AI-written admissions essays are widespread but penalized

Calvin Isley
Harvard University

Johann D. Gaebler
New York University

Sharad Goel
Harvard University

Abstract

AI is rapidly transforming higher education, including the application process, yet relatively little is known about its use and consequences. To help close this gap, we analyze nearly 8,000 applications submitted between 2020 and 2025 to a large public policy master's program in the United States. We find that in the 2025 admissions cycle, the majority of applicants submitted at least one essay that was primarily AI-generated—despite an explicit prohibition against using AI. Leveraging the abrupt introduction of ChatGPT in November 2022, we find that the availability of AI assistants improved the writing quality of submitted essays. These improvements, however, came with an apparent AI penalty: applicants submitting AI-written essays were admitted less often than comparable non-users. To help explain this penalty, we conduct an experiment with admissions officers, finding that they can often recognize AI writing and rate essays they believe to be AI generated lower than essays they believe to be human generated. These findings indicate that AI is changing both how applicants write and how that writing is evaluated, raising questions about whether admissions practices and policies designed for a pre-AI era remain appropriate.

Acknowledgments We thank the admissions staff at our partner institution for data access and guidance, GPTZero for providing free API access, and Alex Adam for helpful conversations.

1 Introduction

When ChatGPT was released in November 2022 [1], it fundamentally altered the way people write. AI-generated writing has since proliferated across a variety of domains, ranging from academic journal submissions, to job applications, to UN press briefings [2–6]. University admissions essays have been no exception [7–9]. Historically, admissions essays have been a channel for applicants to demonstrate their writing ability, motivation, and program fit, beyond grades and test scores [10–12]. Now, however, university admissions offices must contend with the possibility that applicants' submitted essays do not reflect their abilities or authentic voice [7]. Recent survey evidence suggests that AI use may already be widespread, with 30% of college applicants reporting using generative AI to help write their personal essays, 20% of whom reported using AI to produce a final draft [13]. Similarly, Lee et al. [9] estimated that approximately 8–10% of the text in Common Application essays submitted to a competitive engineering school exhibited linguistic patterns characteristic of LLM-generated content in the year after ChatGPT's release.

Despite the speed, scale, and significance of this change, few universities provide formal guidelines on the use of AI in applications [14], perhaps reflecting a lack of evidence on the prevalence and consequences of AI usage in admissions. Recent work has considered the detectability of AI in admissions materials [15–18], the stylistic properties of AI-generated admissions essays [8, 9, 19], and exploratory analysis on the impacts of AI use on admissions outcomes [9]. But it remains unclear how often applicants are using AI, how AI impacts the quality of their applications, and whether AI use affects applicants' admission decisions. Without answers to these questions, admissions offices may struggle to set and enforce sensible AI policies. This lack of clarity leaves applicants to make individual choices on the appropriate use of AI, potentially exacerbating disparities in the admissions process.

Here we begin to close this information gap by analyzing a novel dataset of nearly 8,000 applications to a large American public policy master's program submitted between 2020 and 2025—covering three admissions cycles before and three after the release of ChatGPT. Using commercial AI detection tools, we find that AI usage is widespread and growing, with 56% of applicants submitting at least one essay that was likely primarily AI-generated in the most recent application cycle. Comparing pre- and post-ChatGPT essays, we find that the broad availability of AI substantially improved the quality of applicants' essays, particularly for international applicants.

Despite these improvements in essay quality, applicants who submitted AI-written essays were nevertheless admitted at lower rates than otherwise similar applicants who did not use AI. To understand the source of this apparent penalty, we conduct an experiment to measure admissions officers' ability to detect AI-written essays. We find that not only can admissions officers reliably discriminate between AI and human-written text, but that they also rate essays they suspect to be AI-written more harshly. Together, our findings suggest that AI use is both common and consequential, at least among the applicants we study, but likely more broadly. These results underscore the need for universities to reevaluate the admissions process in response to this new reality, and to develop sensible and enforceable AI policies for it.

2 Data

To study the role of AI writing in admissions, we assemble a novel dataset of 7,462 applications to a competitive U.S. public policy master's program submitted across six admissions cycles between 2020 and 2025. In addition to five admissions essays, prospective students provide basic

demographic information, GRE or GMAT scores,[1] academic transcripts from previous undergraduate or graduate institutions, and three letters of recommendation from university faculty members or immediate professional supervisors. The essays ranged in length from 250–500 words and cover topics ranging from applicants' motivation for applying to the program, to their personal and professional background, to their plans to make an impact through public service.[2] Applicants who did not attend an English-language undergraduate institution were also required to submit TOEFL, IELTS or Cambridge English scores. These materials comprise the totality of information about applicants available to admissions officers during their decision-making process.

To facilitate statistical analysis, we designed an LLM-based pipeline to convert the unstructured application materials to structured measures [20, 21]. These measures include, for example: length of prior public service or non-profit work experience; performance in previous program-relevant coursework, standardized to 4.0 scale; the number and severity of grammatical errors in essays; and the strength of endorsement in submitted recommendation letters. In total, we extract 216 such measures from the unstructured materials. These measures were designed in collaboration with admissions officers in the program we analyze, with the aim of capturing the key information informing admissions decisions. (For full details, see Appendix Section A.1.)

3 Results

For all six admissions cycles in our data, the program we study required applications to be completed personally and in English. In particular, before final submission, applicants signed an attestation that they did not receive prohibited writing or translation assistance. Starting in 2023—the first cycle after ChatGPT was released—this attestation was expanded to exclude use of generative AI.

3.1 AI-Written essays are common

We use a popular commercial AI-detection tool, GPTZero [22], to identify AI-generated essays. GPTZero and other tools like it work by identifying subtle stylistic patterns in text to determine its origin. GPTZero classifies documents into one of three categories: (1) "AI written," meaning documents primarily produced by AI, including human-written text that was rewritten by AI; (2) "human," meaning documents written exclusively by humans; and (3) "mixed," meaning documents containing both human-written and AI-generated text. GPTZero first estimates the probability that text comes from each of these three categories, and then assigns documents to the highest-probability category. It further assigns confidence scores to the classifications indicating their level of certainty. For our primary analysis, we use its "AI written" classification with "high" confidence, suggesting that flagged essays were primarily generated by AI—and we throughout refer to such essays as "AI-generated." We replicate our primary results with Pangram, as well as with looser AI-use classification thresholds (e.g., including documents classified as "mixed"), finding qualitatively similar results. See Appendix E for details.

For each of the six application cycles we study, Figure 1 shows the proportion of applicants flagged as submitting at least one AI-generated essay. In the three cycles prior to the release of ChatGPT in late 2022, virtually no essays were identified as AI-written. That pattern suggests the detector has a low false-positive rate, consistent with the company's stated error rate as well as with estimates from independent audits [22, 23].

[1] Applicants were not required to submit GRE or GMAT scores if they applied in 2020 or 2021.

[2] The prompts for the required essays were largely the same throughout our sample period, barring a substantive change to one essay's prompt in 2023. Our results persist regardless of whether we include or exclude that essay. See Appendix Section E for discussion of the modified results.

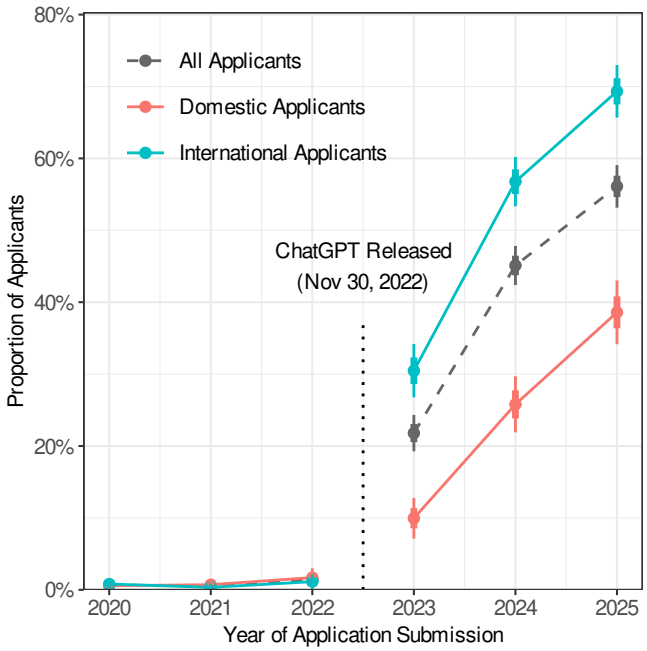

Figure 1: *Proportion of applicants who submitted one or more essays detected by GPTZero as written by AI between 2020 and 2025. Solid lines are disaggregated by nationality, the dashed line aggregates across all applicants. ChatGPT was released on Nov. 30, 2022 [1], one day before the December 1 application deadline for the 2022 cycle, suggesting few, if any, applications were written with AI by applicants that year.*

Beginning in 2023, the proportion of applicants submitting at least one AI-written essay flagged climbs by around 20 p.p. per year on average, to nearly 60% in the most recent cycle. We find that AI usage is also considerably higher for international applicants (69.3%) than for domestic applicants (38.6%).[3] We further find that AI-use is higher among younger applicants, but is largely uncorrelated with performance on standardized test scores and undergraduate major (see Appendix Fig. A1).

Though already relatively high, our estimate likely understates the true prevalence of AI use. AI detectors are typically calibrated to ensure few human-written responses are erroneously flagged as AI-generated, consistent with the low false-positive rate we see in the pre-ChatGPT years in Figure 1. As a result, however, these detectors could have relatively high false-negative rates, mistakenly classifying AI-generated text as human. Further, to aid interpretation, we have focused on text that was likely primarily generated by AI. As a result, our counts miss instances where applicants use AI in part without submitting fully AI-generated text, behavior that is still prohibited by the program we consider.

3.2 AI improves essay quality

We next consider the impact of AI on the quality of submitted essays. To do so, we develop an automated procedure to evaluate essays along seven dimensions: grammar, style, clarity, prompt

[3]In contrast to prior work on the quality of AI detection software, we find no evidence of a differential false positive rate for international applicants or for TOEFL and IELTS takers [24]. The usage gap between domestic and international applicants is larger than the gap between applicants who submit language test scores and those who do not.

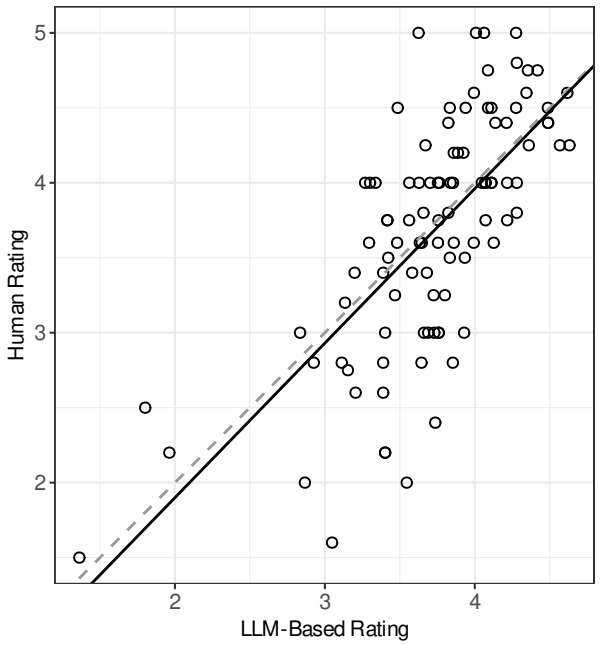

Figure 2: *Comparison of expert and LLM-based ratings of essay quality. For each essay, its expert rating is the average rating given by admissions officers, on a six-point scale; the LLM-based rating is a weighted combination of seven dimensions of essay quality, scored by an LLM. Higher values indicate higher quality. The solid line is the line of best fit, and the dashed line indicates equality between human and model scores. The two rating methods have a correlation of 70%.*

responsiveness, voice, commitment to public service, and demonstrated leadership. We instruct an LLM to grade each dimension on a five-point scale designed to minimize interpretive ambiguity. For example, each essay's grammar score reflects the number and severity of errors, with 5 points awarded for 0-1 surface errors (e.g., spelling, punctuation, and subject-verb misalignment), and 1 point for either 11+ such errors or multiple meaning-blocking errors. To construct a single, composite score of essay quality, we then had five admissions officers from the partner program each rate up to 100 submitted essays—50 that were likely generated by humans and 50 by AI—on a six-point scale, from "much weaker" to "much stronger" than a typical essay. Finally, we fit a model to predict these expert ratings from the LLM-scored measures, yielding weights for each of the seven dimensions of essay quality we consider—and, accordingly, a composite essay quality score. Figure 2 shows that the automated and expert ratings are in relatively close agreement, with a correlation of 70%. Further details are provided in Appendix Sections B and C.

We apply our automated evaluation procedure to infer overall and dimension-specific quality scores for each essay in our dataset. Then, for each applicant, we average the scores of their submitted essays to produce applicant-level measures of writing quality. Figure 3 shows a sharp improvement in the overall quality of essays (blue line) after the introduction of ChatGPT, particularly for international applicants and especially in the most recent cycles, when use of AI was greatest. These gains come from large improvements in mechanical aspects of writing (grammar, style, and clarity, indicated by the red line), with more modest improvement along substantive dimensions (prompt responsiveness, voice, commitment to public service, and demonstrated leadership, indicated by the green line). Further, these overall and dimension-specific gains in essay quality persist when we include controls for applicant quality, suggesting these changes are not driven by changes to the

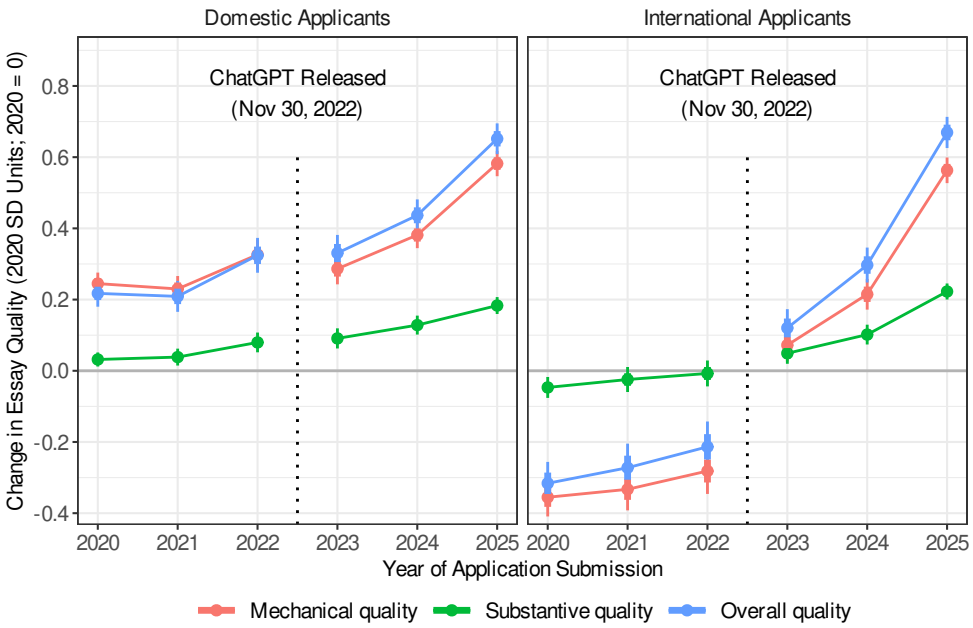

Figure 3: *Change in human-calibrated and LLM-assessed essay quality over time, disaggregated by residency status, in 2020 standard deviation units. LLM quality ratings are aggregated into substantive quality and mechanical quality. Each series is independently standardized to the mean and standard deviation of the full applicant pool applying in 2020.*

composition of the applicant pool. (See *SI Appendix* Section C and Fig. A4.)

3.3 AI use is penalized

The widespread availability of AI meaningfully improved the quality of submitted essays, raising the possibility that applicants who use it gain an advantage over those who do not. To investigate this possibility, we next estimate the impact of submitting AI-written essays on admissions decisions.

In the 2023 and 2024 admissions cycles, applicants with at least one essay flagged as AI-written were admitted substantially less often than those without any such essays.[4] But applicants who use AI may differ from those who do not in a variety of ways, complicating a causal interpretation of the different admissions rates. By leveraging full application files reviewed by admissions officers, we can account for a wide set of potential confounders, though we cannot completely eliminate the possibility of omitted-variable bias. To adjust our estimates for observable differences between applicants, we apply double machine learning (DML) [25], a popular method for estimating causal effects with flexible models that incorporate a large number of covariates—around 400 in our case.

The results are shown in Table 1. Each column reports the estimated percentage-point change in admission probability for each essay flagged as AI-generated submitted by an applicant. The first column reports simple OLS estimates, without adjusting for any applicant characteristics. Our main DML estimates are given in the second and third columns. In the first row of the second column, we estimate a 1.5 p.p. decrease in admission probability for each submitted essay that was AI-written,

[4]In the 2025 admissions cycle, after seeing our results, the program we analyze overhauled its admissions process. Given the potential influence of our results on the decision-making process, we do not examine outcomes for that most recent cycle.

Table 1: AI Usage and Admission

	OLS	DML	
	(1)	(2)	(3)
AI Count	-0.077***	-0.015**	-0.026***
	(0.007)	(0.006)	(0.006)
AI Count (2023)	-0.081***	-0.015	-0.021
	(0.013)	(0.012)	(0.013)
AI Count (2024)	-0.076***	-0.013*	-0.021**
	(0.008)	(0.006)	(0.007)
Applicant controls		✓	✓
Essay quality			✓

Note: Standard errors are reported in parentheses. The first row reports estimates from models without year interactions; the final two rows report estimates from separate year-specific models. Column (1) reports unadjusted OLS estimates from linear probability models. Columns (2) and (3) report DML estimates using random forest nuisance functions and 5-fold cross-fitting. Across 20 independently seeded fits, coefficients and standard errors are aggregated using DoubleML's median-based repeated-splitting procedure [25]. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

after controlling for our rich set of covariates, but excluding measures of essay quality because of potential post-treatment bias. The rows below show the results from two models that separately estimate the effect for each of the two cycles. The estimated effects are negative in both cycles but more precisely estimated in 2024, likely because AI use was less common in 2023; consequently, the year-specific estimate is not statistically significant in 2023. Combining both cycles, applicants who used AI submitted, on average, 2.32 AI-written essays, corresponding to an aggregate 3.48 p.p. decrease in admission probability.

In the third column of Table 1, we estimate the effect of AI use on admissions decisions after additionally adjusting for the dimensions of essay quality that we constructed above—in addition to the myriad application features we previously adjusted for. This model specification aims to isolate the penalty for AI use itself, accounting for substantive differences in the AI-written and non-AI-written essays, though it may be subject to post-treatment bias. In this case, the estimated effect increases to a penalty of 2.6 p.p. for each AI-written essay, or about 6.0 p.p. for the typical applicant who used AI. That increase is expected, since AI-written essays are, on average, stronger, and stronger essays themselves improve an applicant's odds of admission. As a result, not accounting for essay strength masks some of the penalty from AI use.

The apparent AI penalty might be expected given that AI use was explicitly prohibited by the program we study. But the admissions office we study did not systematically attempt to detect, adjudicate, or penalize violations of the policy. In particular, submitted essays were not screened with any AI detection tools. Readers might have informally downgraded applicants who used AI, though doing so would require readers to identify AI-written essays.

To understand how well admissions officers can identify AI use in application essays, during the essay quality calibration procedure described above, we additionally asked admissions staff to assess the likelihood that essays were written by a human or by AI, on a six-point scale from "very likely human" to "very likely AI". (Further details are provided in Appendix Section B.) Admissions staff achieved an AUC of 0.70 (95% CI: $[0.65, 0.75]$) on this detection task, meaningfully above chance though far below the accuracy achieved by commercial detectors.

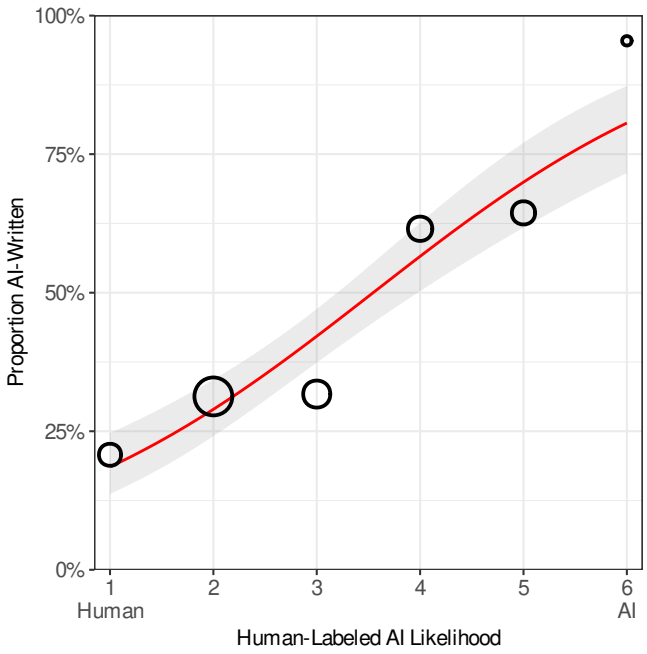

Figure 4: *Proportion of essays that are AI-generated by admissions officers' perceived likelihood of AI authorship. Point size reflects number of essays in each response category. The red line is a smoothed fit with 95% confidence bands. The AI-likelihood scale ranges from 1 ("very likely human") to 6 ("very likely AI"), with intermediate values reflecting increasing likelihood of AI authorship.*

Figure 4 provides further detail, plotting the relationship between admissions officers' perceived likelihood of AI authorship against the essay's machine-inferred provenance. We find that the human judgments are directionally informative but imprecise. For example, among essays rated on the human side of the scale (1-3), 70.4% were classified as human-generated by GPTZero, and among those rated on the AI side, 67.8% were classified as AI-generated by GPTZero. It thus seems that admissions officers, while imperfect, are able to identify AI writing reasonably well.

Finally, we consider whether admissions officers assign lower scores to essays they perceive to be AI-generated, a pattern that would be consistent with past work suggesting that AI-writing is penalized in both admissions and more broadly [9, 26]. The black points in Figure 5 show the average rating assigned to essays as a function of perceived likelihood that it was AI-generated. Essays deemed by admissions staff to be "very likely" AI-generated receive markedly lower scores. These essays were also rated lower by our automated evaluation procedure, shown by the red points—though not as low as by admissions staff. More generally, it is possible that perception of AI correlates with substantive dimensions of essay quality, and those aspects, not perceived AI use, drive the lower ratings. In Table 2, we estimate staff ratings after adjusting for a variety of covariates, including our machine-generated measures of essay quality, whether the essay was in reality AI-generated, and basic demographic and academic information about the applicant. After this adjustment, we still find that staff ratings are negatively correlated with perceived AI use, providing further evidence that AI use itself was penalized by admissions staff. In sum, the lower observed acceptance rate for applicants who use AI is consistent with admission officers detecting and penalizing AI writing. For more detailed discussion of our experiment and reports of individual rater performance, see Appendix Section B and Figs. A2 and A3.

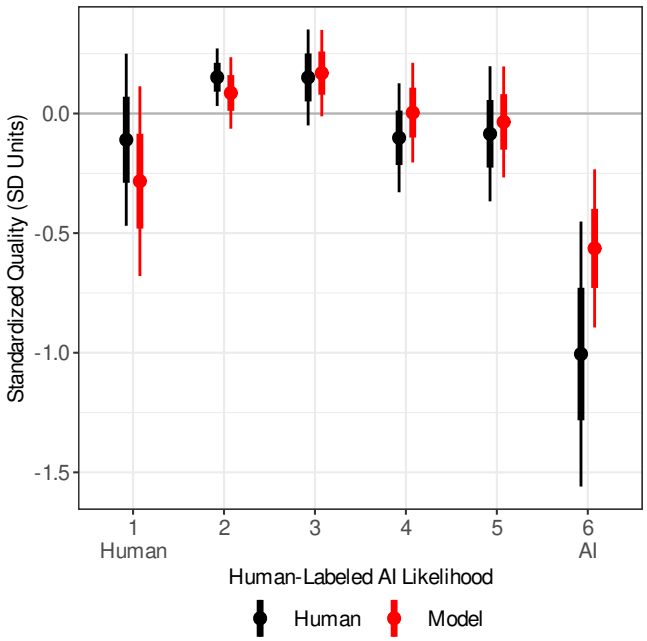

Figure 5: *Normalized human and human-calibrated model essay quality scores as a function of admissions officer perception of AI usage. The AI-likelihood scale ranges from 1 ("very likely human") to 6 ("very likely AI"), with intermediate values reflecting increasing likelihood of AI authorship.*

4 Discussion

Drawing on nearly 8,000 applications across six admissions cycles, we document a dramatic shift in applicant behavior. By 2025—just three years after the introduction of ChatGPT—a majority of applicants submitted at least one essay written primarily by AI, despite explicit prohibitions on its use. Among international applicants, the rate approached nearly 70%. We further find that AI use improved the overall quality of submitted essays, particularly the more mechanical aspects of writing. Yet admissions staff could often recognize AI-written essays and appear to have penalized them.

Our findings expose a central tension for university admissions. AI can improve the quality of submitted essays while weakening the connection between those essays and applicants' unaided writing ability—and potentially subjecting applicants to idiosyncratic penalties for using AI. Even before AI, application essays did not necessarily reflect an applicant's unaided effort, as students frequently turned to mentors, teachers, and others for help. But AI assistants have likely widened the gap between an applicant's individual writing ability and the quality of the final submitted product. At the same time, by reducing the burden of writing, AI may help applicants communicate their ideas more clearly, even as uncritical use of these tools may also distort or displace those ideas. Ultimately, universities must clarify what they seek to learn from application essays and reconsider whether the purposes those essays have historically served remain appropriate in a world where AI-assisted writing is becoming standard professional practice.

These results and their implications should be considered in light of several limitations. First, our analysis is limited to applications to a single graduate program in public policy, and the patterns we document may differ, potentially substantially, across contexts. For example, in settings where

Table 2: AI-Likeness and Human Quality Ratings

	Human Quality Rating		
	(1)	(2)	(3)
AI Likelihood	-0.131*	-0.119*	-0.109*
	(0.059)	(0.047)	(0.047)
LLM-Based Quality		1.029***	0.995***
		(0.090)	(0.093)
AI-Written		0.019	-0.033
		(0.101)	(0.102)
Applicant controls			✓
Reviewer fixed effects	✓	✓	✓

Note: Standard errors are reported in parentheses and clustered by applicant. Applicant controls in column (3) include undergraduate GPA, quantitative and verbal GRE scores, residency status, gender, age, and an indicator for whether the applicant submitted a TOEFL score. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

AI use is not expressly prohibited, its prevalence may be even higher and its consequences less severe. Second, our estimates of AI use rely on commercial detectors calibrated to limit false positives, and are thus likely conservative. Moreover, to facilitate interpretation, we consider as "AI generated" only essays that were likely primarily written by AI, excluding less extensive uses of AI that were also prohibited. Third, our automated measure of essay quality is an imperfect proxy for expert judgment. Although it generally tracks ratings by admissions staff, it may systematically miss dimensions of quality that matter in admissions. Finally, our estimate of the apparent penalty associated with AI use may be affected by unmeasured confounding. Our double machine learning approach adjusts for hundreds of covariates derived from submitted application materials, but we cannot rule out omitted factors associated with both AI use and admissions decisions.

Looking ahead, our findings highlight the need for institutions to move toward deliberate, transparent policies around AI use. Many institutions have yet to establish policies governing AI use in admissions, and even where formal restrictions exist, the boundaries of permissible AI assistance may be unclear and enforcement inconsistent. Our empirical results, on their own, do not determine what role, if any, AI-assisted writing should play in a well-designed admissions process. Rather, by providing systematic evidence on the prevalence and consequences of AI use, we hope to help institutions make more informed decisions about what they seek to learn from applicants' written materials and how those materials should be evaluated in an increasingly AI-mediated world.

References

[1] OpenAI. Introducing ChatGPT, March 2022. URL `https://openai.com/index/chatgpt/`.

[2] Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, and James Zou. Quantifying large language model usage in scientific papers. *Nature Human Behaviour*, 9(12):2599–2609, December 2025. doi: 10.1038/s41562-025-02273-8. URL `https://www.nature.com/articles/s41562-025-02273-8`.

[3] Sarah Kessler. Employers Are Buried in A.I.-Generated Resumes. *The New York*

Times, June 2025. URL https://www.nytimes.com/2025/06/21/business/dealbook/ai-job-applications.html.

[4] Dirk HR Spennemann. Delving into: the quantification of Ai-generated content on the internet (synthetic data), March 2025. URL http://arxiv.org/abs/2504.08755. arXiv:2504.08755 [cs.IR].

[5] Jenna Russell, Marzena Karpinska, Destiny Akinode, Katherine Thai, Bradley Emi, Max Spero, and Mohit Iyyer. AI use in American newspapers is widespread, uneven, and rarely disclosed, April 2026. URL http://arxiv.org/abs/2510.18774. arXiv:2510.18774 [cs.CL].

[6] Claudine Gartenberg, Sharique Hasan, Alex Murray, and Lamar Pierce. More versus better: Artificial intelligence, incentives, and the emerging crisis in peer review. *Organization Science*, 37(3):795–812, 2026. doi: 10.1287/orsc.2026.ed.v37.n3. URL https://doi.org/10.1287/orsc.2026.ed.v37.n3.

[7] Natasha Singer. Ban or Embrace? Colleges Wrestle With A.I.-Generated Admissions Essays. *The New York Times*, September 2023. URL https://www.nytimes.com/2023/09/01/business/college-admissions-essay-ai-chatbots.html.

[8] Kibum Moon, Kostadin Kushlev, Andrew Bank, Benjamin L Luttges, Indre Viskontas, James C Kaufman, Dan R Johnson, Angela L Duckworth, and Adam Green. The Creative Link Between Words and Ideas is Weakening in the AI Era, February 2026. URL https://osf.io/preprints/psyarxiv/jsz58_v6/.

[9] Jinsook Lee, Conrad Borchers, AJ Alvero, Thorsten Joachims, and Rene F. Kizilcec. The digital divide in generative AI: Evidence from large language model use in college admissions essays. In *Proceedings of the Thirteenth ACM Conference on Learning @ Scale*, L@S '26, page 179–190, New York, NY, USA, 2026. Association for Computing Machinery. doi: 10.1145/3774398.3811620. URL https://doi.org/10.1145/3774398.3811620.

[10] Kelly Mae Ross and Josh Moody. What Admissions Officers Think of 3 Common College Essay Topics, February 2020. URL https://www.usnews.com/education/best-colleges/articles/2018-07-09/what-admissions-officers-think-of-3-common-college-essay-topics.

[11] Michael N. Bastedo, Nicholas A. Bowman, Kristen M. Glasener, and Jandi L. Kelly. What are We Talking About When We Talk About Holistic Review? Selective College Admissions and its Effects on Low-SES Students. *The Journal of Higher Education*, 89(5):782–805, September 2018. doi: 10.1080/00221546.2018.1442633. URL https://doi.org/10.1080/00221546.2018.1442633.

[12] Gretchen W. Rigol. Admissions Decision-Making Models: How U.S. Institutions of Higher Education Select Undergraduate Students. Technical report, College Entrance Examination Board, 2003. URL https://eric.ed.gov/?id=ED562589. ERIC Number: ED562589.

[13] Jennifer Rubin, Ella Lombard, J., Katharine Chen, and Riddhi Divanji. Navigating College Applications with AI How High School Teachers and Students Use Tools Like ChatGPT[White Paper]. Technical report, foundry10, July 2024. URL https://s3.us-west-2.amazonaws.com/craft-resource-files/foundry10_Navigating_College_Admissions_AI_White_Paper.pdf.

[14] Kaplan. Kaplan Survey: Colleges Clarify Rules on AI Use in Admissions Essays, November 2025. URL https://kaplan.com/about/press-media/college-admissions-officers-survey-2025-essays-ai.

[15] Yijun Zhao, Alexander Borelli, Fernando Martinez, Haoran Xue, and Gary M. Weiss. Admissions in the age of AI: detecting AI-generated application materials in higher education. *Scientific Reports*, 14(1):26411, November 2024. doi: 10.1038/s41598-024-77847-z. URL https://www.nature.com/articles/s41598-024-77847-z.

[16] Hugh Burke, Rebecca Kazinka, Raghu Gandhi, and Aimee Murray. Artificial Intelligence-Generated Writing in the ERAS Personal Statement: An Emerging Quandary for Post-graduate Medical Education. *Academic Psychiatry: The Journal of the American Association of Directors of Psychiatric Residency Training and the Association for Academic Psychiatry*, 49(1):13–17, February 2025. doi: 10.1007/s40596-024-02080-9.

[17] Nicole Cumbo, Whitney Williams, Joseph C Canterino, Noelle Aikman, and Jonathan D Baum. Can Residency Programs Detect Artificial Intelligence Use in Personal Statements? *Cureus*, 17(7):e88969, 2025. doi: 10.7759/cureus.88969. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12392696/.

[18] Zachary C. Lum, Lohitha Guntupalli, Augustine M. Saiz, Holly Leshikar, Hai V. Le, John P. Meehan, and Eric G. Huish. Can Artificial Intelligence Fool Residency Selection Committees? Analysis of Personal Statements by Real Applicants and Generative AI, a Randomized, Single-Blind Multicenter Study. *JBJS Open Access*, 9(4):e24.00028, October 2024. doi: 10.2106/JBJS.OA.24.00028. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11498924/.

[19] Jinsook Lee, AJ Alvero, Thorsten Joachims, and René Kizilcec. Poor alignment and steerability of large language models: Evidence from college admission essays, 2026. URL https://arxiv.org/abs/2503.20062.

[20] Johann D Gaebler, Calvin Isley, Christopher Avery, and Sharad Goel. Reassessing the role of standardized tests in university admissions. Working Paper 35407, National Bureau of Economic Research, July 2026. URL http://www.nber.org/papers/w35407.

[21] Calvin Isley, Johann D. Gaebler, and Sharad Goel. Mitigating label bias with interpretable rubric embeddings, 2026. URL https://arxiv.org/abs/2605.21455.

[22] Edward Tian and Alexander Cui. Gptzero: Towards detection of ai-generated text using zero-shot and supervised methods, 2026. URL https://gptzero.me.

[23] Brian Jabarian and Alex Imas. Artificial writing and automated detection. Working Paper 34223, September 2025. URL http://www.nber.org/papers/w34223.

[24] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. GPT detectors are biased against non-native English writers. *Patterns*, 4(7):100779, July 2023. doi: 10.1016/j.patter.2023.100779. URL https://www.sciencedirect.com/science/article/pii/S2666389923001307.

[25] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, February 2018. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

[26] Manav Raj, Justin M. Berg, and Rob Seamans. The artificial intelligence disclosure penalty: Humans persistently devalue AI-generated creative writing. *Journal of Experimental Psychology: General*, 155(4):896–915, 2026. doi: 10.1037/xge0001889.

[27] SFFA v. Harvard. Students for Fair Admissions, Inc., Petitioner, v. President and Fellows of Harvard College. Students for Fair Admissions, Inc., Petitioner, v. University of North Carolina, et al., 2023. https://www.supremecourt.gov/opinions/22pdf/20-1199_l6gn.pdf.

[28] Educational Testing Service. GRE concordance tables: Verbal reasoning and quantitative reasoning. Technical report, Educational Testing Service, Princeton, NJ, 2014. URL https://web.archive.org/web/20140823203631/ets.org/s/gre/pdf/concordance_information.pdf.

[29] QS Quacquarelli Symonds. QS World University Rankings 2025, 2024. URL https://www.topuniversities.com/world-university-rankings.

[30] National Center for Education Statistics. Integrated Postsecondary Education Data System (IPEDS), 2023, 2024. URL https://nces.ed.gov/ipeds/use-the-data. Admissions, graduation rates, and institutional characteristics data from the 2023 collection year, representing the 2022-23 academic year.

[31] OpenAI. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

[32] Microsoft. Azure AI Document Intelligence, 2024. URL https://azure.microsoft.com/en-us/products/ai-services/ai-document-intelligence.

[33] Scholaro, Inc. Scholaro database: International grade conversion, 2025. URL https://www.scholaro.com/db/Countries. Accessed: 2025-09-13.

[34] Bradley Emi and Max Spero. Technical Report on the Pangram AI-Generated Text Classifier, July 2024. URL http://arxiv.org/abs/2402.14873. arXiv:2402.14873 [cs].

[35] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 1877–1901, Red Hook, NY, USA, December 2020. Curran Associates Inc. URL https://dl.acm.org/doi/10.5555/3495724.3495883.

[36] Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011. doi: 10.18637/jss.v042.i08.

[37] Laurent Bergé. Efficient estimation of generalized linear models with high-dimensional fixed effects. *Journal of Statistical Software*, 86(1):1–28, 2018. doi: 10.18637/jss.v086.i01.

[38] Philipp Bach, Malte S. Kurz, Victor Chernozhukov, Martin Spindler, and Sven Klaassen. DoubleML: An object-oriented implementation of double machine learning in R. *Journal of Statistical Software*, 108(3):1–56, 2024. doi: 10.18637/jss.v108.i03.

[39] Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017. doi: 10.18637/jss.v077.i01.

[40] Michel Lang, Martin Binder, Jakob Richter, Patrick Schratz, Florian Pfisterer, Stefan Coors, Quay Au, Giuseppe Casalicchio, Lars Kotthoff, and Bernd Bischl. mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, Dec 2019. doi: 10.21105/joss.01903. URL `https://joss.theoj.org/papers/10.21105/joss.01903`.

A Supplementary notes on dataset curation

The data collection and featurization pipeline used for our analysis is described in greater detail in concurrent work [20, 21]. Here, we summarize the key elements relevant for this work, and provide further information about unique data used for this paper. In particular, Section A.1 details the construction of our university application dataset, the covariates we use in our predictive models, and the extraction and featurization pipeline used to structure applicants' transcripts and other admissions materials. Section A.2 discusses how we created our AI usage data.

A.1 Admissions data

Our data include applications to a two-year public policy master's program at a competitive American professional school submitted between 2020 and 2025, the most recent application cycle, totaling 7,462 applications. Each application contains a variety of information about an applicant, including biographical information and applicant metadata, admissions essays, previous or concurrent academic transcripts, resumes, letters of recommendation, and test scores. This set of materials comprises all of the information available to admissions officers at the time of admissions decision-making.

In collaboration with admissions officers at the university we partner with, we translate this structured and unstructured data into a rich dataset of 427 covariates. 30 of these features come from pre-structured application data. 216 come from our extraction pipeline described below. We use reported school information to join in another 181 features with information on an applicant's prior institution(s). Before fitting any models, we impute missing values, add indicators for missingness when appropriate, and remove non-varying and collinear columns. Further details are presented in Section D.

A.1.1 Structured data

Biographical information and applicant metadata In filling out their application, applicants report basic information about themselves. This includes concurrent applications to other programs at our partner institution, self-reported age, gender, race, military service, and other demographics. This information is captured in a structured format at the time of application submission, and presented to admissions officers at decision time.[5] Admissions officers are also presented with information regarding previous admissions cycles in which the applicant has applied, and whether they were admitted and enrolled.

Test scores As part of their application, students are required to submit GRE or GMAT scores.[6] Additionally, applicants who are not native English speakers and did not attend an English-language undergraduate institution are required to submit IELTS, TOEFL, or Cambridge English scores. Among the 1,580 applicants who report language test scores, roughly 63% submit TOEFL scores and 37% submit IELTS scores. No students report Cambridge English scores.

Between the GRE and GMAT, applicants primarily submit GRE scores, from which we include quantitative and verbal section scores into our predictive models. To include in our models the 6.4% of applicants who submit GMAT scores instead, we convert GMAT quantitative and verbal scores to equivalent percentile scores for the GRE to place GMAT and GRE scores on a common

[5] Following the 2023 Supreme Court decision in *SFFA v. Harvard*, legally protected characteristics such as race generally cannot be lawfully considered in admissions in the U.S. [27]. Following this decision, this data has since been unavailable to admissions officers. However, the data are still collected for administrative use.

[6] In 2020 and 2021, our partner institution suspended testing requirements because of the COVID-19 pandemic.

scale [28]. To compensate for this conversion, we additionally include an indicator for whether an applicant's "GRE" scores instead represent a converted GMAT score.

University rankings As part of their application materials, applicants are required to list each university in which they previously enrolled, in addition to submitting an official transcript from each school. To supplement our covariate set with metadata on the quality of an applicant's prior education, we leverage the 2025 QS World University Rankings [29] and IPEDS survey data [30]. The inclusion of these datasets allows us to incorporate university rankings, size, level of research activity, admission rates, graduation rates, financial information, demographic information, average test scores, and employment outcomes as covariates in our models. In total, we consider 44 features from IPEDS and 46 from QS data. In our models, we include this data for *both* their previous undergraduate and graduate degrees separately. Additionally, from the institutional information available in the IPEDS survey data, we predict QS rankings for schools with otherwise missing rankings. We split these rankings (real or predicted) into quintiles, and include that quintile as an additional feature, granting a total of 181 features from IPEDS and QS data.

A.1.2 Unstructured data

Transcripts Transcripts provide a rich source of information about an applicant's academic performance at previous institutions attended. However, the variety of formats and grading conventions makes them difficult to process at the scale required for predictive modeling. To incorporate this information into our models, we created the following process for processing and extracting features from this valuable unstructured data.

The pipeline for processing transcripts consisted of four steps:

1. First, we preprocessed the transcript files to enhance the quality of our subsequent extraction. In particular, we use OpenAI's `gpt-5-nano`'s [31] multimodal capabilities to identify the rotation of each page in the transcript. We un-rotate any rotated pages so that the text reads horizontally. We additionally increase the text contrast on each page and remove any watermarks.

2. Once preprocessed, each transcript is transcribed to markdown via Microsoft Document Intelligence [32].

3. Then, we prompt `gpt-5-mini` to extract course-level information (e.g., term, title, course codes, credits, grade, instructor name, and level) and further term- and degree-specific information (e.g., official institution name and address, matriculation and graduation dates, and majors and minors) in a loosely structured format, primarily to remove extraneous information and reduce errors downstream.

4. Finally, this loosely structured representation of an applicant's transcript is converted into structured output in accordance with a fixed JSON schema, again by `gpt-5-mini`.

After processing the raw transcript files, we then featurized this newly structured data into a set of 155 measures of previous academic performance. These include overall GPAs, subject area GPAs (e.g., mathematics and statistics, economics), course levels (e.g., upper- and lower-undergraduate), GPA trends over time, performance in specific courses (e.g., introductory micro- and macroeconomics), indicators for majors, minors, and honors, and other similar measures. For applicants with multiple transcripts, we calculate credit-weighted GPAs across institutions.

To ensure a constant 4.0 scale across all applicants in our sample, approximately half of whom come from international institutions, we prompt `gpt-5-mini` in Step 4 to convert the GPA to a 4.0 scale, if a conversion key was extracted from the raw transcript. If not, we include in our prompt all available conversion keys from the Scholaro database [33] for the relevant country (the same resource our partner admissions office uses to convert GPAs when no scale is provided).

Letters of recommendation Letters of recommendation are submitted in a more standardized format than transcripts. The applicant must submit a total of three letters of recommendation from prior professors, professional supervisors, or other individuals who can speak to the applicant's readiness for the program. Together with our partner admissions office, we developed two rubrics for evaluating the written text submitted in the letter. The first focused on the relevance of a recommender's relationship to the applicant, the strength of their recommendation, the recommender's profession, seniority, and employing organization, their apparent depth of knowledge about the candidate and the length of their relationship, as well as the letter writer's account of the applicant's academic preparedness, professional leadership, commitment to public service, and curiosity. The other rubric focuses specifically on whether the letter makes specific reference to applied quantitative skills or other evidence of quantitative preparation for the program.

In addition to the written letter, recommenders also submit a cover sheet, providing information about themselves, their relationship with the applicant, the number of other letters they are writing, and the individuals against whom the applicant is being compared. Additionally, recommenders answer Likert-style questions about an applicant's academic preparation, ability to express themselves, work ethic, personality traits, and other dimensions of their preparedness for graduate school.

As with transcripts, we convert an applicant's three letter/cover sheet pairs into markdown using Microsoft Document Intelligence [32]. In three separate prompts, we then prompt `gpt-5-nano` to extract the structured applicant ratings from the cover sheet and evaluate it according to our two rubrics. These three extractions grant a total of 39 measures related to an applicant's letters of recommendation.

Resumes Once again, we leverage Microsoft Document Intelligence [32] to convert an applicant's submitted resume PDF into markdown text. We then prompt `gpt-5-nano` to structure the applicant's work history. To do so, for each position listed in their resume, we extract the applicant's title and tenure, employment status (intern, part-time, full-time) and seniority, sector (for-profit, non-profit, government), size and prominence of their employer, leadership, public service, or quantitative responsibilities, and other key information. We also prompt the model to classify the applicant's career trajectory according to the promotions, demotions, and lateral career moves contained in their resume. From each applicant's structured work history, we calculate 8 summary measures of program-relevant work experience, such as their number of years of work in the public sector.

Essays Finally, as part of their application, applicants are required to submit at least 5 essays of around 250 to 500 words, covering topics like their interest in the specific program, the unique perspective they bring to the school, and their interest in public service. This core set of essay topics remains constant throughout our six-year sample, barring one essay. The essay's prompt changes significantly in 2023, though the core question remains qualitatively similar.[7]

To evaluate the quality of these essays in a way that matches that of an admissions officer, we develop a rubric that mirrors the rubric used in the admissions process to evaluate seven

[7]We reproduce our analysis without this essay in Section E, and find qualitatively similar results.

aspects of the essays: their grammatical correctness, stylistic quality, clarity and organization, adherence to the prompt and formal requirements, voice and originality, public service merit, and evidence of leadership potential. To do so, after converting to markdown with Microsoft Document Intelligence [32], we again prompt `gpt-5-mini` to grade each essay according to our rubric, leveraging concrete evaluation criteria for each dimension. For instance, grammatical quality is scored according to the number and severity of errors. Evidence of leadership potential was rated in accordance with whether the applicant mentioned a specific future leadership role, and the number of details they provided about that position and their plans to attain it.

Additionally, admissions officers report placing particular emphasis on the contents of one of the five essays. To account for this reported heightened attention to this essay, we include the LLM-assessed quality scores for that essay as an additional seven features, granting 14 total essay quality features.

A.1.3 Featurization

After compiling this expansive list of applicant covariates, we adopt standard machine learning approaches to handle data missingness. In particular, we impute any missing numerical covariates with their median, and impute any missing logical covariates with `FALSE`. To allow the model to learn if missingness itself is important, we additionally include a missingness indicator for features where missingness may be informative. In total, from our featurization pipeline, we are left with 427 complete covariates and 420 missingness indicators.

A.1.4 Outcomes

Our two primary outcomes of concern are essay quality and admission probability. In Section A.1.2 and in Section C, we detail our process for generating essay quality ratings, which, as we show in the main text (Figure 2), align well with human ratings of essay quality. Therefore here we only detail our admissions-related outcome.

In addition to each applicant's application materials, we have access to their admissions decisions. We use a binary indicator of whether or not the applicant was admitted to the program as our main outcome of interest. Importantly, though we have access to application materials from the most recent round (submitted in 2025) the admissions office we partner with substantially changed its admissions process in light of our preliminary results, and thus we do not analyze outcomes for that cycle.

A.2 AI detection

Because we cannot observe true AI authorship in the application data, our empirical approach depends on the accuracy of AI detection tools, as we use their detection to serve as a proxy for true AI usage. To minimize measurement error in our estimates of potential benefits or penalties associated with AI usage, we explore two detectors recently shown to have impressively low false positive and false negative rates: GPTZero and Pangram [22, 23, 34]. In our main analysis, we consider GPTZero's model `3.9`. In our robustness checks, we use Pangram's version `3.1`.

A.2.1 Additional text cleaning

Before running an essay through any of our detectors, we completed an additional layer of text cleaning to remove any markdown artifacts from Microsoft Document Intelligence [32] and to strip any headers or essay prompts from the essay text, ensuring the AI detection classifiers only ran

on the applicant's submitted writing. To do so, we prompted `gpt-5-nano` [31] to reproduce each essay without any markdown tags, student metadata included in the document's header, or prompt text. To ensure the model did not alter the text of the applicant's essay, we validated the cleaning by computing character-level diffs between the original and cleaned versions. For any essay with a non-zero diff, we prompted `gpt-5-nano` to flag any instances where the cleaning introduced errors, altered meaning, or changed the original text. We manually reviewed all flagged cases ($n \approx 1200$), and confirmed changes made were strictly subtractive of non-essay text, manually reverting LLM-made changes for a small number of essays. In addition to validating that the LLM-cleaner did not alter the essay text, we further confirm the resulting labels do not change significantly. Following cleaning, labels change for 1,015 (2.7%) essays, with a large majority (865) changing classification from AI to non-AI.

A.2.2 Detector error rate validation

As discussed in the main text, GPTZero classifies text into one of three provenances: "AI," "Mixed," or "Human." GPTZero additionally reports the confidence of their classification as "Low," "Medium," or "High." The other detector we consider here, Pangram, uses similar "AI," "Human," and "Mixed" labels, but does not provide document level confidence assessments. To determine which, if any, of these provenances achieved an error rate appropriate for our analysis, we first investigated each detector's error rates on our data for a variety of different thresholds signifying "AI use."

False positive rates Leveraging our historical data, we confirm the false positive rates of each detector on the essays submitted prior to the release of ChatGPT.[8] Tables A1 and A2 report the AI detection rates by year of material submission across a variety of classification thresholds, at the applicant and essay level, respectively. At the essay level, both GPTZero and Pangram display impressively low false positive rates for their "AI" labels, matching prior analyses of AI detectors [23]. For GPTZero, this low false positive rate increases marginally when the required confidence threshold is relaxed, and slightly more significantly when "mixed" labels are considered as AI usage. When both "mixed" labels are used and confidence levels are relaxed, the essay-level false positive rate increases by an order of magnitude. Pangram also displays a much higher false positive rate when "mixed" labels are considered. As applicants submit multiple essays in their application, these false positive rates increase when aggregated to the applicant level, as expected, though the general patterns remain the same.

False negative rates In our context, we cannot meaningfully estimate false negative rates, as we do not have access to ground truth applicant AI usage. To the extent to which AI usage goes undetected, non-AI users in our data will include some true AI users, and therefore our estimates may reflect the effect of detectable AI usage rather than AI usage at large. However, prior work reports low false negative rates on text entirely written by AI [23]—precisely the provenance captured by GPTZero and Pangram's "AI" labels. This suggests that our labels are unlikely to systematically miss essays that were fully AI-authored, limiting measurement error concerns for the treatment we consider in our main analysis.

[8]The application deadline for applications submitted in 2022 was one day after November 30th, so we assume no essays submitted in 2022 were written using LLMs. While large language models existed prior to the commercial release of ChatGPT [e.g. as in 35], we assume their use in applicant essays is plausibly negligible in practice, as models were less accessible to broad audiences at the time.

A.2.3 Detector selection

As our primary estimands are applicant-level, we base our detector selection on applicant-level false positive rates. As shown in Table A1, only GPTZero "AI" labels with "high" confidence and Pangram "AI" labels maintain false positive rates below 1% (on average) on the pre-ChatGPT data (0.8% and 0.3%, respectively). To prevent overstating our prevalence estimates and to mitigate noise in treatment identification, we therefore only consider these two thresholds as indicative of AI authorship in our main analysis. We selected GPTZero for our main text results, though Pangram's "AI" labels produce qualitatively similar results. Section E provides more commentary on the robustness of our results to different threshold choices and AI detection software.

A.2.4 Usage by demographics

Figure A1 reports AI usage rates across demographic subgroups. Most strikingly, we find a large and persistent gap in usage between domestic and international applicants. International applicants show substantially higher AI usage across every subgroup we examine. This gap is most pronounced for applicants below the age of 24. AI usage is also weakly correlated with quantitative GRE scores for both groups, and flat across verbal scores, suggesting AI usage is not concentrated among weaker applicants.

B Experimental details

To assess how well admissions officers can detect AI-written content and to better understand how they determine an essay's quality, we ran an experiment on the admissions officers at our partner university. (Our study, protocol number IRB24-0242, was granted IRB approval by the Harvard University-Area Committee on the Use of Human Subjects. Consent was obtained in person, and data were analyzed anonymously.) From the admissions office, we recruited five staff members who were going to review applications in the upcoming admissions cycle (2025–2026 cycle). We selected 100 essays (50 human, 50 AI) submitted in either 2023 or 2024, all in response to the same prompt. To ensure AI and human essays came from applicants with similar observable profiles, we used `MatchIt` [36] to construct a 1:1 nearest-neighbor propensity score matching on applicant characteristics. We then randomly sampled 50 matched pairs from the resulting set, granting 50 human- and 50 AI-written essays from applicants with similar observable characteristics.

Each officer was tasked to review all 100 essays sequentially, in a random order, with the option to pause and resume their reviews at any time. After reading each essay, participants were asked to rank, on a six-point Likert scale, both how likely they thought the essay was AI-generated and how they would rate its quality in comparison to essays they had reviewed as part of the admissions process. The options on the AI-likelihood scale included: "Very likely human," "Moderately likely human," "Somewhat likely human," "Somewhat likely AI," "Moderately likely AI," and "Very likely AI." The options on essay quality scale included: "Much weaker," "Moderately weaker," "Somewhat weaker," "Somewhat stronger," "Moderately stronger," and "Much stronger."

Of the participants, Raters 2–5 rated all 100 essays. Rater 1 only rated 60 (32 human, 28 AI), which grants a total of 460 ratings. Figures A2 and A3 show the distributions of each rater's AI likelihood ratings and essay quality ratings, respectively. Most raters demonstrate the ability to discriminate between AI and human essays, though imperfectly: Human essays cluster at lower values and AI essays receive higher ratings on average. That said, Rater 5 assigned nearly all essays a rating of 2 on the AI likelihood scale regardless of provenance, suggesting limited ability

or willingness to discriminate between AI and human text. In terms of quality, ratings are more uniform across raters and essay types, consistent with quality being a more subjective judgment.

C Essay quality

Here, we provide further detail on our results pertaining to essay quality. This section contains two parts. First, we detail how we construct human-calibrated quality for all essays in our dataset. Then, we discuss our replication of the results of the essay quality time series analysis, controlling for applicant observables.

C.1 Human calibrated essay quality

After completing the experiment detailed in Section B, we fit a linear regression model in R to extract human weights on our different LLM-extracted dimensions of essay quality. As we report in the main text, we observe a harsh quality penalty when a rater believes an essay to be "very likely AI," so we drop those ratings ($n = 22$) to remove that bias from our estimates. We then regress each rater's quality rating on our seven LLM-assessed dimensions of essay quality, with a rater fixed effect to absorb individual differences in the raters. Table A3 displays the coefficients on each quality dimension as estimated by this model. We use these coefficients as weights for computing our human-calibrated composite quality score, used throughout the main text. As Figure 2 shows, our human-calibrated composite reflects human preferences well.

C.2 The effect of AI use on essay quality

In the main text, we report a timeseries of essay quality over the six years in our sample, to demonstrate the large improvement in writing quality we observe following the release of ChatGPT. However, it is possible that these gains were instead driven by a shift in the applicant pool. To explore this counterfactual, we produce a similar time series to that of Figure 3 for each dimension of LLM-assessed essay quality while controlling for the rich set of applicant observables described in Section A.1, except for essay quality ratings. We estimate separate models for domestic and international applicants. Instead of the group means as reported in Figure 3, Figure A4 reports the coefficients for each year in those models, providing estimates of essay quality by year controlling for applicant characteristics. These coefficients are then normalized by the 2020 *within-group* standard deviation to ensure a common scale across years and quality dimensions, and track changes relative to the group's baseline. Models are estimated using `feols` from the `fixest` package [37] in R.

We find the gains in essay quality reported in the main text persist after controlling for applicant observables, suggesting the trends observed in Figure 3 are not driven solely by changes in the composition of the applicant pool. Pre-ChatGPT estimates for mechanical quality (grammar, style and clarity), are statistically indistinguishable from zero, followed by dramatic post-ChatGPT gains—particularly for international applicants, whose mechanical gains are roughly double those of domestic applicants by 2025. Substantive dimensions also trend upward, however some (e.g., leadership, prompt responsiveness, and public service) exhibit trends that begin prior to the release of ChatGPT, making attribution to AI-use less clear. These patterns are most consistent with AI improving the mechanical polish of essays, particularly for international applicants, with less clear evidence of the effects on substantive content.

D Admissions penalties

In this section, we provide further details on our DML estimates of the effect of AI usage on admissions outcomes. We estimate partially linear models using the `DoubleML` R package [25, 38]. Our treatment variable is the number of an applicant's essays classified as AI-written, and the outcome is an indicator for admission. While in our prevalence estimates we define AI users as applicants who submitted at least one AI-written essay, in the real admissions process (without access to AI detection tools) admissions officers' demonstrated imperfect detection ability suggests their detection signal of AI use likely scales in the number of AI-written essays submitted by an applicant. To capture this effect, we model a continuous count treatment. The coefficient from these models thus represents the estimated change in admission probability associated with one additional essay classified as AI-written, conditional on observable characteristics. To confirm our results are not an artifact of our continuous treatment effect, we estimate a binary treatment effect of *any* AI usage in Section E, and find qualitatively similar results.

We report both pooled and admissions-cycle specific estimates for applications submitted in 2023 and 2024. The pooled model combines these two cohorts, and includes an admissions cycle indicator among other applicant-level controls. The cycle-specific estimates are fit only on that cycle's data (e.g., the coefficient for AI Count (2023) is only estimated on the 2023 applicant pool.) We fit separate nuisance functions per cycle.

The estimates presented in columns (2) and (3) of Table 1 in the main text arise from two highly similar specifications. Column (2) controls for the large array of covariates described in Section A.1, but excludes essay quality measures. Because AI usage may affect essay quality, conditioning on these measures could introduce post-treatment bias. Column (3) includes these 14 controls as a descriptive check. While the post-treatment bias complicates the causal interpretation of these estimates, that the penalty grows larger after adjusting for essay quality suggests the admissions penalty is not explained by differences in essay quality between AI and non-AI users.

Before fitting any models, we first eliminate any columns that are single valued, eliminate any remaining collinear columns by fitting a linear regression on the full covariate set, dropping any features whose coefficients could not be estimated, and finally remove any columns that do not exhibit variation in both 2023 and 2024, separately. After these exclusions, we retain a common set of 411 covariates that exhibit variation across both admissions cohorts, 70 of which are missingness indicators. Our primary specification excludes the 14 essay-level quality measures, and thus includes 397 features. In the pooled models, we add an application cycle term that increases both counts to 398 and 412, respectively.

We estimate the treatment nuisance function and the outcome nuisance function using random forests via `mlr3` with `ranger` backend [39, 40]. We tune the two nuisance functions separately for each estimation sample and covariate specification using 5-fold cross validation. To select hyperparameters, we search over 30 random combinations of `mtry`, minimum node size, maximum tree depth, and sampling fraction using 200 tree forests. We then fit the final nuisance models using the selected hyperparameters and 500 trees, with 5-fold cross-fitting.

Cross-seed aggregation. To account for simulation variability introduced by the random cross-fitting fold assignment, we repeat each DML specification across 20 random seeds and aggregate the results using the repeated splitting procedure implemented by the DoubleML package [38]. All DML estimates, both in the main text and those provided in Section E reflect this repeated-splitting aggregation procedure.

Residual variation checks. Identification of the partially linear model requires the AI essay count to retain variation after conditioning on our included sets of covariates. Table A4 reports the residual variance and cross-fitted R^2 for each treatment nuisance model. For the specifications presented in the main text, residual variances range from 0.80 to 2.04, and cross-fitted $R^2 \leq 0.32$, indicating ample residual treatment variation after conditioning on our covariate set. Across all of our continuous-treatment robustness checks, we similarly find sufficient residual variation. For our binary treatment effect robustness check, we report overlap plots in Figures A5 and A6.

E Robustness

To evaluate the robustness of our admissions results, we vary our analysis along four dimensions. Across all of these dimensions of variation, our results persist.

1. **Detector choice:** We first vary which AI detection software's labels we consider. Table A5 replicates our main results using Pangram [34]. The effects are qualitatively identical to those we find with GPTZero.

2. **Detection threshold:** In addition to investigating robustness for different detection software, we additionally report estimates for looser "AI generated" classification thresholds. In particular, columns (1)-(4) in Table A6 report estimates of the AI-use penalty when we include "mixed" label classifications or, in the case of GPTZero, lower confidence classifications. Our pooled estimates are stable across all variations, whereas the year-specific estimates are noisier.

3. **Omission of essay with changed prompt:** As mentioned in the main text, one essay prompt changed significantly in 2023. In Table A6, column (5), we reproduce our main results excluding this essay entirely. We find qualitatively similar estimates, suggesting the change in this essay's prompt is not impacting our estimated effect.

4. **Binary treatment:** Finally, we replace the count treatment with a binary indicator for whether an applicant submitted any AI-written essay. As seen in Table A7, we find a penalty on any AI usage that is consistent with our main result. We provide support of the overlap assumption for binary treatment in Figures A5 and A6.

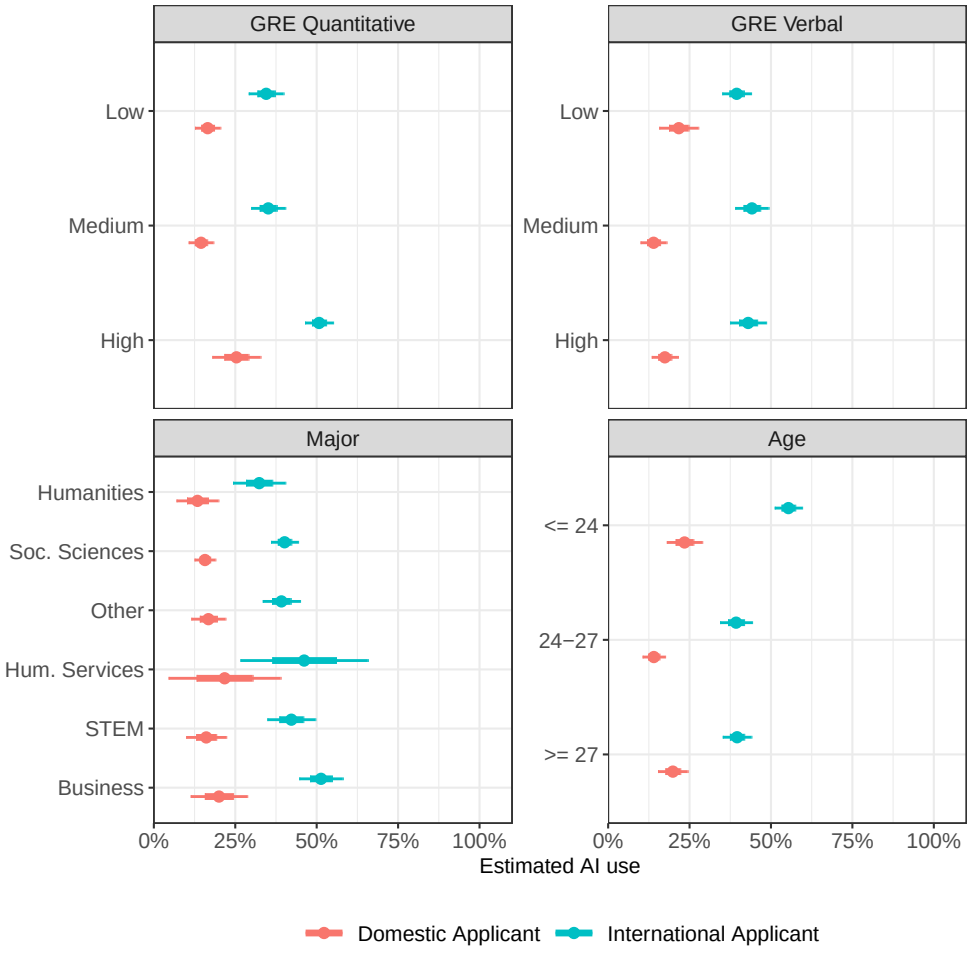

Figure A1: GPTZero estimated applicant-level AI usage by demographic characteristics, disaggregated by nationality. The "Low," "Medium," and "High" bins in the GRE plots contain the lower, middle, and upper tertiles of GRE scores, respectively. Age bins each contain roughly one third of the applicant pool.

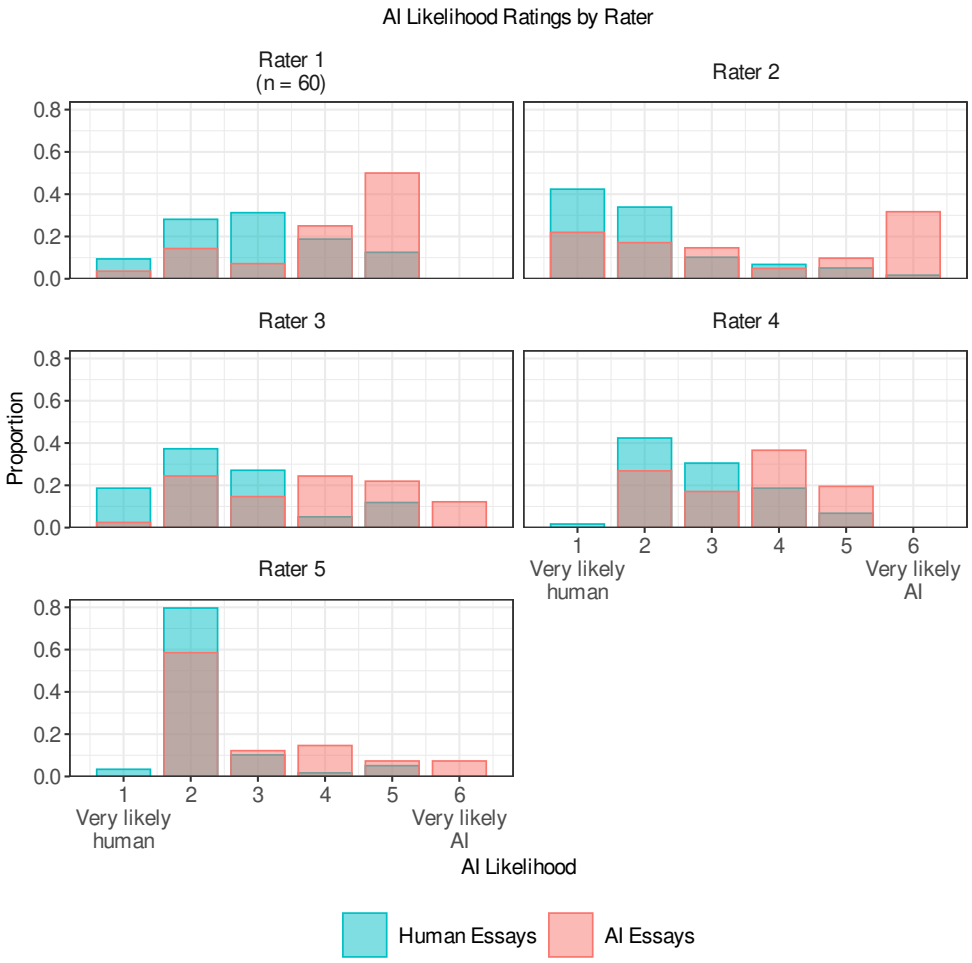

Figure A2: Distribution of AI likelihood ratings by rater, disaggregated by essay provenance. Each panel shows one admissions officer's ratings on a six-point scale ranging from 1 ("Very likely human") to 6 ("Very likely AI"). Bars show the proportion of AI-written and human-written essays receiving each rating. Rater 1 completed 60 of the 100 essays.

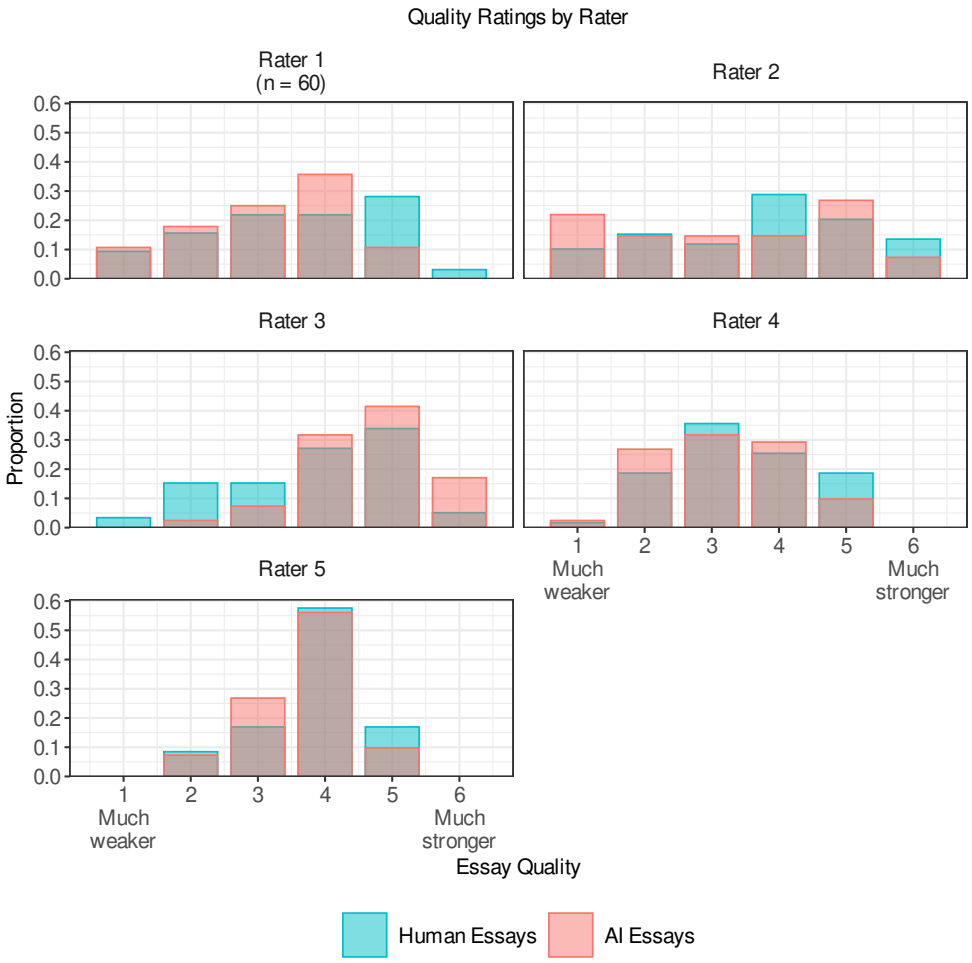

Figure A3: Distribution of essay quality ratings by rater, disaggregated by essay provenance. Each panel shows one admissions officer's ratings on a six-point scale ranging from 1 ("Much weaker") to 6 ("Much stronger") relative to a typical applicant. Bars show the proportion of AI-written and human-written essays receiving each rating. Rater 1 completed 60 of the 100 essays.

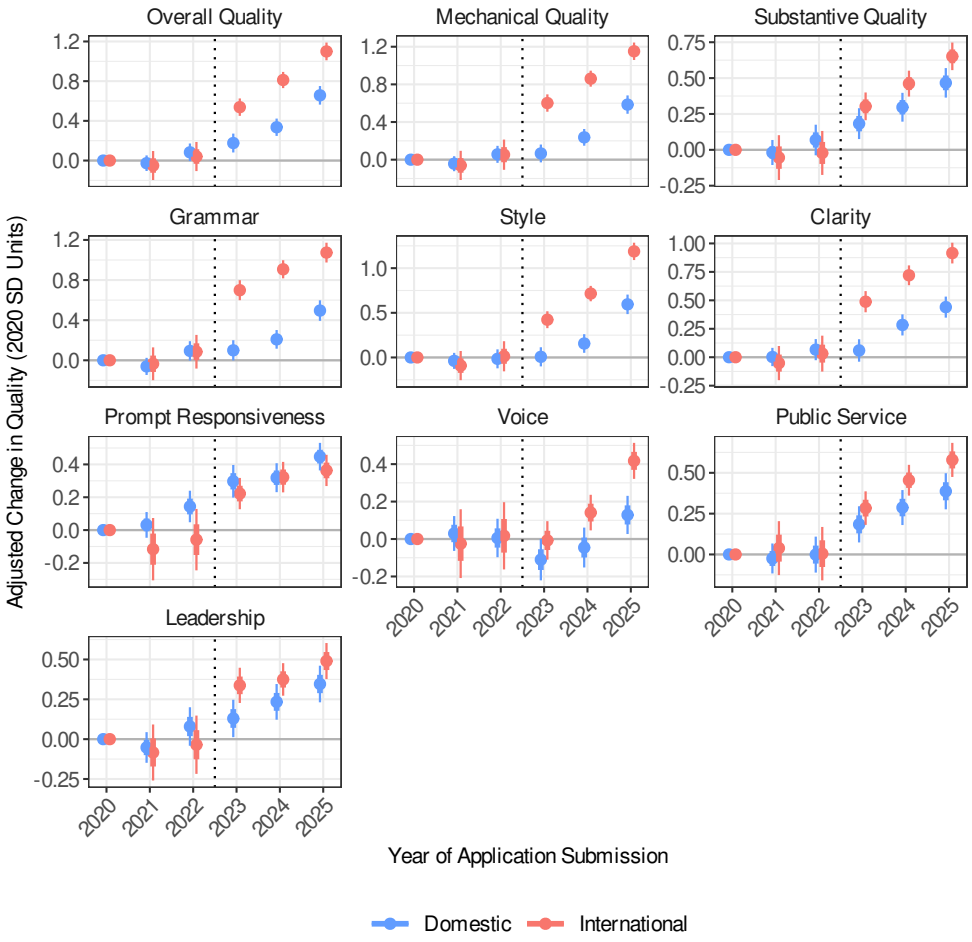

Figure A4: Change in LLM-assessed essay quality over time, disaggregated by residency status, conditional on applicant covariates. Points show OLS estimates of round coefficients relative to applicants in 2020, normalized to 2020 SD units. Error bars show ±1 and ±2 SE, clustered at the applicant level. Baseline year (2020) is normalized to zero within group. The dashed line marks the release of ChatGPT in late 2022.

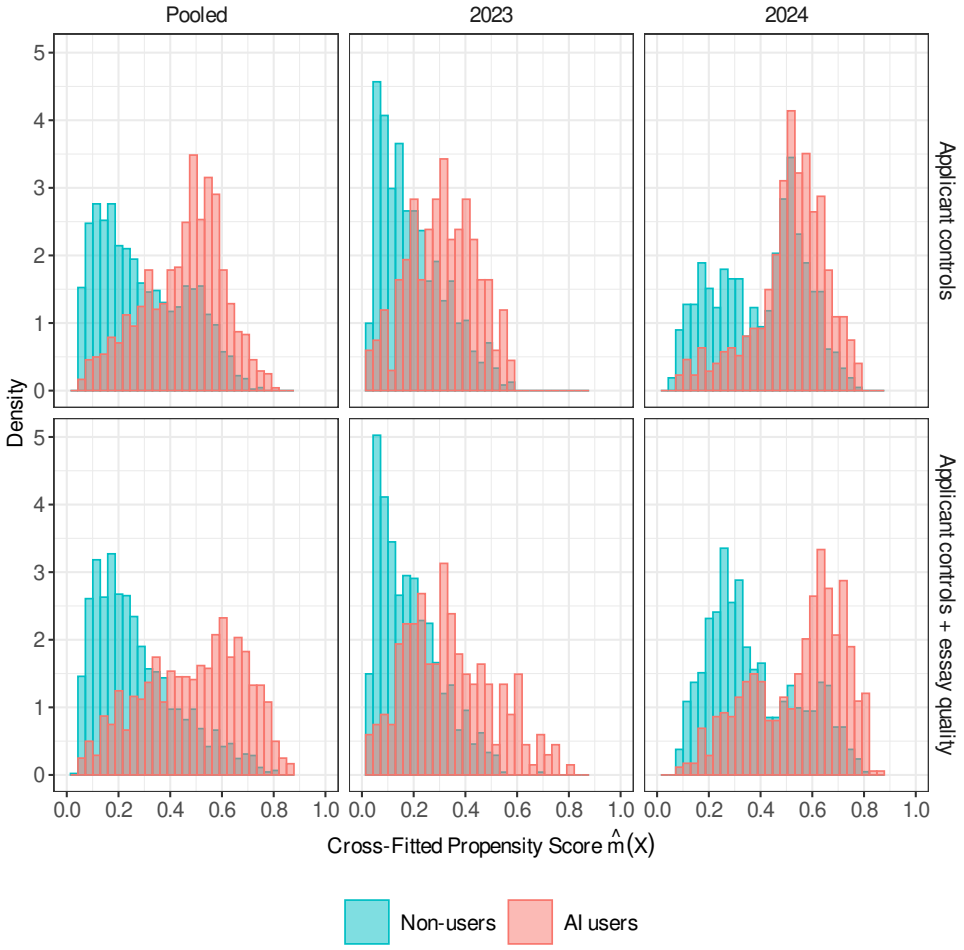

Figure A5: Overlap diagnostics for the GPTZero binary treatment specification. Each panel shows the distribution of $\hat{P}(D = 1|X)$ from the DML treatment nuisance model (random forest), for AI-using and non-AI-using applicants. Columns correspond to application cohort (2023, 2024), rows correspond to specification.

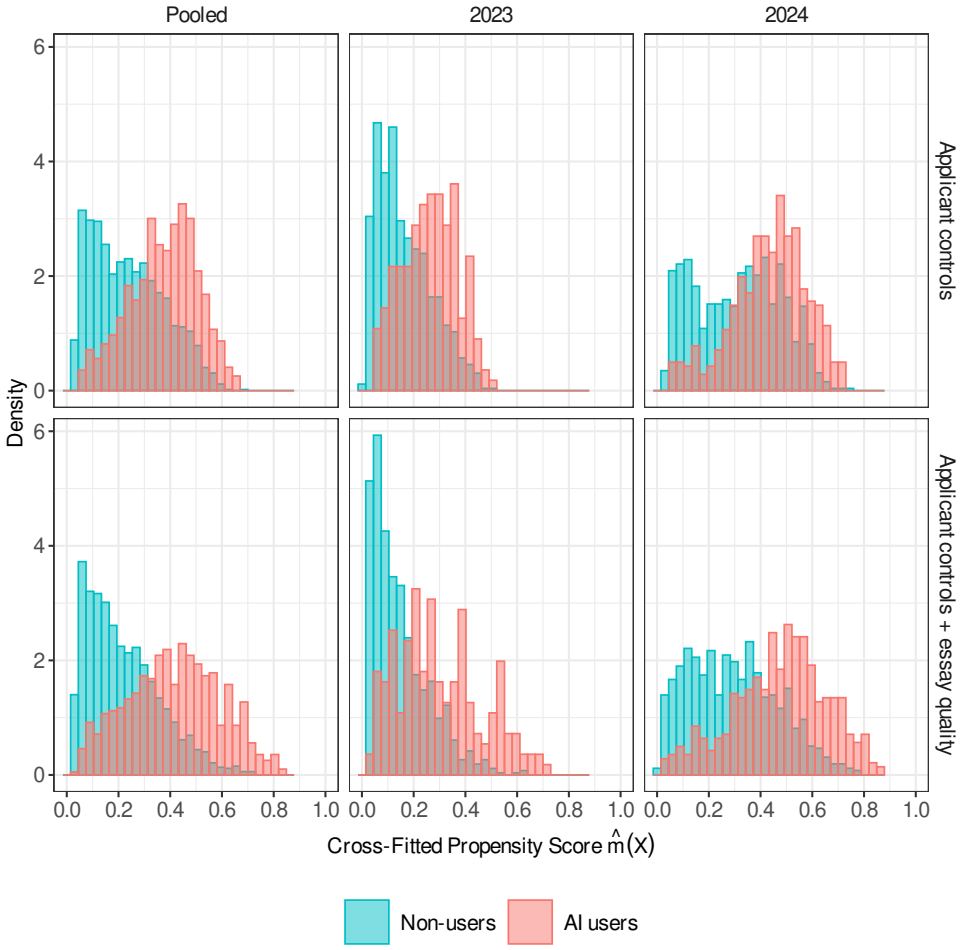

Figure A6: Overlap diagnostics for the Pangram binary treatment specification. Each panel shows the distribution of $\hat{P}(D = 1|X)$ from the DML treatment nuisance model (random forest), for AI-using and non-AI-using applicants. Columns correspond to application cohort (2023, 2024), rows correspond to specification.

Table A1: AI Detection Rates by Application Year and Detector (Applicant Level)

Year	GPTZero AI Only High Confidence (1)	GPTZero AI + Mixed High Confidence (2)	GPTZero AI Only Any Confidence (3)	GPTZero AI + Mixed Any Confidence (4)	Pangram AI Only (5)	Pangram AI + Mixed (6)	GPTZero AI Only Excl. Changed Essay (7)
False-positive rates (pre-ChatGPT)							
2020	*0.7% (0.2%)*	*4.0% (0.5%)*	*1.7% (0.3%)*	*6.4% (0.6%)*	*0.2% (0.1%)*	*6.3% (0.6%)*	*0.5% (0.2%)*
2021	*0.5% (0.2%)*	*2.9% (0.5%)*	*1.3% (0.3%)*	*4.7% (0.6%)*	*0.4% (0.2%)*	*7.2% (0.7%)*	*0.5% (0.2%)*
2022	*1.4% (0.4%)*	*4.3% (0.7%)*	*1.8% (0.5%)*	*6.8% (0.9%)*	*0.5% (0.2%)*	*11.9% (1.1%)*	*0.8% (0.3%)*
Detection rates (post-ChatGPT)							
2023	21.8% (1.3%)	38.2% (1.5%)	25.8% (1.3%)	43.2% (1.5%)	17.4% (1.2%)	47.1% (1.5%)	18.0% (1.2%)
2024	45.1% (1.4%)	62.6% (1.3%)	47.9% (1.4%)	65.9% (1.3%)	35.3% (1.3%)	64.7% (1.3%)	39.3% (1.3%)
2025	56.1% (1.5%)	71.5% (1.3%)	60.0% (1.5%)	75.6% (1.3%)	45.9% (1.5%)	63.0% (1.4%)	50.4% (1.5%)

Note: Entries report the percentage of applicants classified as AI users, with binomial standard errors in parentheses. An applicant is classified as an AI user if at least one submitted essay satisfies the detector definition shown in the column heading. Italicized rows correspond to pre-ChatGPT application cycles and are interpreted as false-positive rates. Application year refers to the year in which the application was submitted. Column (7) excludes the changed essay from the classification.

Table A2: AI Detection Rates by Application Year and Detector (Essay Level)

Year	GPTZero AI Only High Confidence (1)	GPTZero AI + Mixed High Confidence (2)	GPTZero AI Only Any Confidence (3)	GPTZero AI + Mixed Any Confidence (4)	Pangram AI Only (5)	Pangram AI + Mixed (6)	GPTZero AI Only Excl. Changed Essay (7)
False-positive rates (pre-ChatGPT)							
2020	*0.1% (0.0%)*	*0.8% (0.1%)*	*0.3% (0.1%)*	*1.4% (0.1%)*	*0.0% (0.0%)*	*1.4% (0.1%)*	*0.1% (0.0%)*
2021	*0.1% (0.0%)*	*0.6% (0.1%)*	*0.3% (0.1%)*	*1.0% (0.1%)*	*0.1% (0.0%)*	*1.6% (0.2%)*	*0.1% (0.0%)*
2022	*0.3% (0.1%)*	*1.0% (0.2%)*	*0.4% (0.1%)*	*1.6% (0.2%)*	*0.1% (0.0%)*	*2.8% (0.3%)*	*0.2% (0.1%)*
Detection rates (post-ChatGPT)							
2023	9.0% (0.4%)	19.3% (0.5%)	10.5% (0.4%)	24.0% (0.6%)	7.4% (0.4%)	27.5% (0.6%)	8.1% (0.4%)
2024	21.9% (0.5%)	38.0% (0.6%)	24.0% (0.5%)	45.4% (0.6%)	17.8% (0.5%)	40.7% (0.6%)	20.5% (0.6%)
2025	29.4% (0.6%)	44.9% (0.7%)	32.6% (0.6%)	53.7% (0.7%)	23.2% (0.6%)	34.5% (0.6%)	28.0% (0.7%)

Note: Entries report the percentage of essays classified as AI-written, with binomial standard errors in parentheses. Italicized rows correspond to pre-ChatGPT application cycles and are interpreted as false-positive rates. Application year refers to the year in which the application was submitted. Column (7) excludes the changed essay from both the numerator and denominator.

Table A3: Quality Model Dimension Weights

Dimension	Coefficient	95% confidence interval
Grammar	0.208	[0.063, 0.354]
Style	0.207	[0.023, 0.391]
Clarity	0.105	[-0.132, 0.343]
Prompt responsiveness	0.479	[0.014, 0.944]
Voice	0.226	[0.107, 0.346]
Public service	0.213	[0.059, 0.368]
Leadership	0.064	[-0.073, 0.201]

Note: Coefficients are estimated from an OLS regression of human quality ratings on the seven LLM-based quality dimensions with rater fixed effects. Brackets report 95% confidence intervals. The intercept and rater fixed effects are omitted.

Table A4: Residual Variation in AI Essay Count

		Applicant controls only		Applicant controls + essay quality	
Treatment	Sample	$\widehat{E[V^2]}$	R^2	$\widehat{E[V^2]}$	R^2
GPTZero: AI Only High Confidence	Pooled	1.550	0.172	1.276	0.319
	2023	0.960	0.113	0.804	0.257
	2024	2.042	0.121	1.719	0.260
GPTZero: AI + Mixed High Confidence	Pooled	2.446	0.216	1.851	0.407
	2023	1.898	0.150	1.495	0.330
	2024	2.887	0.161	2.178	0.367
GPTZero: AI Only Any Confidence	Pooled	1.670	0.184	1.394	0.319
	2023	1.049	0.127	0.894	0.256
	2024	2.206	0.125	1.864	0.260
GPTZero: AI + Mixed Any Confidence	Pooled	2.959	0.226	2.243	0.414
	2023	2.374	0.171	1.905	0.335
	2024	3.430	0.161	2.541	0.378
Pangram: AI Only	Pooled	1.438	0.165	1.231	0.285
	2023	0.866	0.115	0.757	0.227
	2024	1.908	0.131	1.713	0.220
Pangram: AI + Mixed	Pooled	2.761	0.246	2.187	0.403
	2023	2.485	0.241	2.118	0.353
	2024	2.988	0.209	2.342	0.380
GPTZero: AI Only Excl. Changed Essay	Pooled	1.019	0.148	0.853	0.286
	2023	0.605	0.102	0.517	0.233
	2024	1.352	0.101	1.170	0.222

Note: All specifications include applicant controls; the right panel additionally includes essay-quality measures. Residual variation is the mean squared out-of-fold treatment residual, $\frac{1}{n}\sum_i (D_i - \widehat{m}(X_i))^2$*. Entries are medians across 20 independently seeded fits. Pooled models include an admissions-cycle indicator; year-specific models are estimated separately within each cohort.*

Table A5: Pangram AI Usage and Admission

	OLS	DML	
	(1)	(2)	(3)
AI Count	-0.083***	-0.016**	-0.022***
	(0.007)	(0.006)	(0.006)
AI Count (2023)	-0.086***	-0.026*	-0.019
	(0.014)	(0.011)	(0.012)
AI Count (2024)	-0.082***	-0.016*	-0.020**
	(0.008)	(0.007)	(0.007)
Applicant controls		✓	✓
Essay quality			✓

Note: Standard errors are reported in parentheses. The first row reports estimates from models without year interactions; the final two rows report estimates from separate year-specific models. Column (1) reports unadjusted OLS estimates from linear probability models. Columns (2) and (3) report DML estimates using random forest nuisance functions and 5-fold cross-fitting. Across 20 independently seeded fits, coefficients and standard errors are aggregated using DoubleML's median-based repeated-splitting procedure [25]. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A6: DML Estimates: Robustness Across Detector Specifications

	Applicant controls				
	GPTZero AI + Mixed High Confidence (1)	GPTZero AI Only Any Confidence (2)	GPTZero AI + Mixed Any Confidence (3)	Pangram AI + Mixed (4)	GPTZero AI Only Excl. Changed Essay (5)
Pooled	-0.015**	-0.015**	-0.014**	-0.012**	-0.017*
	(0.005)	(0.005)	(0.004)	(0.005)	(0.007)
AI Count (2023)	-0.026**	-0.017	-0.018*	-0.013	-0.028
	(0.009)	(0.011)	(0.008)	(0.008)	(0.014)
AI Count (2024)	-0.008	-0.010	-0.009	-0.009	-0.014
	(0.006)	(0.006)	(0.005)	(0.006)	(0.008)
	Applicant controls + essay quality				
	GPTZero AI + Mixed High Confidence (1)	GPTZero AI Only Any Confidence (2)	GPTZero AI + Mixed Any Confidence (3)	Pangram AI + Mixed (4)	GPTZero AI Only Excl. Changed Essay (5)
Pooled	-0.027***	-0.021***	-0.025***	-0.021***	-0.025**
	(0.006)	(0.006)	(0.005)	(0.005)	(0.008)
AI Count (2023)	-0.033***	-0.022	-0.024**	-0.015	-0.024
	(0.010)	(0.012)	(0.009)	(0.008)	(0.016)
AI Count (2024)	-0.020**	-0.018**	-0.022***	-0.019**	-0.025**
	(0.007)	(0.007)	(0.006)	(0.007)	(0.008)

Note: Standard errors are reported in parentheses. The first row of each panel reports pooled estimates; the final two rows report estimates from separate year-specific models. All models use random forest nuisance functions and 5-fold cross-fitting. Across 20 independently seeded fits, coefficients and standard errors are aggregated using DoubleML's median-based repeated-splitting procedure [25]. Column (5) excludes the changed essay from treatment construction and from the essay-quality controls in the lower panel. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A7: DML Estimates: Binary Treatment Robustness

	Applicant controls		Applicant controls + essay quality	
	GPTZero Binary	Pangram Binary	GPTZero Binary	Pangram Binary
	(1)	(2)	(3)	(4)
Pooled	-0.036*	-0.060**	-0.056**	-0.076***
	(0.018)	(0.019)	(0.019)	(0.020)
Any AI (2023)	-0.032	-0.050	-0.043	-0.053
	(0.031)	(0.032)	(0.032)	(0.034)
Any AI (2024)	-0.035	-0.049*	-0.061**	-0.070**
	(0.022)	(0.023)	(0.023)	(0.024)

Note: The treatment is an indicator for submitting at least one essay classified as AI-written. Columns (1) and (2) include applicant controls; columns (3) and (4) add essay-quality measures. The pooled models include an admissions-cycle indicator, while the 2023 and 2024 estimates are obtained from separate year-specific models. All models use random forest nuisance functions and 5-fold cross-fitting. Across 20 independently seeded fits, coefficients and standard errors are aggregated using DoubleML's median-based repeated-splitting procedure [25]. Standard errors are reported in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.