Surveying the Impact of Generative Artificial Intelligence on Political Science Education

Nicole Wu, *University of Toronto, Canada*
Patrick Y. Wu, *New York University, USA*

ABSTRACT   Recent applications of new innovations in artificial intelligence have brought up questions about how this new technology will change the landscape and practices in a wide range of industries and sectors. This article focuses on the impact of generative large language models on teaching, learning, and academic assessment in political science education by analyzing two novel surveys administered by the discipline's major professional body, the American Political Science Association. We present the results of these surveys and conclude with recommendations.

Generative large language models (LLMs) can produce original content in response to custom prompts at user-specified lengths in seconds. These tools are capable of writing human-like prose and answering a wide variety of questions (Zhao et al. 2023). Popular platforms such as ChatGPT (Brown et al. 2020) and Google Gemini are free or low cost, sparking concerns that these tools will be used by students to automatically generate essays and answers to tests. Some are concerned that LLMs would aggravate issues stemming from "contract cheating"—that is, purchasing work from a commercial third party—because artificial intelligence (AI) tools are inexpensive. Although it is difficult to obtain reliable data on changes in LLM usage by students due to social-desirability bias and sampling issues, the awareness and popularity of these tools have increased rapidly since late 2022.

Responses from universities and instructors vary. Key debates center on whether the use of LLMs constitutes plagiarism; the extent to which these tools would improve or impede student learning and career readiness; and whether to explicitly ban, incorporate, or encourage the use of AI tools. These differences in philosophy are apparent from university statements. For example, Harvard Law School's "Statement on Use of AI Large Language Models" states that the use of AI "in preparing to write, or writing, academic work for courses…is prohibited unless expressly identified in writing by the instructor as an appropriate resource…." Conversely, the Sandra Day O'Connor College of Law at Arizona State University explicitly permits their applicants to use generative AI, citing its prevalence in the legal field and the school's mission "to educate and prepare the next generation of lawyers and leaders."

In classrooms, instructors often have leeway in making generative AI policies. Princeton University states that the university "doesn't intend to ban ChatGPT or to levy a top-down edict about how each instructor should address the AI program…." Likewise, at DePaul University, "Faculty have the discretion to allow or encourage students to use AI in class activities and/or assignments." In many ways, instructors across institutional types (e.g., R1, R2, and liberal arts colleges) are left on their own to design their own assessments and evaluate students. The goal of this article is not to provide a definitive guide on the best response to LLMs—it is a determination that must be made by individual instructors. We provide data on what political science educators are doing in response to these new applications of AI and describe key considerations to guide these decisions using insights from colleagues that we surveyed. This study aligns with a growing literature in pedagogy about the opportunities and challenges of LLMs in education (e.g., Kasneci et al. 2023; Tlili et al. 2023).

Specifically, we present data from two original surveys collected in collaboration with the American Political Science Association (APSA) from April to August 2023 (Wu and Wu 2024). These surveys examined how colleagues perceive the utility and

Corresponding author: Nicole Wu 🔟 *is assistant professor of political science at the University of Toronto. She can be reached at nicolek.wu@utoronto.ca.*
Patrick Y. Wu 🔟 *is a postdoctoral fellow at New York University's Center for Social Media and Politics. He can be reached at pyw230@nyu.edu.*
Both authors made equal contributions.

impact of generative LLMs in their classroom. The first and shorter survey was sent to current and recent members of the professional body as part of APSA's 2023 membership survey. In the second survey, a subset of political science academics was asked further about whether they have taken new measures in response to the proliferation of these tools and to determine whether written responses from our essay bank were produced by students or bots. We find that (1) their overall evaluation of AI's impact on political science education is slightly negative but that they deem certain applications to be positive (e.g., copyediting and formatting); (2) educators plan to change their teaching approaches, but any changes primarily are on enforcement rather than integrating AI into their courses; and (3) respondents who participated in the bot- versus student-written essays identification exercise were able to make a correct determination no better than a coin toss (with a success rate between 45% and 53%).

The following two sections describe our data-collection method and present more detailed results from our two surveys. The final section evaluates common tactics reported by colleagues in response to the increase in LLMs.

DATA AND METHODS

Between April and August 2023, we conducted two surveys jointly with APSA to better understand how political science academics perceive the utility and impact of generative LLMs. The surveys were approved by the University of Toronto's Research Ethics Board (Protocol #00044085). The first survey, consisting of four questions, was embedded within APSA's annual membership survey and distributed to approximately 11,000 individuals. We received 1,615 responses: 62.6% were faculty, 11.8% were graduate students, 4.6% had an adjunct or visiting professor lecturer position, and 3.7% were postdoctoral fellows. Of these respondents, 81.3% were located primarily in the United States. Of the total number of respondents, 938 reported that they were from PhD-granting institutions, 146 from MA-granting institutions, 224 from four-year (BA-granting) institutions, 31 from two-year institutions or community colleges, and 18 from other institutions; 258 did not respond to this question. Among those who answered this question, the average respondent received their PhD in 2004. We refer to this first survey as the "membership survey." We conducted a second, longer survey targeting a subset of APSA members who had participated in the Teaching and Learning Conference (TLC) within the past five years. This "TLC survey" was sent to 3,442 individuals; we received 198 responses. TLC participants may be more engaged and interested in pedagogical issues than an average political science instructor in higher education. The TLC survey allowed us to introduce more questions and components, including requesting respondents to identify bot- and human-produced essays from an essay bank.

These two surveys complemented one another by allowing us to make better observations about a large and important subset of the discipline while enabling a more detailed solicitation of views on these new technologies. APSA administered both opt-in surveys. Although we cannot observe the counterfactual outcome, surveys distributed by the discipline's professional body likely yielded higher response rates than if the surveys were distributed by us alone. Due to this data-collection method, our sampling frame is limited to current and recent APSA members; however, the surveys reached a larger group of political science educators to whom we otherwise would not have had access.

Because the LLMs impact written work mostly in terms of course assignments, we first explored the pedagogical importance of written assignments as perceived by colleagues. We then collected data on attitudes toward LLMs and views on specific applications of these tools. In the TLC survey, we also asked participants about their responses to these generative AI tools in their classroom and to identify bot- versus student-written essays in our essay bank. The next section presents the results of the two surveys.

RESULTS

This section describes our six main findings from the two surveys. The surveys reveal a mix of skepticism, uncertainty, and recognition of the potential impact of AI on teaching and learning.

Written Assignments Are Deemed Important in Political Science Courses

Political science educators believe that political science is a writing-heavy major. All TLC respondents stated that written work is important for upper-level students; only 1.27% stated that it is "not at all important" for lower-level students. This trend held across the membership survey. Figure 1 shows that improving writing skills, writing argumentatively, and generating original arguments are among the most important pedagogical goals of written assignments among respondents on the membership survey. We did not include those who responded "Don't know or prefer not to answer" or skipped the question.

We also allowed respondents to add, as a free response, what they considered important about writing assignments. Common themes included thinking critically and analytically, demonstrating comprehension of materials, synthesizing knowledge, evaluating concepts, and building professional skills.
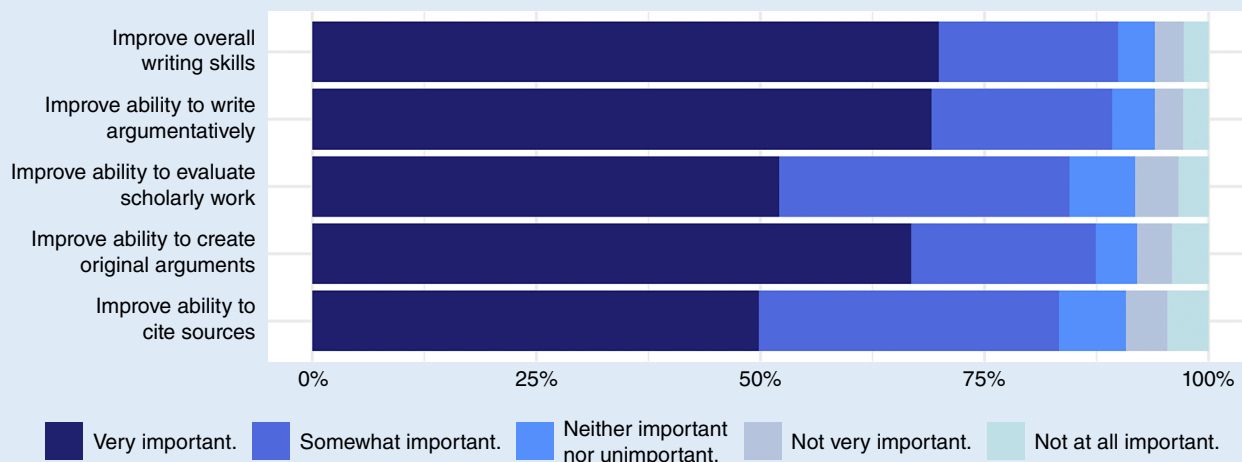
Of the TLC participants surveyed, averages of 62.1% and 45.5% of students' grades were based on take-home written assignments in upper- and lower-level courses, respectively. If instructors considered the use of LLMs to be problematic, take-home written assignments represent clear regulatory challenges. In-class written assignments were not considered effective solutions because they may limit the time that students can spend formulating and executing their ideas.

Educators Are More Pessimistic or Uncertain About AI Tools

In general, colleagues were more pessimistic than optimistic about these tools. On average, respondents to the membership survey rated AI tools at -0.45 on a scale of -2 to 2, with -2 being "very bad" for education and 2 being "very good." The plurality of colleagues surveyed (41.5%) believed that AI tools are "neither good nor bad" for courses (figure 2). Only about 13.6% believed that AI tools are "good" or "very good." These results are not surprising given the current and projected capabilities of these tools and how they might hinder educators' effectiveness in pedagogical goals that they deem important through assigning written work to their students. Results from the smaller
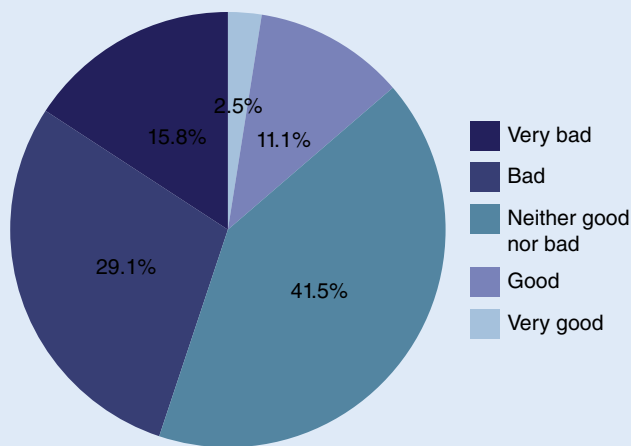
Figure 1
Pedagogical Value of Written Assignments (Membership Survey)

Legend: Very important. | Somewhat important. | Neither important nor unimportant. | Not very important. | Not at all important.

Note: Responses to the question: "Generally speaking, what educational value do you see in written assessments (e.g., research essays, reflection papers) for yourself or your students?"

Figure 2

Responses to the Question: "Overall, Do You Think These AI Tools Are Good or Bad for Your Political Science Classes?" (Membership Survey)

Legend: Very bad | Bad | Neither good nor bad | Good | Very good

(Values shown: 2.5%, 15.8%, 11.1%, 29.1%, 41.5%)

TLC survey were largely consistent with the broader membership survey.

Using an open-ended question, we further probed why respondents thought AI tools were good, bad, or neither good nor bad for political science courses; 562 respondents responded to this question. Several themes emerged: there were concerns about independent thinking and originality (e.g., "It does nothing to teach them how to learn, think critically and independently, and formulate evidence-based views"); a lack of critical engagement (e.g., "They are bad because they provide low-threshold opportunities to avoid engagement with hard and time-consuming learning activities"); academic integrity (e.g.,

"Students use AI tools exclusively to violate academic integrity"); and misinformation and biases in answers generated from LLMs (e.g., "Generative AI tools are inherently extremely biased and will reinforce existing biases along many different dimensions"). Other responses were more optimistic. Some educators perceived these tools as beneficial for tasks such as editing, grammar checking, and formatting (e.g., "They can be very useful for copyediting, catching grammatical mistakes, and improving prose in papers"). Others also recognized that these generative LLMs could help non-native English speakers who have difficulty with language structure (e.g., "They can help non-native English speakers formulate their arguments"). Other respondents were more neutral: they viewed these tools as an inevitability (e.g., "We professors need to incorporate and use AI to combat their inevitable use by students so, in this respect, they are 'neither good nor bad'").

Support for AI Tools Varies Based on Application

Although the instructors' overall opinion of AI tools was relatively negative, it is important to consider their views about the different functions of AI tools. AI tools can be used to generate human-like answers based on the massive corpus of text data in the training set, but they also can be used in an assistive manner to correct grammar and format an essay. Figure 3 shows the average support for various applications of LLMs. The order in which these functions are displayed was randomized.

Respondents in the membership survey were most opposed to using AI tools for writing essays from scratch and answering multiple-choice and fill-in-the-blank questions. They also were against the use of LLMs to generate ideas, outlines, and citations and to provide supporting evidence. Generative LLMs synthesize vast amounts of text data from the Internet and digitized media, which allows them to respond to questions and prompts across various subjects. However, LLMs do not cite sources reliably and often generate random and incorrect information in their responses, a phenomenon known as "hallucination" (Gao et al.

Figure 3
Support for AI Tool Usages (Membership Survey)
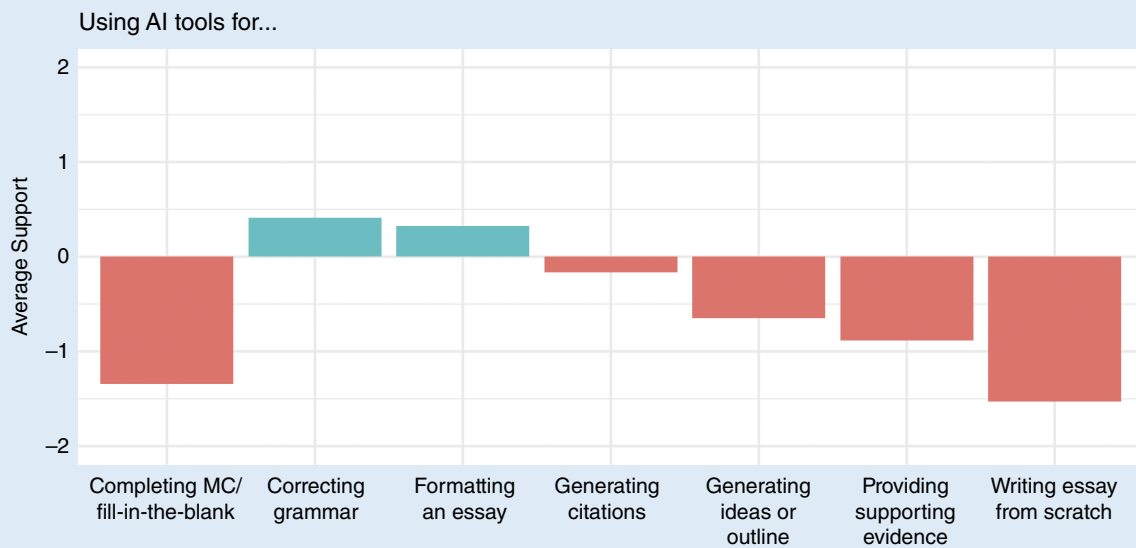
Note: Average responses to the question: "These AI tools can be used to improve and/or complete student essays in a number of ways. Do you favor, oppose, or neither favor nor oppose the following applications for yourself or your students?"

2023). These aspects of generative AI may suggest why some of its applications are considered more problematic.

In contrast, we found support for AI to correct grammar and format essays. Using AI to increase the accuracy of writing or otherwise improve the prose was not perceived to be as problematic as other uses of AI. Such applications are perhaps akin to the usage of more accepted software (e.g., Grammarly) or the use of a university's writing center. The active promotion of AI tools could help English as a Second Language students to improve the quality of their writing and to streamline tasks that do not promote active learning (e.g., harmonizing citation styles). Overall, instructors are supportive of AI when it is used in an assistive as opposed to a generative manner. These attitudes toward AI applications were distributed evenly and did not differ meaningfully among individuals from different institution types and their role in academia (see the online appendix).

response to the increase of LLMs; more than 45% did not plan to change the number of written assignments. Most respondents reported considering running questions through ChatGPT as well as implementing AI-detection systems. This may allow instructors to have a sense of a specific LLM's approach to a topic; however, LLM responses are sensitive to the wording used in the prompt and any guidance that a student might provide (e.g., prompts to cite specific authors). Additionally, as we discuss in our recommendations, AI-detection systems are prone to false positives. Therefore, it is prudent to note that a punitive system based on a problematic detection system may risk increasing faculty burden and stress on the part of students.

Most Recognize the Importance of AI Tools but Are Not Enthusiastic About AI Tools Integration in Courses

Of the TLC respondents, 73% stated that it was at least moderately

...instructors are supportive of AI when used in an assistive as opposed to a generative manner.

Educators Plan to Change Teaching Approaches, but Largely on the Side of Detection and Prevention

More than 60% of the TLC survey respondents believed that 51% or more of their students will use AI to help with their assignments in the next five years. Figure 4 reports support for various response strategies commonly considered.
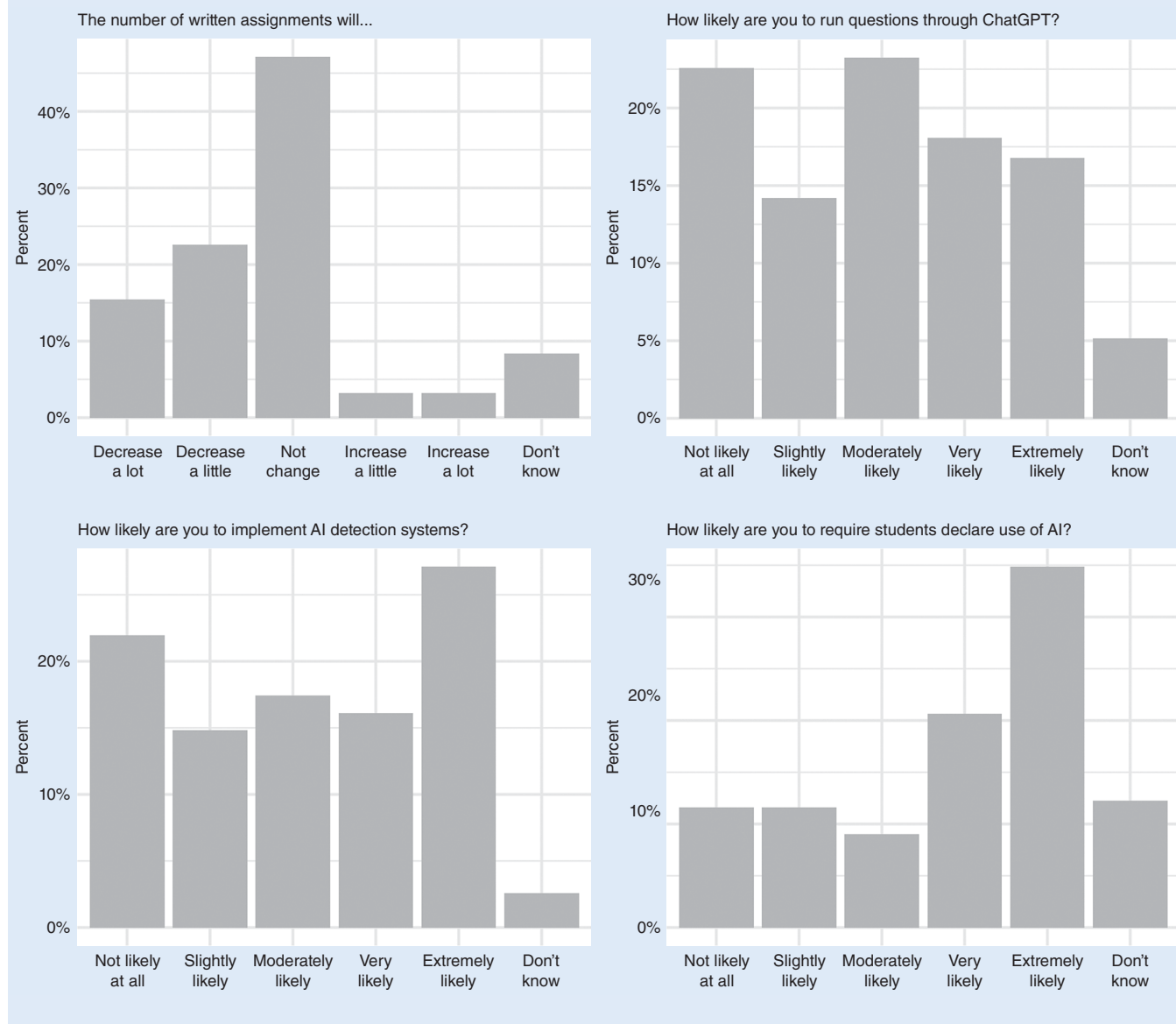
Despite the professed importance of written assignments, more than 30% of TLC survey respondents stated that they are likely to decrease the number of written assignments in

important that students learn how to use generative AI tools. At the same time, more than 50% of TLC respondents said that they either were not likely or only slightly likely to integrate AI tools as part of their courses. Among the categories for this question, more than 30% of respondents chose "not likely at all." Figure 5 presents the full results for the two survey questions.

These outcomes, combined with previous results, suggest that most educators who are skeptical of LLMs plan to address generative AI on the enforcement side.

Figure 4
Planned Changes Made to Teaching Approaches to Address Usage of AI Tools (TLC Survey)

The number of written assignments will...

How likely are you to run questions through ChatGPT?

How likely are you to implement AI detection systems?

How likely are you to require students declare use of AI?

Educators Are Not Always Reliable Adjudicators of Student- and AI-Produced Essays

One main issue with enforcement is that there currently are no accurate means of detecting AI-generated prose. How good or bad are humans at recognizing AI-produced essays? In our TLC
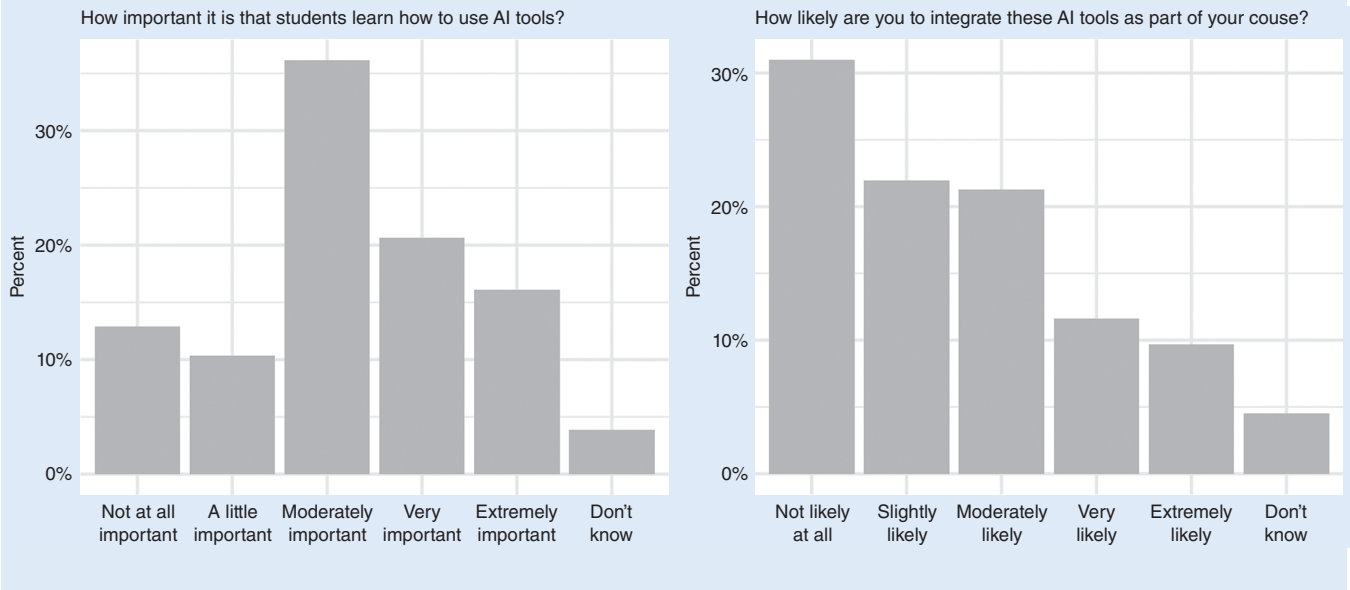
Most recognize the importance of AI tools but are not enthusiastic about AI tools integration in courses.

survey, we investigated this issue by asking respondents to assess two essays drawn from a pool of three topics, which were taken from actual undergraduate class assignments. We asked respondents to determine whether each essay was written by a student or by an LLM and how confident they were with their judgments.

There were a different number of essays for each topic. The first topic, an international relations (IR) question,[1] had two student essays and two GPT-4 essays. The second topic, an American politics (AP) question,[2] had one student essay and two essays written by GPT-4. The third topic, a political economy question,[3]

had two student essays and two essays written by GPT-4. Essays 1 and 3 were more typical, substantive essay topics; essay 2 featured a review of a documentary. Each essay had two GPT-written essays: one was a version produced using the essay question as it was written for the prompt; the other was produced using both the

How important it is that students learn how to use AI tools?

How likely are you to integrate these AI tools as part of your couse?

essay question and substantive knowledge guidance in the prompt.

For each essay, only approximately half of the respondents were able to correctly determine whether the essay was written by GPT-4 or a student; the other half was either incorrect or

politics instructors were correct. It is possible that instructors who have more carefully designed their questions to be "AI-proof" would fare considerably better in a similar exercise in real life. However, the survey suggests that instructors may not always be reliable adjudicators of student- versus AI-produced

Educators are not always reliable adjudicators of student- and AI-produced essays.

could not make a determination (table 1). We find evidence to suggest that domain-specific knowledge may modestly improve the accuracy of determination. For essay 1 (the IR essay), 54.5% of IR instructors were correct. For essay 2 (the AP essay), 52.2% of American politics instructors were correct. For essay 3 (the political economy essay), 50.0% of comparative

essays for conventional topics. Regarding this survey, confidence levels (around 5 out of 10) were approximately equal across the groups making determinations on the first essay. For the second and third essays, the group that correctly determined whether the essay was written by a student or by AI had slightly higher confidence levels. Additional findings are in the online appendix. These findings underscore how difficult it can be to determine whether an essay was written by a student or by GPT-4.

DISCUSSION AND RECOMMENDATIONS

We conducted two surveys to better understand the impact of AI on political science education. Our first survey targeted APSA members. A second longer survey targeted participants of TLC in the past five years. The results largely align with expectations: educators view writing assignments as important and they are pessimistic about recent AI tools; however, they also support the use of AI tools for certain applications (e.g., editing). Although most TLC respondents believed that students will use AI tools to help them with their assignments in the next few years, most do not plan to integrate them as part of their courses. Rather, most changes focus on the detection and prevention of AI-written work. To conclude, we offer a few recommendations that respond to the trends we observed across the two surveys.

Table 1
Results of Assessing Essays as Student-Written or GPT-4 Written

	ESSAY 1	ESSAY 2	ESSAY 3
Correct Determination	52.94% (5.37)	47.52% (5.91)	47.06% (5.94)
Incorrect Determination	30.39% (5.50)	36.63% (4.97)	33.33% (5.32)
Don't Know/ Prefer Not to Answer	16.67%	15.84%	19.61%
N	102	101	102

Note: The numbers in parentheses are confidence levels of the respondents who made a correct or incorrect determination.

Care Must Be Taken When Using AI-Detection Software

A large percentage of respondents stated that they are "extremely likely" to implement AI-detection systems. Detecting AI-generated work is an active area of research (for an overview, see Tang, Chuang, and Hu 2023). However, as of the time of writing, there does not exist reliable AI-detection software (OpenAI 2023). ChatGPT, for example, is unable to determine whether or not something was written by ChatGPT (OpenAI 2023). Tools such as GPTZero claim to accurately detect AI- versus human-written work. However, these tools have been shown in case studies to suffer from high false positive rates (Edwards 2023; Fowler 2023). A student being falsely accused of cheating can face long-term consequences, raising the stakes of using such tools to make assessments of written work. Studies also have shown that AI-detection software can be fairly easy to evade. Taloni, Scorcia, and Giannaccare (2024) demonstrate that slightly modifying abstracts can dramatically increase false negative rates using existing AI-detection software. OpenAI also debuted an AI classifier for indicating AI-written work, but it quietly deprecated the software after only eight months due to its low accuracy rate (Kirchner et al. 2023).

Declaration Statements to Collect Data and Encourage Transparency

For those colleagues who want to ban AI tools, the use of a declaration statement may not deter the use of AI. There is not much empirical support that declarations of honesty alter behavior (Kristal et al. 2020). However, declaration statements can set expectations for students for using generative LLMs (e.g., what is and is not allowed) and enable educators to collect data on how tools are used by their students on assignments. This helps instructors to stay current and be apprised of the creative ways in which some students are using AI tools. These statements may combine a descriptive paragraph with a response to a multiple-choice question on how AI was used for quick processing.

Redesigning Course Components

Realistically, redesigning materials to respond to emerging AI technologies will require trial and error. For those instructors who want to limit students' use of AI tools, colleagues have proposed alternative activities (other than essay writing) that promote creativity and critical thinking before the advent of LLMs, which might be of additional value today. Pedagogical innovations and interventions including group concept mapping (Wilson et al. 2023), creating memes with accompanying essays explaining the thinking behind the meme (Wells 2018), reflections on in-class simulations and games (Handby 2021), and a "Learning Record" that includes notes submission and an open-ended final project (Lawton and Kenner 2023) are not conducive to AI-produced outputs. Thus, students may be less inclined to use AI tools to complete these assignments.

However, most colleagues surveyed indicated that they believe most students will use AI tools and that it is moderately important that they learn how to use them. Some instructors may want to integrate these tools into their courses. We summarized the ideas shared by colleagues in the surveys as follows. Some take a more liberal stance and allow students to use AI tools as long as proper citation is provided. Another tasks students to use generative LLMs to "resolve complex technical issues for a data-

visualization assignment." Another colleague imagines a real-world scenario in which a student—in the hypothetical role of an advocacy or political group staffer—generates a first draft using ChatGPT and annotates it for accuracy, depth, and relevance. Others will use scaffolding strategies and move toward a more iterative model of writing—that is, assigning fewer assignments but with more components and/or drafts—to be more explicit about the skills or knowledge that the individual part seeks to develop (while also recognizing that LLMs might be capable of finishing these components). Ultimately, many respondents (and we) believe that it is beneficial to have an open dialogue with students about how these new tools will help or hurt their skill development.

CONCLUSION

Recent applications of AI will alter teaching, learning, and academic assessment. Instructors who want to ban AI tools may find themselves unable to do so because there are no credible ways of detection. Punitive measures based on problematic detection systems may risk increasing stress and burden for everyone involved. Whereas declaration statements will not stop the unauthorized use of AI tools, they may encourage transparency and operate as a data-collection mechanism while instructors learn about how students use AI tools. Those who want to take a more proactive role in this changing educational landscape may take steps to integrate generative LLMs into the pipeline of writing assignments or replace or supplement traditional academic essay writing with innovative assessments. Our discipline would benefit from a continuous discussion of the opportunities and pitfalls of AI tools in higher education as these technologies become more prevalent in the workplace and in our society.

NOTES

1. Evaluate the statement: "State weakness creates opportunities for states or rebel groups to start military conflicts. This explains the occurrence of the WWII and separatist conflicts in Chechnya."

2. Write a review of the documentary "American Factory."

3. Evaluate the statement: "Economic incentives for investments do more harm than good." Use examples as necessary. Name two political explanations for corporate welfare. In other words, explain the political motivations to provide subsidies or tax breaks.

REFERENCES

Brown, Tom, Benjamin Mann, Nick Ryder, et al. 2020. "Language Models Are Few-Shot Learners." In *Advances in Neural Information Processing Systems*, ed. Hugo Larochelle, et al., 1877–901. New York: Curran Associates, Inc.

Edwards, Benj. 2023. *Why AI Detectors Think the US Constitution Was Written by AI.* https://arstechnica.com/information-technology/2023/07/why-ai-detectors-think-the-us-constitution-was-written-by-ai.

Fowler, Geoffrey A. 2023. "What to Do When You're Accused of AI Cheating." *Washington Post*, August 14. www.washingtonpost.com/technology/2023/08/14/prove-false-positive-ai-detection-turnitin-gptzero.

Gao, Tianyu, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. *Enabling Large Language Models to Generate Text with Citations.* arXiv:2305.14627 [cs.CL]. https://arxiv.org/abs/2305.14627.

Handby, Edmund. 2021. "Classroom Games to Teach Contemporary Political Theory." *Journal of Political Science Education* 17 (Sup1): 23–31.

Kasneci, Enkelejda, Kathrin Sessler, Stefan Küchemann, et al. 2023. "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education." *Learning and Individual Differences* 103:102274.

Kirchner, Jan Hendrik, Lama Ahmad, Scott Aaronson, and Jan Leike. 2023. *New AI Classifier for Indicating AI-Written Text.* https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

Kristal, Ariella S., Ashley V. Whillans, Max H. Bazerman, and Dan Ariely. 2020. "Signing at the Beginning Versus at the End Does Not Decrease Dishonesty." *Proceedings of the National Academy of Sciences* 117 (13): 7103–107.

Lawton, Sally F., and Alison Kenner. 2023. *The Learning Record: Portfolio-Style Assessment in Political Science Courses.* Los Angeles: Teaching and Learning Conference at APSA, September 3.

OpenAI. 2023. *Teaching with AI.* August 31. https://openai.com/blog/teaching-with-ai.

Taloni, Andrea, Vincenzo Scorcia, and Giuseppe Giannaccare. 2024. "Modern Threats in Academia: Evaluating Plagiarism and Artificial Intelligence Detection Scores of ChatGPT." *Eye* 38 (2): 397–400. DOI:10.1038/s41433-023-02678-7.

Tang, Ruixiang, Yu-Neng Chuang, and Xia Hu. 2023. *The Science of Detecting LLM-Generated Texts.* arXiv:2303.07205 [cs.CL].

Tlili, Ahmed, Boulus Shehata, Michael Agyemang Adarkwah, et al. 2023. "What If the Devil Is My Guardian Angel: ChatGPT as a Case Study of Using Chatbots in Education." *Smart Learning Environments* 10 (1): 15.

Wells, Dominic D. 2018. "You All Made Dank Memes: Using Internet Memes to Promote Critical Thinking." *Journal of Political Science Education* 14 (2): 240–48.

Wilson, Matthew, Christopher Howell, Kelsey Martin-Morales, and Sanghoon Park. 2023. "Concept Mapping and Reading Comprehension." *Journal of Political Science Education*:1–24. https://doi.org/10.1080/15512169.2023.2164861.

Wu, Nicole, and Patrick Y. Wu. 2024. "Replication Data for 'Surveying the Impact of Generative Artificial Intelligence on Political Science Education.'" *PS: Political Science & Politics.* DOI:10.7910/DVN/FNZQ06.

Zhao, Wayne Xin, Kun Zhou, Junyi Li, et al. 2023. *A Survey of Large Language Models.* arXiv:2303 .18223 [cs.CL].