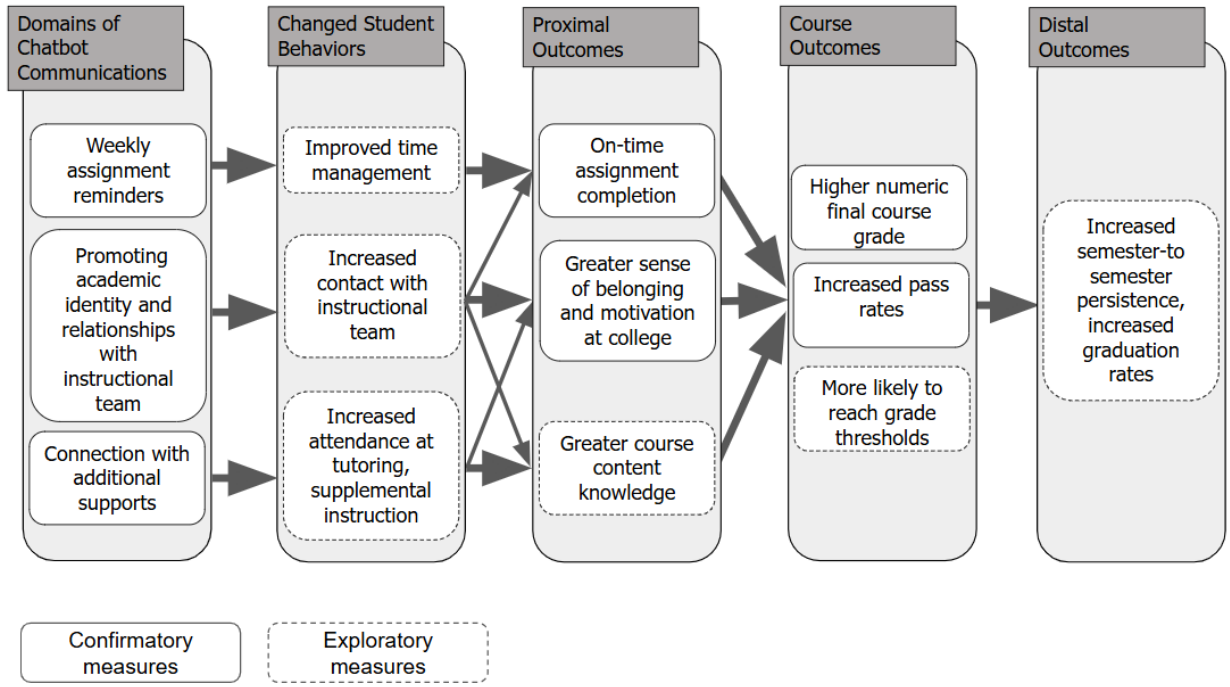
FIGURES

Figure 1. Course chatbot workflow



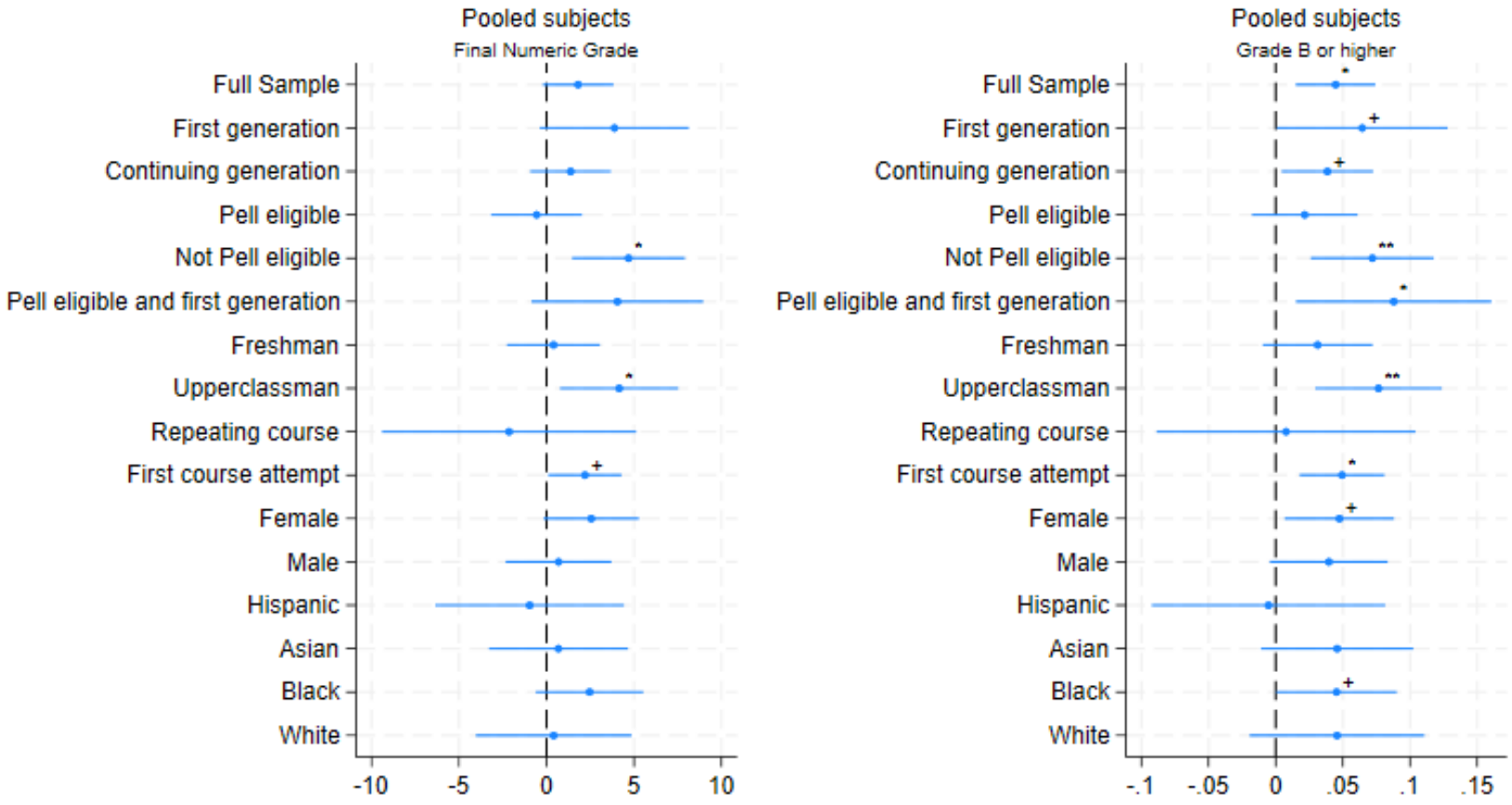
Notes: Figure illustrates how the academic course chatbot reviews student messages and how the combination of bot assessment of content knowledge and human teaching assistant review determines the messages students receive.

Figure 2. Course chatbot theory of change



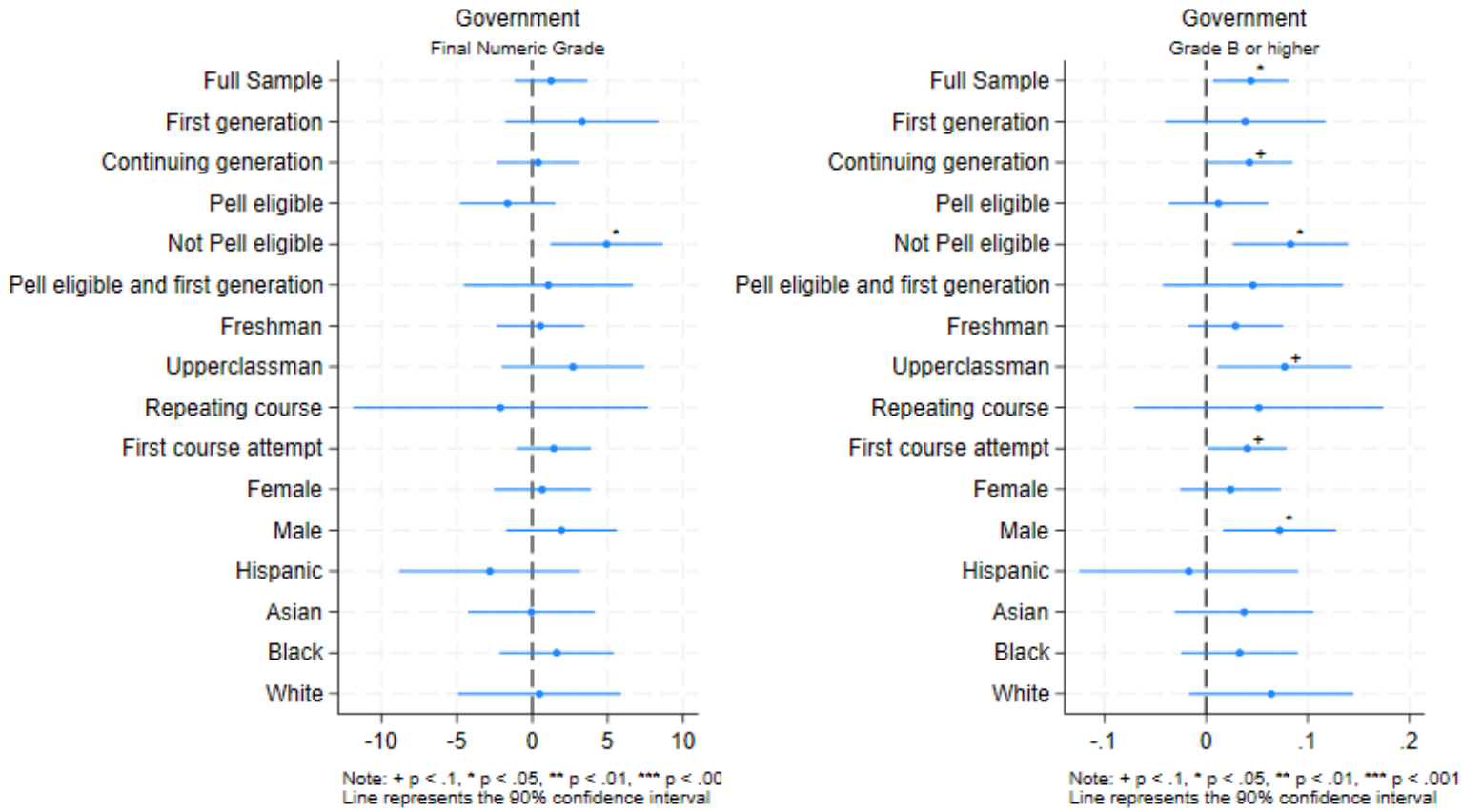
Notes: Figure illustrates the theory of changes for the course chatbot, outlining the primary domains of the chatbot, the intended changes to student behavior in response to each type of message, the proximal within-course outcomes that would reflect those changed behaviors, the final course outcomes, and the distal college outcomes hypothesized to be affected by the chatbot communications.

Figure 3. Heterogeneous treatment effect of course chatbot, pooled across subjects



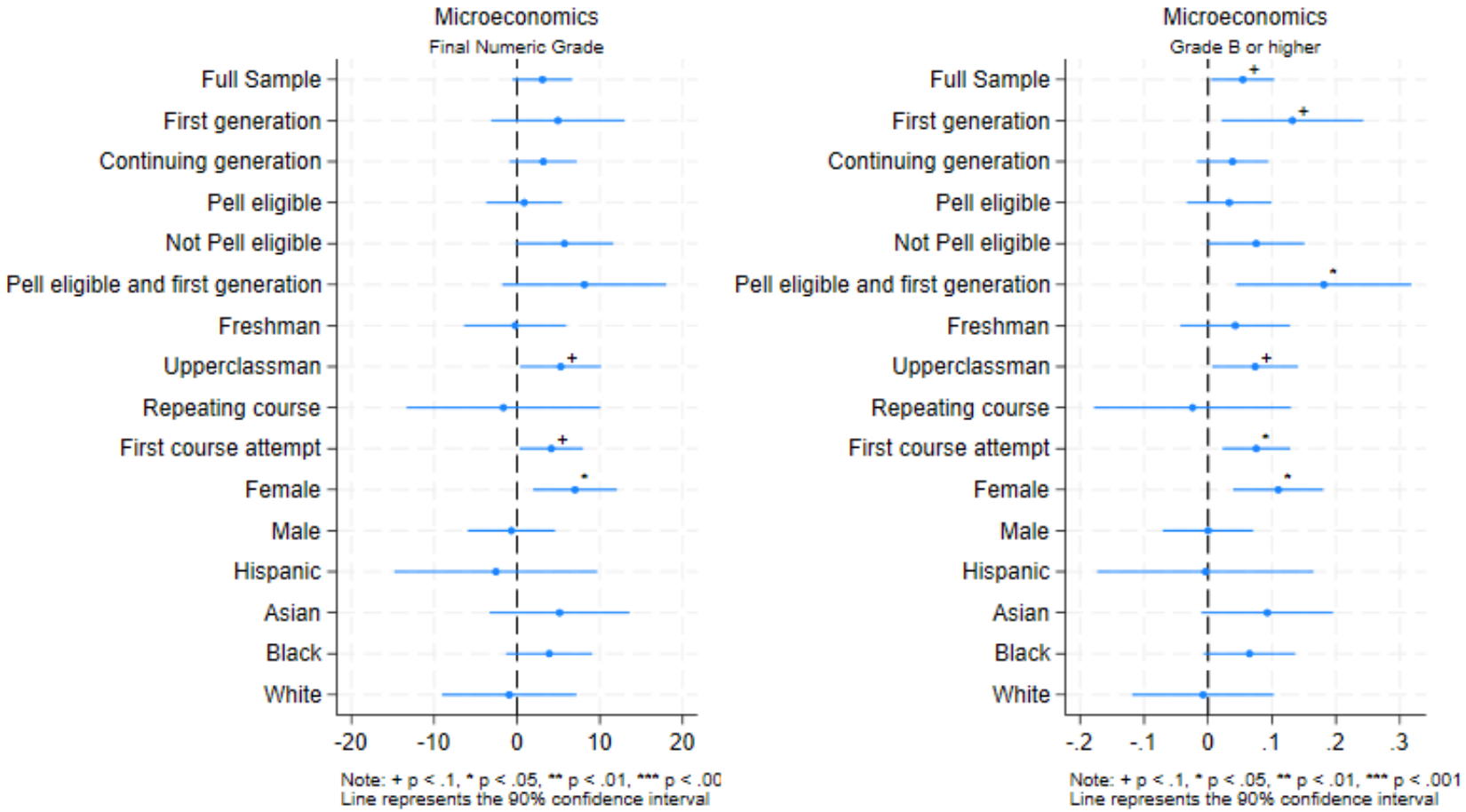
Notes: Each coefficient plotted represents the treatment coefficient from the main analytic model for the listed group of students, including the full set of other covariates and randomization blocks described in equation 1, with the outcome of interest being students' final numeric grade in the course or whether the student received an A or B in the course. The line surrounding each point estimate represents the 90% confidence interval. +p<0.10, *p<0.05, **p<0.01, ***p<0.001

Figure 4a. Heterogeneous treatment effect of course chatbot, Government



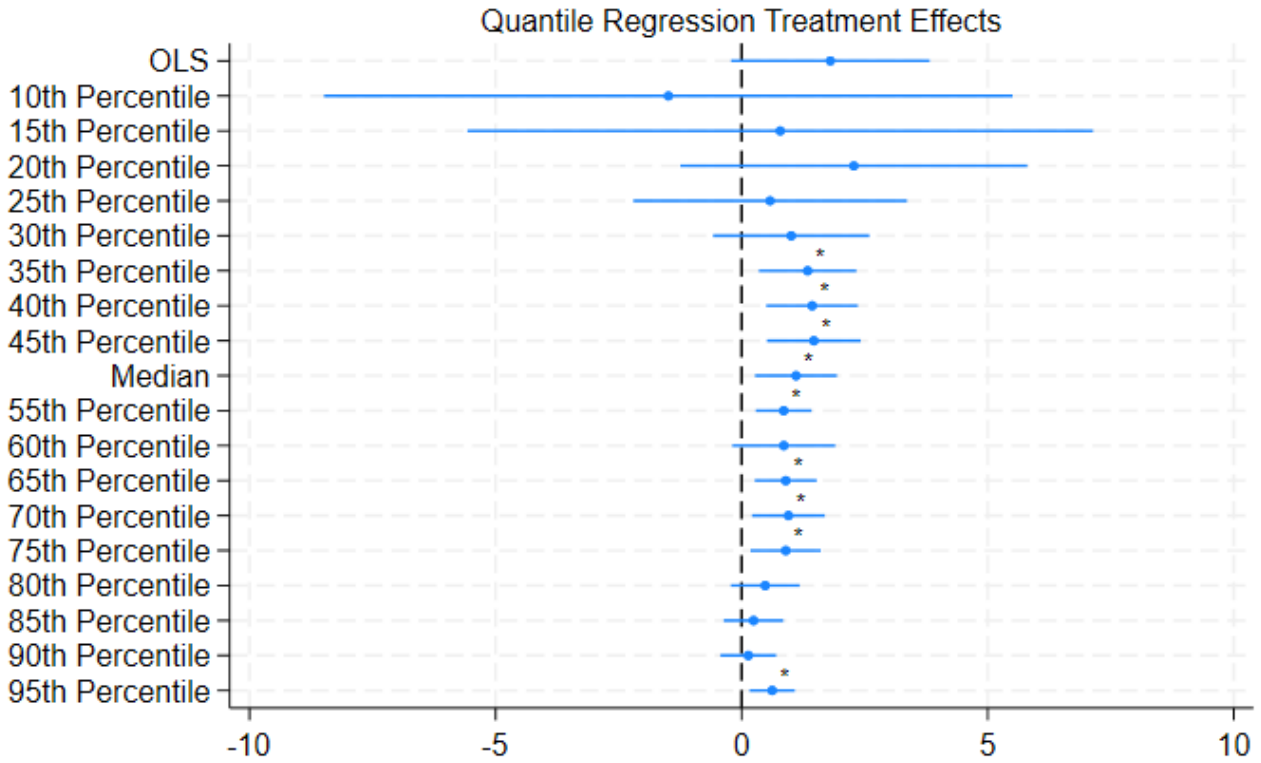
Notes: Each coefficient plotted represents the treatment coefficient from the main analytic model for the listed group of students enrolled in the Government course, including the full set of other covariates and randomization blocks described in equation 1, with the outcome of interest being students' final numeric grade in the course or whether the student received an A or B in the course. The line surrounding each point estimate represents the 90% confidence interval. +p<0.10, *p<0.05, **p<0.01, ***p<0.001

Figure 4b. Heterogeneous treatment effect of course chatbot, Microeconomics



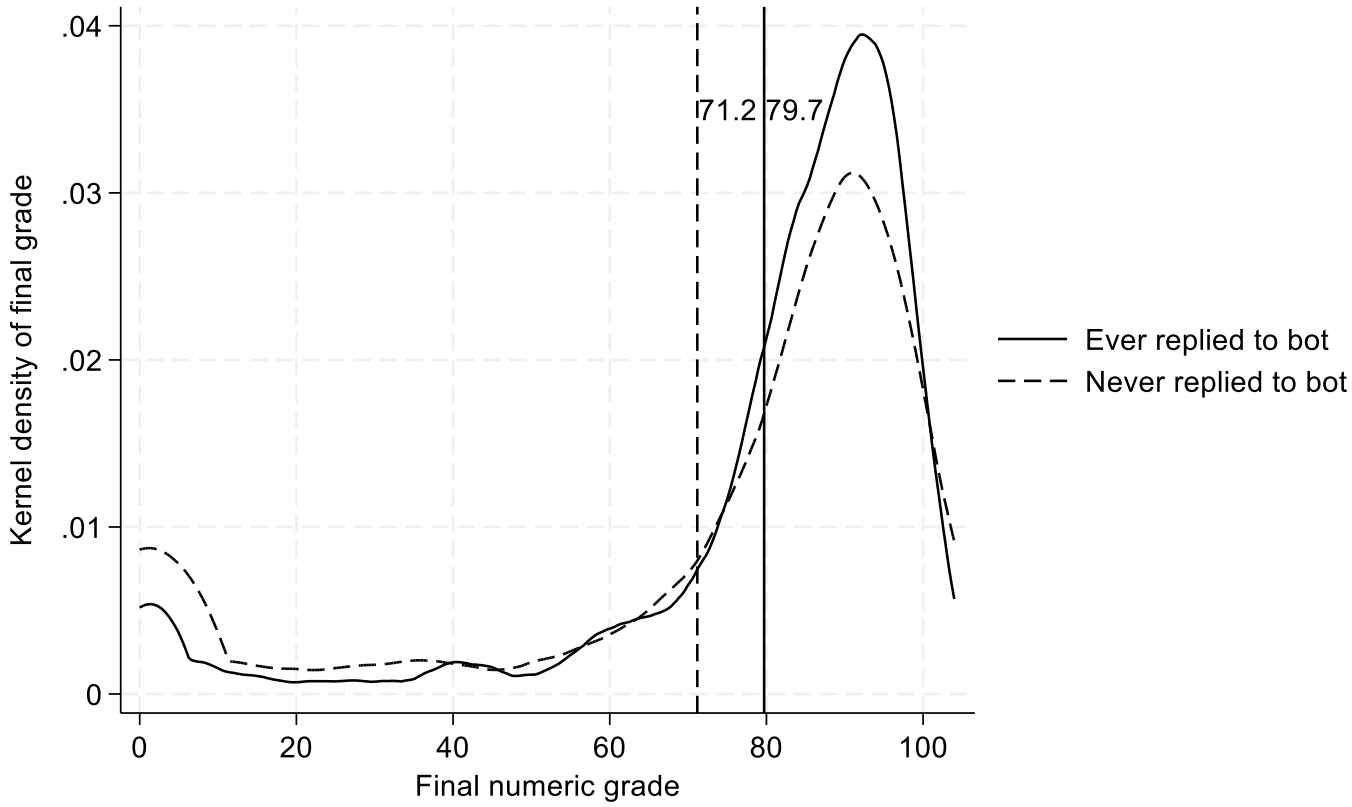
Notes: Each coefficient plotted represents the treatment coefficient from the main analytic model for the listed group of students enrolled in the Microeconomics course, including the full set of other covariates and randomization blocks described in equation 1, with the outcome of interest being students' final numeric grade in the course or whether the student received an A or B in the course. The line surrounding each point estimate represents the 90% confidence interval. +p<0.10, *p<0.05, **p<0.01, ***p<0.001

Figure 5. Quantile treatment effects on final grade, full sample



Notes: Each coefficient plotted represents the treatment coefficient from a quantile regression model for each listed percentile point in the distribution, including the full set of other covariates and randomization blocks described in equation 1, with the outcome of interest being students' final numeric grade in the course. The line surrounding each point estimate represents the 90% confidence interval. +p<0.10, *p<0.05, **p<0.01, ***p<0.001

Figure 6. Final numeric grade by chatbot engagement, among treated students



Notes: Figure plots the final numeric grade students received in the course as reported in the course section gradebook, pooling Government and Microeconomics sections together. Sample limited to students assigned to treatment who remained enrolled in the course until the end of the semester, with final grade density separately plotted by whether students ever sent at least one message to the bot (51% of treated students). N=1,201.

APPENDIX A

Appendix Table A1: Analytic Sample and Randomization Balance, Government

	Fall 2021		Spring 2022		Fall 2022		Pooled	
	Control	Treatment difference	Control	Treatment difference	Control	Treatment difference	Control	Treatment difference
Female	0.62	0.01	0.60	0.06	0.47	0.02	0.56	0.03
		(0.043)		(0.044)		(0.042)		(0.025)
Asian	0.19	-0.01	0.27	-0.01	0.24	0.01	0.23	0.00
		(0.034)		(0.040)		(0.036)		(0.021)
Black	0.46	0.06	0.43	0.04	0.40	0.03	0.43	0.04
		(0.044)		(0.045)		(0.041)		(0.025)
White	0.27	-0.05	0.20	-0.01	0.23	-0.02	0.23	-0.03
		(0.038)		(0.036)		(0.035)		(0.021)
Hispanic	0.15	-0.01	0.15	0.01	0.15	-0.03	0.15	-0.01
		(0.031)		(0.033)		(0.029)		(0.018)
First Generation	0.24	0.01	0.24	0.04	0.23	0.02	0.24	0.02
		(0.039)		(0.040)		(0.035)		(0.022)
Pell Eligible	0.63	-0.04	0.59	0.01	0.53	-0.02	0.58	-0.02
		(0.043)		(0.045)		(0.042)		(0.025)
Course Re-takers	0.07	0.01	0.13	0.02	0.07	0.00	0.08	0.01
		(0.023)		(0.031)		(0.019)		(0.014)
Freshman	0.43	0.05	0.65	0.00	0.78	-0.01	0.63	0.01
		(0.043)		(0.043)		(0.028)		(0.022)
Upperclassman	0.49	-0.04	0.27	-0.01	0.17	0.00	0.30	-0.01
		(0.043)		(0.040)		(0.027)		(0.021)
Transfer	0.07	0.00	0.08	0.01	0.05	0.01	0.07	0.00
		(0.023)		(0.025)		(0.018)		(0.013)
High School GPA	3.45	0.02	3.46	0.05	3.61	-0.02	3.52	0.01
		(0.035)		(0.036)		(0.032)		(0.020)
Joint F-Test		0.878		0.743		0.934		0.475
N students		509		481		578		1568

Notes: Robust standard errors in parentheses. Includes randomization blocks. High school GPA reported here excludes missing cases.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Appendix Table A2: Analytic Sample and Randomization Balance, Microeconomics

	Fall 2022		Spring 2023		Pooled	
	Control	Treatment difference	Control	Treatment difference	Control	Treatment difference
Female	0.54	-0.02	0.47	0.02	0.51	0.00
		(0.048)		(0.045)		(0.033)
Asian	0.20	-0.01	0.19	0.02	0.20	0.01
		(0.038)		(0.036)		(0.026)
Black	0.49	0.04	0.54	-0.02	0.51	0.01
		(0.048)		(0.045)		(0.033)
White	0.19	-0.03	0.20	-0.02	0.20	-0.02
		(0.037)		(0.035)		(0.025)
Hispanic	0.13	0.01	0.11	0.00	0.12	0.00
		(0.033)		(0.028)		(0.021)
First Generation	0.20	0.03	0.21	0.00	0.20	0.01
		(0.040)		(0.037)		(0.027)
Pell Eligible	0.57	-0.02	0.54	0.05	0.55	0.02
		(0.048)		(0.045)		(0.033)
Course Re-takers	0.14	-0.01	0.16	0.02	0.15	0.01
		(0.034)		(0.034)		(0.024)
Freshman	0.13	-0.03	0.50	-0.06	0.33	-0.04
		(0.031)		(0.045)		(0.028)
Upperclassman	0.70	0.00	0.42	0.05	0.55	0.03
		(0.044)		(0.045)		(0.032)
Transfer	0.16	0.03	0.07	0.01	0.12	0.02
		(0.037)		(0.024)		(0.021)
High School GPA	3.40	0.03	3.36	0.06	3.37	0.05
		(0.053)		(0.042)		(0.033)
Joint F-Test		0.9918		0.7759		0.7442
N students		426		489		915

Notes: Robust standard errors in parentheses. Includes randomization blocks. High school GPA means reported here include zeros for missing cases.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Appendix Table A3: Randomization inference p-values

	Pooled Sample				Panel B: American Government				Panel C: Microeconomics			
	(1)		(2)		(1)		(2)		(1)		(2)	
	Treatment Effect		Inference p-value	N	Treatment Effect		Inference p-value	N	Treatment Effect		Inference p-value	N
Final Grade	1.80		0.131	2483	1.24		0.399	2483	3.07		0.135	2483
	(1.226)				(1.463)				(2.193)			
Earned A	0.04	*	0.036	2483	0.04	+	0.069	2483	0.04		0.260	2483
	(0.018)				(0.022)				(0.031)			
Earned B or higher	0.04	**	0.009	2483	0.04	*	0.061	2483	0.05	+	0.076	2483
	(0.018)				(0.022)				(0.030)			
Earned C or higher	0.02		0.123	2483	0.02		0.345	2483	0.04		0.141	2483
	(0.017)				(0.021)				(0.028)			
D or F	-0.02		0.233	2483	-0.02		0.132	2483	0.00		0.865	2483
	(0.013)				(0.017)				(0.023)			
Withdrew	0.00		0.580	2483	0.00		0.584	2483	-0.01		0.718	2483
	(0.008)				(0.008)				(0.016)			
DFW	-0.02		0.167	2483	-0.03	+	0.110	2483	-0.01		0.672	2483
	(0.015)				(0.018)				(0.026)			
Dropped	0.00		0.682	2483	0.01		0.406	2483	-0.03	+	0.075	2483
	(0.010)				(0.014)				(0.016)			
Covariates included	X		X		X		X		X		X	

Notes: Robust standard errors in parentheses calculated using randomization inference with the Stata command *ritest*. Includes randomization blocks. Models pooling across subjects include subject fixed effects. "DFW" stands for earning a D or F in the course or withdrawing from the course. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA.

+p<0.10, *p<0.05, **p<0.01, ***p<0.001

APPENDIX B – Sample chatbot Messages

LAUNCH MESSAGE_08.23.2021

Department / Office	Political Science 1101
Purpose	Launching TA Pounce to students in POLS 1101 (group 1)
Target Population	180
Successful Contacts	176

Script

Hi name_first! I'm the chatbot for American Government.
This term I'm working with Dr. Evans to help you stay on track. I'll send you course reminders and tips to succeed. You can text me questions anytime! So hit me up and I'll do my best to get you the answer.

Contacts without this profile information receive a backup text.

Pro-tip: Start each week with Dr. Evans' announcement.
👉 bit.ly/pols1101ann
If you don't want these messages, just text #PAUSE to stop (but I hope you'll give me a chance).

WEEK 1 GENERAL_08.24.2021

Department / Office	Political Science 1101
Purpose	Weekly reminder of upcoming due dates sent to all students
Target Population	178
Successful Contacts	174

Script

WEEKLY DIGEST 😈
Hi - each week I'll send you a reminder of upcoming due dates. Last year, almost all students found these weekly digest messages helpful. They help you have all the info before you make a plan to complete your coursework.
Use this link to access Dr. Evans' announcement
👉 bit.ly/F21pols1101ann

PRO-TIP: Download the Exam 1 study guide and fill out Ch. 1 this week as you read.
DUE THIS WEEK: You have 6 tasks (2hrs. total) due 🧩 WK1 Due Date.
1 Watch COURSE INTRO VIDEO
2 Read Syllabus
3 Take Syllabus Quiz
4 Watch AREA9 INTRO VIDEO
5 Read Chapter 1
6 Take Pre-Course Survey
👉 bit.ly/F21pols1101toc

WEEK 3 CUSTOMIZED DIGEST_ALL COMPLETE_09.07.2021

Department / Office	Political Science 1101
Purpose	Weekly reminder of upcoming due dates + personalized message to students who have completed all previously due graded requirements
Target Population	180
Successful Contacts	173

Script

WEEKLY DIGEST

Hi name_first ! I see you've completed all your reading so far this semester. Keep up the good work!
Reviewing the readings will help with Exam 1 next week. Supplemental Instruction (SI) can also help.

Contacts without this profile information receive a backup text.

Did you know that students who regularly attend SI score an average of one letter grade higher than students who don't? Come check it out!
BITLY: SI

PRO-TIP: Download & fill out the Exam 1 study guide. Schedule time in your calendar to fill out Chs. 1&2 if you haven't already.
DUE THIS WEEK: You have 2 tasks (about 2hrs total) due WK3 Due Date .
1 Read Chapter 3
2 Complete "Activity: Know Thy Political Self?"
BITLY: TOC

WEEK 3 CUSTOMIZED DIGEST_MISSING_09.07.2021

Department / Office	Political Science 1101
Purpose	Weekly reminder of upcoming due dates + personalized message to students who have at least missing graded requirement (<70%)
Target Population	24
Successful Contacts	23

Script

WEEKLY DIGEST

Hi name_first ! You seem to have a missing assignment. Check your iCollege email to see how to make it up. I'm here to help, so text me with any questions. Exam 1 is next week. Supplemental Instruction (SI) is a great way to prepare.

Contacts without this profile information receive a backup text.

Did you know that students who regularly attend SI score an average of one letter grade higher than students who don't? Can you find an hour this week to attend SI?

BITLY: SI

PRO-TIP: Download & fill out the Exam 1 study guide. Schedule time in your calendar to fill out Chs. 1&2 if you haven't already.

DUE THIS WEEK: You have 2 tasks (about 2hrs total) due WK3 Due Date .

1 Read Chapter 3

2 Complete "Activity: Know Thy Political Self?"

BITLY: TOC

WEEK 3 CUSTOMIZED DIGEST_WORK AHEAD_09.07.2021

Department / Office	Political Science 1101
Purpose	Weekly reminder of upcoming due dates + personalized message to students who have already completed all graded requirements due in course so far including the current week--students have worked ahead
Target Population	6
Successful Contacts	5

Script

WEEKLY DIGEST

Hi name_first ! I see you've already completed the assignments for this week. That's truly awesome—keep up the excellent work!
Exam 1 is next week. Supplemental Instruction (SI) is a great way to review the reading you've done.

Contacts without this profile information receive a backup text.

Did you know that students who regularly attend SI score an average of one letter grade higher than students who don't? Come check it out:
BITLY: SI

PRO-TIP: Download & fill out the Exam 1 study guide this week if you haven't already. This is also a great time to complete the NCCHR assignment early for extra credit.
Have a great week!

LAUNCH #QUIZME/INTRO TYLER_09.10.2021

Department / Office	Political Science 1101
Purpose	Encouraging message to all students introducing Tyler (but not COMMAND #tyler) and introducing COMMAND #quizme
Target Population	239
Successful Contacts	235

Script

Howdy name_first! This is Tyler—the human behind the chatbot for American Gov. Many students have told me it can be hard to know if they have studied enough for a test. So I've set up a feature for you in this chatbot called: #quizme

Contacts without this profile information receive a backup text.

Text back the command #quizme (include the hashtag) anytime to start a short quiz with questions covering concepts on Exam 1 coming up in 5 days. After you take the quiz, hit me up with any questions. Teamwork makes the dream work. Let's be a team!

WEEK 5 CUSTOMIZED DIGEST_ MISSING EXAM 1_ 09.20.2021

Department / Office	Political Science 1101
Purpose	Week 5 message to students who did not complete Exam 1.
Target Population	10
Successful Contacts	10

Script

WEEKLY DIGEST

Hi name_first ! I see you did not take Exam 1. Follow this link to take action on making up the exam. Do this today! Dr. Evans will only allow make-ups for a few days.

bit.ly/exam1makeup

Contacts without this profile information receive a backup text.

PRO-TIP: If you complete the NCCHR assignment this week, it could add up to 3 points to your final course grade.

DUE THIS WEEK: You have 2 tasks (about 2hrs. total) due WK5 Due Date .

1 Read Chapter 4

2 Take Check-in Survey I

BITLY: TOC

ENCOURAGEMENT WK5_09.23.2021

Department / Office	Political Science 1101
Purpose	Encouragement message sent to all students addressing how students may be feeling overwhelmed at this point in the semester.
Target Population	225
Successful Contacts	215

Script

Howdy name_first ! Tyler here—the human behind the chatbot for American Gov. Students have told me they feel overwhelmed at this point in the semester. Especially after the first exam, it's totally normal to feel this way.

Contacts without this profile information receive a backup text.

It's also totally normal for this feeling to pass, so I encourage you to continue working hard. I'm here for you, too. Text in your questions any time and if the bot can't answer them, I will. I wish you the best of luck this semester. I'm rooting for you big time!

ENCOURAGEMENT INTERACTIVE WK10_10.28.2021

Department / Office	Political Science 1101
Purpose	Encouraging message sent to all student asking them to share how they felt about Exam 2.
Target Population	169
Successful Contacts	155
Response Rate	29%

Script

Howdy name_first ! Tyler here! Sooo.. how'd the exam go?! I know Exam 2 is typically the hardest one. How did you feel about it? REPLY 1/2/3
Contacts without this profile information receive a backup text.

[1]: Good! 😁
[2]: Meh... 😐
[3]: Not so good. 😒

1 Good! 😁

That's great to hear! So that we can better support our students, would you mind sharing what you found most helpful during studying or on the exam? I'm sure your peers would love to hear any advice you have to share. We won't include your name!

2 Meh... 😐

So that we can better support our students, would you mind sharing what went well and what didn't go so well during studying or on the exam? Is there anything Dr. Evans can do to help support you on exams?

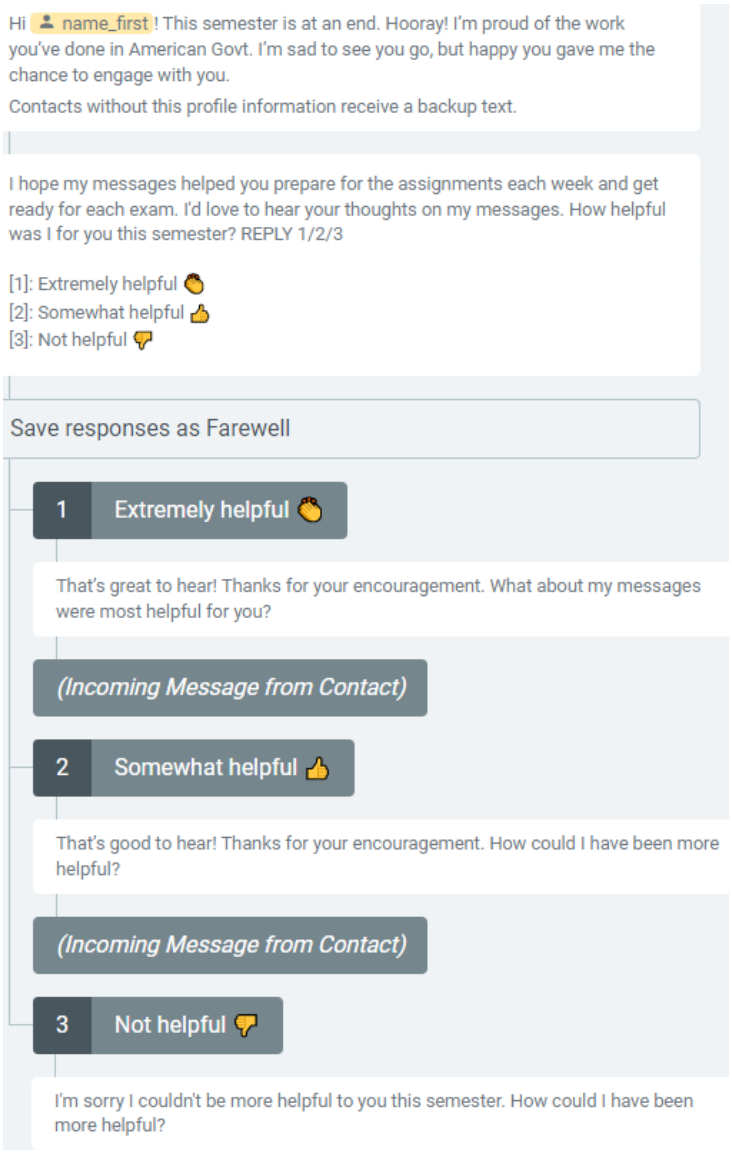
3 Not so good. 😒

So that we can better support our students, would you mind sharing what didn't go so well during studying or on the exam? Is there anything Dr. Evans can do to help support you on exams?

FAREWELL INTERACTIVE MESSAGE_12.13.2021

Department / Office	Political Science 1101
Purpose	Farewell message wishing them well and asking for their quick feedback on how helpful the bot was for them this semester.
Target Population	225
Successful Contacts	207

Script



APPENDIX C – Attitudinal Survey Measures

Organizational Support (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I know how the new things we're learning in this class connect to what we've learned before.
- This instructor regularly checks in to make sure we understand the class material.
- I feel like this class is organized to help me do well.
- It's clear what we're supposed to be doing in this class.
- I can communicate with this instructor about this class as needed.

Institutional Growth Mindset (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- This instructor seems to believe that students have a certain amount of intelligence, and they really can't do much to change it.

Self-Efficacy (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I have felt confident about my ability to do well in this class.

Inspiring Expectations (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I feel like this instructor trusts I can persist through challenging course material.
- I feel like this instructor thinks I can learn anything that is taught in classes.
- I feel like this instructor expects I will keep improving as a student.
- I feel like this instructor believes I have real potential in school.
- I feel like this instructor sees me as someone who could be successful in academics.
- I feel like this instructor recognizes that I can earn good grades if I put the effort in.

Adaptive Student Attributions (1 = "Not at all likely"; 5 = "Extremely likely")
If the following situation occurred during this course, how likely is it that you would have the thoughts below?

- You have to miss an exam for personal reasons.
    - I would think, "This instructor will be inflexible or unsupportive"
    - I would think, "This instructor will be understanding and helpful"
- You fall behind on the coursework one week, and the instructor messages you to say they noticed you still needed to turn things in.
    - I would think, "The instructor thinks I don't care about my education"
    - I would think, "The instructor is concerned about how I'm doing"
- You are doing poorly in the course and are at risk of failing.
    - I would think, "The instructor probably thinks I should drop the course."
    - I would think, "The instructor probably thinks I can pick my grade up."

GSU Challenge/Threat Ratio (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I feel like GSU will be a positive challenge for me.
- I feel like I have what I need to be successful at GSU.

- I am worried that some of the work at GSU will be stressful or overwhelming. (reverse-coded)
- I am uncertain if I could perform well in future GSU courses (reverse-coded)

GSU Social Belonging and Belonging Uncertainty (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I feel like I belong at GSU.
- I feel comfortable in classes at GSU.
- I feel accepted at GSU.
- I feel like I can be myself at GSU.
- Sometimes I feel that I belong at GSU, and sometimes I feel that I don't belong. (reverse-coded)

Trust and Fairness (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- This instructor treats me with respect.
- I trust this instructor to treat me fairly.
- I feel like the instructor truly has the best interest of their students in mind. (Eric added)

Meaningful Work (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- In this class, we do meaningful work, not busy work.
- What we learn in this class is connected to real-life.
- This teacher makes what we're learning really interesting.
- I feel like the course material to be relevant or useful to my life.
- I have been able to connect the course material to my interests or values.

Nervousness with Instructor (1= "Not at all nervous"; 5= "Extremely nervous")

- Imagine you decided to meet one-on-one with the instructor.
    - How nervous would you be about meeting this instructor?
    - How nervous would you be about having something to talk about?
    - How nervous would you be that they might judge you if you ask a "dumb" question?

Teacher Caring (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I feel like this instructor is glad that I am in their class.

Broad Regard (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I feel like this instructor would like to learn about my life outside of school.
- I feel like this instructor cares about what I do outside of my coursework.
- I feel like this instructor recognizes I have many identities beyond being a student.
- I feel like this instructor sees me as a person with many goals and values.
- I feel like this instructor welcomes my personal background and history.
- I feel like this instructor appreciates that I spend time on interests outside of schoolwork.

Appendix Table C1: Selection into end-of-course survey completion, Government

	Didn't complete survey	Completed Survey	Difference	
Treatment	0.493	0.505	0.013	
			(0.028)	
Female	0.554	0.582	0.041	
			(0.028)	
Asian	0.141	0.276	0.124	***
			(0.022)	
Black	0.552	0.398	-0.148	***
			(0.028)	
White	0.204	0.229	0.035	
			(0.023)	
Hispanic	0.117	0.155	0.039	*
			(0.019)	
First Generation	0.259	0.244	-0.016	
			(0.024)	
Pell Eligible	0.604	0.558	-0.038	
			(0.028)	
Course Re-takers	0.147	0.060	-0.100	***
			(0.018)	
Freshman	0.578	0.661	0.022	
			(0.025)	
Upperclassman	0.358	0.268	-0.033	
			(0.025)	
Transfer	0.063	0.071	0.010	
			(0.015)	
High School GPA	0.513	0.395	-0.132	***
			(0.028)	
N students	505	1063	1568	

Notes: Robust standard errors in parentheses for model reporting difference in characteristics among the survey completers and non-completers; mode includes randomization blocks.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Appendix Table C2: Treatment effect on student attitudes, Government

	(1)	(2)	(3)	(4)
	Control Mean	Treatment	Treatment	N
Completed survey	0.67	0.01	0.01	1568
		(0.023)	(0.022)	
Organizational Support (1-6)	4.45	0.04	0.03	1063
		(0.073)	(0.072)	
Self-Efficacy (1-6)	4.67	0.00	0.00	1063
		(0.073)	(0.072)	
Adaptive Student Attributions (1-5)	3.81	0.01	0.00	1063
		(0.047)	(0.047)	
Perception of Achievable Challenge (1-6)	3.78	0.00	-0.02	1061
		(0.048)	(0.048)	
Sense of Social Belonging (1-6)	4.20	-0.05	-0.07	1056
		(0.053)	(0.053)	
Trust and Fairness (1-6)	5.01	0.04	0.04	1055
		(0.053)	(0.052)	
Meaningful Work (1-6)	4.73	0.03	0.02	1065
		(0.057)	(0.056)	
Level of Nervousness with Instructor (1-5)	2.52	0.07	0.08	1061
		(0.067)	(0.066)	
Broad Regard (1-6)	4.03	0.04	0.05	1061
		(0.061)	(0.062)	
Covariates included			X	

Notes: Robust standard errors in parentheses. Includes randomization blocks. Item scale in parentheses next to index. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA. Sample size reported separately for each construct measured.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001