Fairness Depends on Assessment: Learning by Teaching with Large Language Models

Sydney Miller[0009−0008−7495−054X], Lan Jiang[0009−0004−2764−0697], and Nigel Bosch[0000−0003−2736−2899]

University of Illinois Urbana–Champaign, Urbana, IL 61820, USA
{srm16,lanj3,pnb}@illinois.edu

Abstract. As large language models (LLMs) are increasingly integrated into college-level writing instruction, it is important to examine whether AI-mediated learning outcomes are equitably distributed across students. This study investigates demographic fairness in Learning by Teaching (LbT) with an LLM, in a college-level writing context with 66 undergraduate students. Each student completed both an LbT condition with an LLM-simulated student and a video lecture condition designed to approximate a traditional lecture format. Learning was measured at multiple levels of Bloom's Taxonomy using immediate post-tests, delayed post-tests, and post-lesson essays assessing rhetorical strategy application. Test-based assessments showed overall learning gains from pre- to post-test across conditions, but no statistically significant differences across demographic groups. Essay-based analyses, however, revealed selective instructional effects: female students applied significantly more rhetorical strategies on average under the LbT condition (2.88) than under the video condition (2.24), a difference not reflected in test-based measures. These findings indicate that assessment modality influences conclusions about fair learning outcomes across students.

Keywords: Learning by Teaching · Large Language Models · Educational Fairness.

1 Introduction

Instructional designs and assessment practices may advantage some students while disadvantaging others, even when overall learning outcomes appear positive [10, 13, 15]. As a result, examining specific learning outcomes across different demographic groups has become an important component of fairness-oriented educational research [2, 12]. Educational technological interventions are often evaluated using summary measures of effectiveness such as average learning gains or overall performance improvements based on student models, particularly when comparing instructional conditions [14]. While such measures are informative for identifying general patterns, they can sometimes obscure differences in how students from different demographic and linguistic backgrounds experience and

benefit from certain instruction [10, 13]. The increasing use of artificial intelligence (AI) systems in educational contexts has further amplified the importance of understanding how learning outcomes are distributed across students [11].

As we show in this paper, assessing fairness in learning outcomes requires consideration of how learning is measured. Bloom's Taxonomy provides a useful framework for distinguishing among levels of learning, such as remembering, understanding, and applying knowledge [3, 1]. Immediate post-tests are commonly used to assess recall, delayed assessments capture retention and conceptual understanding, and written assignments can provide evidence of students' ability to apply knowledge in authentic contexts. Importantly, prior research has shown that assessment types (particularly writing-based assessments) may differentially privilege students based on linguistic background [5, 7].

In this paper, we examine demographic fairness in Learning-by-Teaching (LbT) with a large language model (LLM) in the context of college-level writing instruction by comparing test-based and essay-based learning outcomes across instructional conditions. This paper addresses the following research questions:

RQ1: Do learning gains from the LbT condition differ across demographic groups (gender, race/ethnicity, and linguistic background), as measured by pre–post test scores, and how do these patterns compare to the video condition?

RQ2: Do the patterns observed in RQ1 persist when learning is measured using delayed post-test scores, reflecting retention and conceptual understanding?

RQ3: Do the patterns observed in RQ1 and RQ2 persist when learning is measured using essay-based assessment, reflecting application-level learning?

2 Background

As AI systems, and LLMs in particular, become increasingly integrated into educational contexts, researchers have raised growing concerns about fairness and equity in AI-mediated learning environments [12, 11]. Much of the existing work on fairness in AI has focused on algorithmic bias, representational harms, and disparities in system behavior across demographic groups [17]. In educational applications, fairness-oriented research has often examined predictive models, automated assessments, or student modeling approaches used to estimate performance or risk [2]. While this work has been critical for identifying inequities in machine learning models, less empirical research has examined inconsistent treatment among individual students based on their interactions with such systems and how these differences may propagate to affect demographic equity in learning outcomes [4], particularly in open-ended instructional domains such as writing.

LbT offers a useful framework for examining fairness in AI-mediated learning. In LbT, students deepen their understanding of a topic by taking on the role of a teacher, which requires them to explain concepts, organize knowledge, and anticipate misunderstandings [6, 9, 22]. Prior research on technology-mediated LbT, including teachable agent systems such as *Betty's Brain*, has shown that

students can achieve meaningful learning gains by instructing an artificial agent rather than learning directly from it [21, 24]. However, these systems have largely been confined to well-structured STEM domains, where correct answers and misconceptions can be predefined and automatically evaluated, and have rarely examined whether the benefits of LbT are distributed equitably across students from different demographic or linguistic backgrounds.

Advances in LLMs extend the teachable-agent paradigm to language-based and open-ended domains such as writing. LLMs can engage in flexible, dialogue-driven interactions and respond to student explanations about rhetorical concepts, making them well suited for LbT-based writing instruction [16, 23]. Prior work on LLM prompting has shown that these models can be strategically configured to adopt specific interactional roles, including novice-like behaviors [25, 19], creating new opportunities for LbT approaches in domains like writing instruction.

Writing further complicates questions of fairness because it functions both as a site of learning and as a mode of assessment. Writing tasks require students to retrieve, organize, and apply knowledge in rhetorically meaningful ways, making them a rich source of evidence for application-level learning [8, 20]. At the same time, research on writing assessment has shown that evaluation practices may differentially privilege students based on linguistic background and prior educational experiences [5, 7]. As a result, essay-based assessments may surface learning outcomes differently than test-based measures, particularly for students from linguistically diverse backgrounds.

Evaluating fairness in AI-supported writing instruction therefore requires attention to both instructional design and to how learning is measured and represented. While LbT with LLMs offers a promising approach for engaging students actively in college-level writing instruction, it remains unclear whether its benefits are equitably distributed across students or whether commonly used assessment measures capture learning consistently across demographic and linguistic groups. The present study addresses this gap by examining demographic fairness in LbT with an LLM, using both test-based and essay-based measures of learning.

3 Method

3.1 Dataset

The dataset consists of written essays and test-based assessments from 66 undergraduate students who participated in an in-person, within-subjects experiment comparing two instructional conditions: a LbT condition and a video lecture condition [18]. Participants were recruited from undergraduate courses at a large public university in the United States and received course credit for participation. All procedures were approved by a research ethics board.

The dataset includes demographic information (gender, race/ethnicity, and linguistic background), immediate post-test scores, delayed post-test scores, and

two post-activity essays per participant. Each instructional condition targeted three rhetorical strategies drawn from a set of six foundational strategies in persuasive writing (ethos, pathos, logos, concession and rebuttal, analogies, and rhetorical questions). Strategy assignment and order were rotated using a Latin-square design to balance coverage across conditions and mitigate order effects.

After each instructional condition, students completed a short multiple-choice post-test to assess recognition of rhetorical strategies followed by an essay requiring application of the rhetorical strategies introduced in that condition. A delayed post-test covering all six strategies was administered approximately two weeks later to assess retention and conceptual understanding; 31 students completed the delayed post-test.

The test-based and essay-based assessments captured distinct aspects of learning. Test items assessed recognition and interpretation of rhetorical strategies, reflecting recall and conceptual understanding, while essays required students to generate and apply rhetorical strategies in their own writing, reflecting application-level learning. We treated these measures as providing related but non-equivalent evidence of learning outcomes, allowing us to examine whether conclusions about learning and fairness converged across assessment types.

We calculated the descriptive statistics of the dataset. For race, participants were allowed to select one or more categories. There were 40 White students, 14 Asian students, 5 Black students, and 5 students across categories too small to report. Due to sample size constraints and the distribution of responses, we analyzed race as a binary variable (White and non-White) to enable balanced group-level comparisons.

Similarly, for gender, there were 43 female students, 19 male students, 2 non-binary students, and 1 student who preferred not to say. Because no more than three students reported being non-binary or choosing "Prefer not to say" for gender, we excluded these students from further analyses, as such small samples would not allow for group-specific conclusions or comparisons to other groups. For English as first language, 41 students reported "yes" and 25 reported "no".

3.2 Analysis

We analyzed learning outcomes across instructional conditions (LbT and video lecture) and demographic groups defined by gender, race/ethnicity, and first language. All comparisons evaluated differences between instructional conditions within demographic groups.

To address RQ1, we measured immediate learning outcomes as learning gains by subtracting pre-test scores from post-test scores for each student. We computed mean learning gains for each demographic group under each instructional condition and compared these means using t-tests. We analyzed RQ2 in the same way, but with the delayed post-test only. For RQ3, we manually coded essays for two measures: the total number of rhetorical strategies used and the proportion of distinct strategy types applied (i.e., whether all relevant strategies were employed as directed).

We compared performance across instructional conditions using t-tests, reflecting the within-subject design in which each participant completed both conditions. Because multiple comparisons were conducted across demographic groups and outcome measures, results are interpreted with caution and emphasis on consistent patterns rather than isolated statistically significant findings.

4 Results

Students demonstrated overall learning on both test-based and essay-based measures. Test performance showed near-ceiling scores following both instructional conditions (98-99% accuracy), while essay-based outcomes showed greater variability: students used more rhetorical strategies on average in their essays following the LbT condition (2.70 strategies) than the video condition (2.10 strategies). The analyses that follow examine whether these learning gains differed across demographic groups, instructional conditions, and assessment types.

4.1 RQ1: Immediate Learning Gains

As shown in Table 1, no statistically significant differences were observed between the video condition and the LbT condition across demographic groups. The results suggest that immediate learning gains, as measured by immediate post-tests targeting recall and recognition, did not differ meaningfully across demographic groups or instructional conditions.

Table 1. Immediate recall, measured as the difference between pre-test and post-test scores, by demographic groups under different learning strategies

		Passive instruction	**LbT**	***p*-value**
Gender	Female	0.124	0.147	.665
	Male	0.019	0.035	.764
Race	White	0.120	0.138	.749
	Other	0.027	0.056	.537
English as	Yes	0.111	0.119	.860
first language	No	-0.026	0.056	.186

4.2 RQ2: Retention and Conceptual Understanding

RQ2 examined whether the same patterns observed for immediate learning gains (i.e., RQ1) existed when learning was measured using delayed post-test scores.

Table 2. Conceptual understanding, as measured by post-test scores, across demographic groups under different learning strategies

		Passive instruction	**LbT**	***p*-value**
Gender	Female	0.864	0.903	.511
	Male	0.893	0.929	.611
Race	White	0.884	0.897	.822
	Other	0.825	0.925	.145
English as	Yes	0.877	0.920	.370
first language	No	0.800	0.850	.667

As shown in Table 2, no statistically significant differences were observed between instructional conditions across demographic groups for delayed post-test performance. These results indicate that delayed post-test performance, which was intended to capture retention and conceptual understanding, did not differ significantly across demographic groups or conditions.

4.3 RQ3: Rhetorical Strategy Use and Quality in Essays

As shown in Table 3, analysis of the total number of rhetorical strategies used revealed a significant difference between instructional conditions for some groups. Specifically, female students and students who speak English as their first language used a significantly higher number of rhetorical strategies in the essays following the LbT condition compared to the video condition (p=0.030 and 0.031, respectively). No statistically significant differences were observed for male students or across race and students whose first language was not English for this measure. To ensure that this result was not due to female students all being native English speakers, we calculated the correlation between gender and English as first language ($\phi = .207$, $p = 0.101$), indicating that these groups were related but far from identical.

In contrast, no statistically significant differences were observed across demographic groups in the completeness of rhetorical strategy use for either instructional condition.

5 Discussion

Across RQ1 and RQ2, post-test measures of learning showed no statistically significant differences across instructional conditions or demographic groups. These null results indicate that, when learning is measured through short multiple-choice assessments targeting recall and understanding, LbT with an LLM does not produce differential outcomes across gender, race/ethnicity, or linguistic

Table 3. Rhetorical strategy use and rhetorical quality by demographic groups under different learning strategies. $*$ indicates a statistically significant difference ($p < .05$).

		Number of rhetorical strategies			Complete ratio		
		Video Condition	**LbT Condition**	***p*-value**	**Video Condition**	**LbT Condition**	***p*-value**
Gender	Female	2.238	2.881	.030 *	0.635	0.698	.282
	Male	2.368	2.421	.880	0.667	0.684	.824
Race	White	2.314	2.725	.113	0.592	0.692	.077
	Other	2.308	2.769	.296	0.736	0.708	.716
English as	Yes	2.175	2.800	.031 *	0.634	0.693	.261
first language	No	2.542	2.625	.815	0.692	0.718	.789

background in this sample. Importantly, these null effects occurred in the context of overall learning gains, with average post-test accuracy exceeding 98% across conditions, highlighting the limits of test-based measures for capturing instructional differences in writing-focused, AI-mediated contexts.

In contrast, essay-based analyses (RQ3) revealed instructional effects that were not reflected in test scores. Female students who speak English as their first language demonstrated a significant increase in rhetorical strategy application after the LbT condition. This increase in strategy application exhibits evidence of higher-level learning [1, 3]. These findings indicate that assessment type shapes which learning outcomes, and which differences across students, become visible. Test-based and essay-based assessments did not consistently represent learning outcomes across groups, suggesting that reliance on a single measure obscures meaningful differences in how students express and apply what they have learned. From a fairness perspective, these results emphasize the importance of examining learning using multiple forms of assessment when evaluating AI-supported instructional designs.

6 Limitations

This study has several limitations related to sample size and demographic representation that affect the scope of its fairness claims. First, although the sample size is appropriate for an exploratory study, it limits statistical power for subgroup analyses, particularly for intersectional comparisons across demographic categories. As a result, we operationalized certain demographic variables using coarse groupings (e.g., binary race categories), which enabled balanced comparisons but constrained our ability to examine differences among more specific racial and ethnic groups. Consequently, findings related to demographic fairness should be interpreted as preliminary and descriptive rather than conclusive.

Second, the exclusion of participants identifying as non-binary or selecting "Prefer not to say" for gender reflects a broader challenge in fairness research. While this decision was methodologically necessary due to extremely small group sizes, it highlights how standard quantitative approaches can fail to adequately

represent gender diversity. As a result, our analysis does not capture how LbT with LLMs may differentially affect students outside binary gender categories. Future work should prioritize larger and more diverse samples that support more inclusive analyses and avoid the need for such exclusions.

More broadly, these limitations underscore the importance of aligning methodological choices with inclusivity goals when studying fairness in AI-supported learning environments.

7 Conclusion

While traditional, test-based assessments showed no significant differences across demographic groups, essay-based measures revealed differential instructional effects across students that were not captured by tests. These findings demonstrate that null results on standardized measures do not necessarily indicate uniform learning experiences and that assessment type plays a critical role in determining which learning outcomes, and which learners, become visible. As LLMs continue to be integrated into writing instruction in higher education, future research should attend to instructional effectiveness as measured in multiple ways, including how assessment choices shape conclusions about equity and learning.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anderson, L.W., Krathwohl, D.R.: A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition. Addison Wesley Longman, Inc. (2001)
2. Baker, R.S., Hawn, A.: Algorithmic bias in education. International Journal of Artificial Intelligence in Education **32**(4), 1052–1092 (2022)
3. Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., Krathwohl, D.R., et al.: Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain. Longman New York (1956)
4. Chinta, S.V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Quy, T.L., Zhang, W.: FairAIED: Navigating fairness, bias, and ethics in educational ai applications. arXiv preprint arXiv:2407.18745 (2024)
5. Di Gennaro, K., Brewer, M.: Assessing approaches to writing assessment: Considering claims about fairness. Language Teaching pp. 1–13 (2025)
6. Duran, D.: Learning-by-teaching. Evidence and implications as a pedagogical mechanism. Innovations in Education and Teaching International **54**(5), 476–484 (2017)
7. Elliot, N.: A theory of ethics for writing assessment. Journal of Writing Assessment **9**(1) (2016)
8. Emig, J.: Writing as a mode of learning. College Composition & Communication **28**(2), 122–128 (1977)
9. Fiorella, L., Mayer, R.E.: The relative benefits of learning by teaching and teaching expectancy. Contemporary Educational Psychology **38**(4), 281–288 (2013)

10. Gutiérrez, K.D., Rogoff, B.: Cultural Ways of Learning: Individual Traits or Repertoires of Practice. Educational Researcher **32**(5), 19–25 (Jun 2003). https://doi.org/10.3102/0013189X032005019
11. Holstein, K., Doroudi, S.: Equity and artificial intelligence in education: will "AIEd" amplify or alleviate inequities in education? arXiv preprint arXiv:2104.12920 (2021)
12. Kizilcec, R.F., Lee, H.: Algorithmic fairness in education. In: The ethics of artificial intelligence in education, pp. 174–202. Routledge (2022)
13. Kizilcec, R.F., Saltarelli, A.J., Reich, J., Cohen, G.L.: Closing global achievement gaps in MOOCs. Science **355**(6322), 251–252 (Jan 2017). https://doi.org/10.1126/science.aag2063
14. Koedinger, K.R., D'Mello, S., McLaughlin, E.A., Pardos, Z.A., Rosé, C.P.: Data mining and education. WIREs Cognitive Science **6**(4), 333–353 (Jul 2015). https://doi.org/10.1002/wcs.1350
15. Ladson-Billings, G.: From the Achievement Gap to the Education Debt: Understanding Achievement in U.S. Schools. Educational Researcher **35**(7), 3–12 (Oct 2006). https://doi.org/10.3102/0013189X035007003
16. Malik, A., Mayhew, S., Piech, C., Bicknell, K.: From Tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 15670–15693 (2024)
17. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021)
18. Miller, S., Bosch, N.: Learning by teaching an llm: A new approach to writing instruction. International Conference of the Learning Sciences (ICLS) (2026)
19. Miller, S., Bosch, N.: Prompting for teachability: Designing novice personas in llms for learning by teaching contexts. In: LAK26: 16th International Learning Analytics and Knowledge Conference (in press)
20. Nowacek, R.S.: Agents of integration: Understanding transfer as a rhetorical act. SIU Press (2011)
21. Okita, S.Y., Schwartz, D.L.: Learning by teaching human pupils and teachable agents: The importance of recursive feedback. Journal of the Learning Sciences **22**(3), 375–412 (2013)
22. Roscoe, R.D., Chi, M.T.: Tutor learning: The role of explaining and responding to questions. Instructional Science **36**(4), 321–350 (2008)
23. Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., Akata, Z.: In-context impersonation reveals large language models' strengths and biases. Advances in Neural Information Processing Systems **36**, 72044–72057 (2023)
24. Segedy, J.R., Kinnebrew, J.S., Biswas, G.: Using coherence analysis to characterize self-regulated learning behaviours in open-ended learning environments. Journal of Learning Analytics **2**(1), 13–48 (2015)
25. White, J.: A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv preprint arXiv:2302.11382 (2023)