










# Evaluating an AI Tutor for Bias Across Different Foundation Models

Aditya Vinodh<sup>1</sup>, Emma Harvey<sup>1</sup>, Husni Almoubayyed<sup>2</sup>,  
Renzhe Yu<sup>3</sup>, Christopher Brooks<sup>4</sup>, Allison Koenecke<sup>1</sup>,  
and Rene F. Kizilcec<sup>1</sup>

<sup>1</sup> Cornell University, Ithaca, NY, USA

{av364, evh29, koenecke, kizilcec}@cornell.edu

<sup>2</sup> Carnegie Learning, Pittsburgh, PA, USA

halmoubayyed@carnegielearning.com

<sup>3</sup> Columbia University, New York, NY, USA

renzheyu@tc.columbia.edu

<sup>4</sup> University of Michigan, Ann Arbor, MI, USA

brookschr@umich.edu

**Abstract.** AI tutors are increasingly deployed to diverse groups of learners, raising the need to provide high-quality responses independent of the identity of learners who use them. We present a collaborative audit that assesses whether LIVEHINT AI, a large language model-based AI tutor that is currently under development by Carnegie Learning, meets this goal. We repeatedly prompt LIVEHINT AI with realistic student queries modified to include explicit or implicit statements of identity; e.g., identifying as a particular nationality or writing in a particular dialect. We then assess the responses based on their tone and level of detail. By evaluating different versions of LIVEHINT AI powered by GPT-4, GPT-4o, and Claude-3.5-Sonnet, we found that the choice of foundation model impacts the level of differentiation in responses. This differentiation may reflect pedagogical strategies (e.g., reducing text complexity when observing typos) or it may be undesirable (e.g., responding to an English prompt in a different language). Education researchers can use this approach to select foundation models that best fit their pedagogical approach, and build guardrails around potentially biased, inconsistent, or undesired behavior.

**Keywords:** AI Tutor · LLM · AI Audit · Algorithmic Fairness · Dialect Bias

## 1 Introduction

Large language model (LLM)-based AI tutors are increasingly being used to provide personalized instruction to learners at scale, with the goals of increasing student engagement and academic performance [20]. However, there is a risk that they may inadvertently reinforce or amplify systemic biases in education by

providing unequal levels of assistance to different groups of students [9, 15, 23]. To measure—and facilitate the mitigation of—such risks, researchers and edtech developers can conduct audits of LLM-based AI tutors [17]. As a template for such audits, we present a collaborative audit of LIVEHINT AI, an LLM-based AI tutor that is currently being developed by Carnegie Learning.<sup>1</sup> Our audit is intended to address two key research questions: (1) do semantically equivalent queries containing different statements of identity receive similar responses?; and (2) does the choice of foundation model impact the level of response differentiation? To do this, we repeatedly prompted versions of LIVEHINT AI powered by different foundation models (GPT-4, GPT-4o, and Claude 3.5 Sonnet) with realistic student queries modified to include statements of identity. Those statements could be explicit (i.e., identifying as a particular nationality or ethnicity) or implicit (i.e., writing in a particular dialect or with a particular level of formality). Then, employing lexical metrics commonly used in Natural Language Understanding, we measured whether responses were consistent in tone and level of detail. Overall, while we found no evidence of harmful biases, we found that the choice of foundation model impacts the level of response differentiation. These findings highlight trade-offs between consistency and adaptability in LLM-based edtech systems—there is no single “fairest” model, but some models may align better than others with specific pedagogical goals. Our research contributes (1) actionable insights on foundation model-specific tendencies for educational stakeholders, and (2) a template for conducting future audits of LLM-based AI tutors.

**Related Work.** A growing body of research indicates that LLMs can produce biased outputs [4, 6, 10, 13, 19]. However, despite calls for domain-specific evaluations of AI-based edtech systems, there remains a lack of systematic audits specifically targeting AI tutors [9, 15]. Recently, Harvey et al. [8] proposed a five-step, domain-agnostic framework for auditing LLM-based chatbots for dialect-based quality-of-service harms. Their framework involves selecting a target chatbot, collecting realistic user prompts, perturbing prompts across dialects of interest, prompting the target chatbot and measuring domain-specific components of response quality, and then comparing response quality across dialects. We largely follow this approach, making education-specific modifications throughout.

## 2 Audit Approach

**LIVEHINT AI.** LIVEHINT AI<sup>2</sup> [5] is an LLM-powered interactive chat system by Carnegie Learning in the research and development phase (and currently released to a small number of school districts across the US). LIVEHINT AI has curriculum-specific instruction to provide step-by-step guidance towards solving math problems, and explain concepts when a student makes mistakes, in a way

<sup>1</sup> *Collaborative audits* are conducted by external auditing teams with cooperation from the audit target [14, 24].

<sup>2</sup> <https://discover.carnegielearning.com/livehint-ai>.

consistent with Carnegie Learning’s core curriculum products. LIVEHINT AI can provide analogies and examples and re-frame problems, but does not provide students with final answers. LIVEHINT AI uses an agentic approach [3] to apply a set of guardrails, including toxicity detection using a BERT-based model [7] and rejection of off-task behavior. Prior to this project, the LIVEHINT AI system did not include instructions on what to do when prompted with explicit or implicit demographic identifiers—this study is used to gauge the need of such guardrails and the behavior of different underlying foundation models in such cases.

**Prompt Collection.** To curate a set of realistic student prompts, we drew from the Conversation-Based Math Tutoring Accuracy (CoMTA) dataset [16]. CoMTA includes 188 conversations between students and the Khanmigo chatbot, spanning elementary math to calculus. We selected the subset of student questions that were long enough to be modified to display dialectical and other kinds of linguistic variation. To ensure that prompts were not specific to a given math problem, we identified common question roots (e.g., “I don’t understand”) and wrote new prompts based on those roots. This resulted in four baseline prompts: (1) *I don’t understand this concept.* (2) *Can you help me with this problem?* (3) *I need assistance with my approach.* (4) *This question is confusing.*

**Prompt Perturbation.** We varied the prompts across three facets: stated nationality or ethnicity, dialect, and level of formality. We include examples of baseline and modified prompts in the OSF Online Appendix.<sup>3</sup>

*Stated Identity: Nationality and Ethnicity.* We prepended an explicit statement of nationality or ethnicity to the baseline prompts: “My name is [name] and I am [nationality/ethnicity].” We included the following nationalities and ethnicities: Arabic, Chinese, English, Hawaiian, Indian, Italian, Slavic, and Spanish. We selected this set because it includes groups (a) who are likely to be well-represented in the pool of US learners, and (b) about whom prior research has shown that foundation models can produce text of varying sentiment [18]. We used GPT-4o to select stereotypical names, balancing male and female names.

*Perceived Identity: Dialect.* We modified each of the baseline prompts by ‘translating’ it across different English dialects.<sup>4</sup> To do this, we leveraged Multi-VALUE [26], a rule-based translation system that is built using data on linguistic features and their prevalence in English dialects as determined by a team of trained linguists [12]. We translated each baseline prompt into Colloquial American English<sup>5</sup> (CollE) and Indian English<sup>6</sup> (IndE). For example, “I

<sup>3</sup> <https://osf.io/g8tpb/>.

<sup>4</sup> We did not translate across languages because, at the time of the audit, LIVEHINT AI only has native support for English, with other languages still under development.

<sup>5</sup> <https://ewave-atlas.org/languages/14>.

<sup>6</sup> <https://ewave-atlas.org/languages/52>.

need assistance with my approach.” was transformed into “I am needing assistance with my approach.” (CollE) and “With my approach need the assistance.” (IndE). These statements are semantically equivalent to the original, but represent different ways of speaking that LLM-based systems may implicitly associate with particular identities. We selected this set because it includes dialects that are likely to be well-represented in the pool of US learners. We manually validated the ‘translated’ prompts to ensure that they preserved the meaning of the originals.

*Perceived Identity: Language Formality.* Finally, we progressively decreased prompt formality. We prompted GPT-4 to iteratively introduce misspellings, grammatical errors, and slang to the baseline prompts. Again, we manually validated the perturbed prompts to ensure that they preserved the meaning of the originals.

**Response Collection.** The perturbations above produced 60 distinct prompts. To ensure the robustness of our analysis, we prompted LIVEHINT AI with each variation 25 times. We repeated this process for each of the three foundation models that Carnegie Learning was considering using as part of LIVEHINT AI: GPT-4 (20230613), GPT-4o (20240806), and Claude 3.5 Sonnet (20240620).

**Response Evaluation.** To evaluate the quality of LIVEHINT AI’s responses, we considered (1) their level of detail and (2) the appropriateness of their tone. We evaluated level of detail by measuring length (words) as well as the percentage of bolded words (bolding was typically done to highlight key concepts). We evaluated tone by considering the percentage of non-English words; the percentage of affective, informal, and collective (“we”) language measured using LIWC [21]; FleschKincaid readability score [11]; text complexity as measured by the percentage of words with more than two syllables, and lexical diversity as measured by the ratio of unique word stems to total word count [2]. We measured whether responses varied statistically significantly according to each metric across the nationality/ethnicity, dialect, and formality of prompts using ANOVA, correcting for multiple comparisons using the Benjamini-Hochberg [1] method. We report effect size using Eta-squared ( $\eta^2$ ), which provides a measure of the proportion of variance explained by each factor in an ANOVA.

### 3 Results

We provide an overview of our findings in Table 1 and a comprehensive set of boxplots in the OSF Online Appendix.<sup>7</sup> Overall, we find that Claude 3.5 had the most differentiation, followed by GPT-4, with GPT-4o having the least. Notably, responses produced by the different underlying models differ from one another even when given only the baseline prompts, suggesting that there is less metric

<sup>7</sup> <https://osf.io/g8tpb/>.

variance arising from changing input prompts (by nationality/ethnicity, dialect, or formality) relative to the high variance arising from simply changing the choice of underlying foundation model.

**Stated Identity: Nationality and Ethnicity.** All versions sometimes respond to prompts that are written in English and that identify the prompter as having a particular nationality or ethnicity *in the language associated with that nationality or ethnicity*. In most cases, LIVEHINT AI includes a non-English greeting (e.g., “Aloha!” in response to a self-identified Hawaiian prompter) followed by English text. However, in response to prompts identifying the prompter as Spanish or Italian, all versions sometimes provided responses written entirely in Spanish or Italian, respectively. This effect was most pronounced in the Claude-based version ( $\eta^2 = 0.274$ ,  $p < 0.01$ ), which consistently generates non-English words in response to all stated nationalities/ethnicities. This may represent undesirable differentiation, depending on whether students submitting prompts in English will expect responses in English.

**Perceived Identity: Dialect.** No version of LIVEHINT AI produced responses that varied meaningfully in tone or level of detail based on the dialect a prompt was written in. This indicates that, at least for this set of prompts, LIVEHINT AI does not appear to display dialect bias against speakers of ColLE or IndE.

**Perceived Identity: Formality.** The Claude 3.5 version shows the largest differentiation to the level of formality in the prompt. In particular, it shows significant increases in response length ( $\eta^2 = 0.195$ ,  $p < 0.01$ ) and readability

**Table 1.** Results of our audit, expressed in terms of  $\eta^2$  effect size. Large effect sizes ( $> 0.14$ ) are bolded; statistically significant differences are indicated with asterisks (\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ ). **Baseline** results show *across-model variation* in responses to the baseline prompts. **Nationality/Ethnicity**, **Dialect**, and **Formality** results show *within-model variation* in responses to prompts that vary according to stated nationality/ethnicity, dialect, and formality, respectively.

	Baseline	Nationality/Ethnicity			Dialect			Formality		
	All Models	GPT-4	GPT-4o	Claude	GPT-4	GPT-4o	Claude	GPT-4	GPT-4o	Claude
Length (words)	0.106***	0.025*	0.018	0.050**	0.005	0.001	0.054**	0.035*	0.023	<b>0.195**</b>
% Non-English Words	0.049***	0.069**	0.046	<b>0.274**</b>	0.007	0.000	0.104**	0.000	0.000	0.097**
% Bolded Words	<b>0.194***</b>	0.012	0.049	0.022*	0.046*	0.003	0.030*	0.064**	0.010	0.137**
Affect (LIWC)	<b>0.620***</b>	0.010	0.047	0.030**	0.014	0.019	0.012	0.037*	0.007	0.087**
Informality (LIWC)	0.088***	0.017	0.014	0.047**	0.003	0.019	0.015	0.006	0.012	0.134**
Collectivity (LIWC)	<b>0.225***</b>	0.026*	0.019	0.071**	0.039*	0.004	0.039*	0.012	0.014	0.068*
Flesch-Kincaid	<b>0.391***</b>	<b>0.250**</b>	0.112***	0.087**	0.001	0.014	0.055**	0.005	0.003	<b>0.174**</b>
Text Complexity	<b>0.707***</b>	<b>0.385***</b>	0.025	0.132**	0.001	0.026	0.107**	0.010	0.004	0.078**
Lexical Diversity	<b>0.724***</b>	0.009	0.014	0.057**	0.020	0.007	0.025	0.051**	0.019	0.025

( $\eta^2 = 0.178$ ,  $p < 0.01$ ) as formality decreases. We hypothesize that it may be adapting to informal language by providing increased explanation and simpler language in its responses. Claude 3.5 sometimes explicitly corrected typos from the student prompt, increasing the response length.

## 4 Discussion and Conclusion

Overall, our audit did not find evidence that LIVEHINT AI displays dialect bias against prompts written in ColLE or IndE. However, we did find some differentiation in responses, mostly between differing foundation models, associated with the stated (nationality/ethnicity) or perceived (level of prompt formality) identity of a prompter. While some may consider such differentiation as an instance of personalization, it might be undesired in other situations. For example, a student who submits prompts with multiple typos may benefit from responses that are less complex or more readable. However, some differentiation is likely not desired; for example, an Italian student who writes a prompt in English may prefer to receive a response in English instead of Italian.

**Limitations.** The range of prompts used to evaluate LIVEHINT AI was limited and may not fully represent the diversity of student queries. This constraint arose from our need to use sufficiently long prompts to ensure variation, though future studies could expand this scope for greater representativeness. Our audit was conducted on a math-specific tutor; and results might be different in other subject areas that involve increased cultural context, such as history or social studies [22]. Additionally, we considered only a small set of dialects, which does not fully reflect real-world educational settings. We also did not consider multi-turn conversations in this audit. Finally, we note that collaborative audits are sometimes criticized as instances of corporate capture [25]. In our case, Carnegie Learning provided our research team with an API key and imposed no restrictions on our methodology or the results we present here.

**Future Work.** Edtech providers can build on this audit template to assess their AI tutors for bias, and to systematically select the most suitable foundation model for their use case and pedagogical goals, balancing consistency and adaptability. As tools like LIVEHINT AI play an increasingly important role in providing students with flexible, on-demand support, their design choices can help to bridge gaps and promote educational equity. Future research should expand this methodology to capture broader linguistic and cultural factors. For example, extensions of this audit might focus on responses not automatically translated but written by a diverse group of actual learners. We believe this approach should extend to considering languages as an additional category to detect differentiation, and not just dialects.

**Acknowledgments.** This work was supported by funding from the Learning Engineering Virtual Institute (LEVI) to Carnegie Learning and through the Fairness Analysis and Transfer Learning Hub; additionally, EH and AK were supported by Apple, Inc. Any views, opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as reflecting the views, policies or position, either expressed or implied, of Apple Inc.

## References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B (Methodological)* **57**(1), 289–300 (1995)
2. Benoit, K., et al.: Quanteda: an R package for the quantitative analysis of textual data. *J. Open Source Software* **3**(30), 774 (2018). <https://doi.org/10.21105/joss.00774>, <https://quanteda.io>
3. Christie, S.T., Rafferty, A.N., Lee, Z., Cutler, E., Tian, Y., Almoubayyed, H.: An agentic framework for real-time pedagogical plot generation. In: *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*. Springer (2025)
4. Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J., Chang, K.W.: Harms of gender exclusivity and challenges in non-binary representation in language technologies. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, DR, pp. 1968–1994. ACL (2021), <https://aclanthology.org/2021.emnlp-main.150/>
5. Fisher, J., Almoubayyed, H., Fancsali, S.E., Ritter, S., Ley, L.D., Lee, Z.: Building an instructional design-backed, GPT-driven AI tutor for math homework support. In: *AI for Education: Bridging Innovation and Responsibility at the 38th AAAI Annual Conference on AI* (2024). <https://openreview.net/forum?id=NF1q7Xihrd>
6. Fleisig, E., Smith, G., Bossi, M., Rustagi, I., Yin, X., Klein, D.: Linguistic bias in ChatGPT: language models reinforce dialect discrimination. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, pp. 13541–13564. ACL (2024). <https://aclanthology.org/2024.emnlp-main.750/>
7. Hanu, L.: Unitary team: Detoxify (2020). <https://github.com/unitaryai/detoxify>
8. Harvey, E., Kizilcec, R.F., Koenecke, A.: A framework for auditing chatbots for dialect-based quality-of-service harms. In: *The 2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Athens, Greece (2025). <https://doi.org/10.1145/3715275.3732137>
9. Harvey, E., Koenecke, A., Kizilcec, R.F.: “don’t forget the teachers”: towards an educator-centered understanding of harms from large language models in education. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM (2025). <https://doi.org/10.1145/3706598.3713210>
10. Hofmann, V., Kalluri, P.R., Jurafsky, D., King, S.: AI generates covertly racist decisions about people based on their dialect. *Nature* **633**(8028), 147–154 (2024). <https://doi.org/10.1038/s41586-024-07856-5>
11. Kincaid, J.P., Fishburne, J., Robert P., R., Richard L., C., Brad, S.: Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, Defense Technical Information Center, Fort Belvoir, VA (1975). <https://apps.dtic.mil/sti/citations/tr/ADA006655>

12. Kortmann, B., Lunkenheimer, K., Ehret, K. (eds.): eWAVE (2020). <https://ewave-atlas.org/>
13. Kotek, H., Dockum, R., Sun, D.: Gender bias and stereotypes in large language models. In: Proceedings of The ACM Collective Intelligence Conference. CI '23, New York, NY, USA. ACM (2023), <https://doi.org/10.1145/3582269.3615599>
14. Lam, K., Lange, B., Blili-Hamelin, B., Davidovic, J., Brown, S., Hasan, A.: A framework for assurance audits of algorithmic systems. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency. FAccT '24, pp. 1078–1092. ACM (2024). <http://dx.doi.org/10.1145/3630106.3658957>
15. Lee, J., Hicke, Y., Yu, R., Brooks, C., Kizilcec, R.F.: The life cycle of large language models in education: a framework for understanding sources of bias. *Br. J. Edu. Technol.* **55**(5), 1982–2002 (2024)
16. Miller, P., Dicerbo, K.: LLM based math tutoring: challenges and dataset (2024). <https://osf.io/preprints/edrxiv/5zwv3.v1>
17. Mokander, J., Schuett, J., Kirk, H., Floridi, L.: Auditing large language models: a three-layered approach. *AI Ethics* **4**, 1085–1115 (2023). <https://doi.org/10.1007/s43681-023-00289-2>
18. Narayanan Venkit, P., Gautam, S., Panchanadikar, R., Huang, T.H., Wilson, S.: Nationality bias in text generation. In: Proceedings of the 17th Conference of the European Chapter of the ACL, Dubrovnik, Croatia, pp. 116–122. ACL (2023). <https://aclanthology.org/2023.eacl-main.9/>
19. Sheng, E., Chang, K.W., Natarajan, P., Peng, N.: The woman worked as a babysitter: On biases in language generation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, pp. 3407–3412. ACL (2019). <https://aclanthology.org/D19-1339/>
20. Singer, N.: Will chatbots teach your children? The New York Times (2024). <https://www.nytimes.com/2024/01/11/technology/ai-chatbots-khan-education-tutoring.html>
21. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* **29**(1), 24–54 (2010). <https://doi.org/10.1177/0261927X09351676>
22. Veselovsky, V., et al.: Localized cultural knowledge is conserved and controllable in large language models (2025). <https://arxiv.org/abs/2504.10191>
23. Williamson, B., Molnar, A., Boninger, F.: Time for a pause: Without effective public oversight, AI in schools will do more harm than good. Technical report, National Education Policy Center, Boulder, CO, USA (2024). <http://nepc.colorado.edu/publication/ai>
24. Wilson, C., et al.: Building and auditing fair algorithms: a case study in candidate screening. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, New York, NY, USA, pp. 666–677. ACM (2021), <https://doi.org/10.1145/3442188.3445928>
25. Young, M., Katell, M., Kraft, P.: Confronting power and corporate capture at the FACCT conference. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT '22, New York, NY, USA, p. 1375–1386. ACM (2022). <https://doi.org/10.1145/3531146.3533194>
26. Ziems\*, C., Held\*, W., Yang, J., Dhamala, J., Gupta, R., Yang, D.: Multi-VALUE: A framework for cross-dialectal English NLP. In: Proceedings of the 61st Annual Meeting of the ACL (Volume 1: Long Papers), Toronto, Canada, pp. 744–768. ACL (2023). <https://aclanthology.org/2023.acl-long.44>