Evaluating and Interpreting Gender Bias in LLM Feedback: Span-Level Embedding-Based Evidence from Automated Essay Feedback

Yishan DU[1,*,†], Maria Perez Ortiz[2,‡] and Mutlu Cukurova[1,2,†]

[1]UCL Knowledge Lab, Institute of Education, University College London, UK
[2]UCL Centre for Artificial Intelligence, the Faculty of Engineering, University College London, UK

Abstract

Large language models (LLMs) are used to generate feedback on student writing at a scale; concerns are growing that they may reproduce or amplify gender bias in pedagogically consequential ways. Recent embedding-based benchmarking work has shown that counterfactual gender cues can produce significant semantic divergence in LLM-generated essay feedback, especially under implicit gender conditions. However, such bias is difficult to evaluate because it increasingly emerges through subtle, context-dependent differences in tone, questioning, evaluation, revision guidance, and learner positioning. Thus, a key challenge remains: how can such bias be localised, interpreted, and connected to its downstream pedagogical meaning? This paper addresses this challenge by proposing a span-level embedding-based evaluation framework for analysing gender bias in LLM-generated essay feedback. Using 300 essays from the AES corpus, this work analysed 600 feedback responses generated by GPT-4o mini under an original male-associated condition and a male-to-female counterfactual condition. Feedback responses are segmented into local spans, embedded, and aligned across counterfactual pairs using cosine similarity. The study then estimates each matched span pair's contribution to global cross-condition semantic separation through a leave-one-out cosine influence statistic and assesses significance using one-sided permutation tests. Analysis of the significant span pairs (n=217, p<.05) reveals a systematic shift in pedagogical framing: (1) male-associated feedback is longer, more evaluative, and more strategy-oriented, focusing on argument, organisation, historical context, proofreading, and coherence; (2) counterfactual female feedback is shorter, more interrogative, and more focused on experiential, relational, and affective details. We argue that this pattern represents pedagogical framing bias, where gender cues influence not only what feedback says, but what kinds of learning opportunities it provides. This study contributes an interpretable NLP-based method for connecting embedding-level bias detection with pedagogically sensitive evaluation of automated feedback.

Keywords

LLMs, writing feedback, gender bias, span-level analysis

1. Introduction and Related Work

Large language models (LLMs) are increasingly being used to generate feedback on students' written work [1]. Yet, concerns have intensified over whether it provides equitable learning opportunities, especially when conditioned by gendered cues. Increasing evidence indicates that LLMs generate different content for different learner groups even when all other input is constant [2]; and gender cues can shape model behaviour in scenarios such as collaboration analytics and essay assessment [3, 4]. In LLM-generated feedback for students' writing, LLMs systematically shift the output according to gender cues, and positive feedback bias and feedback withholding bias showed on how student writing is judged and how much substantive critique students receive [5]; Most directly, Du et al. [6] showed that implicit gender manipulations in student essays produced significant semantic divergence across all tested models, with male-coded cues eliciting more autonomy-supportive guidance and female-coded cues more often eliciting controlling. This biased feedback will shape students' self-regulation, motivation, and agentic engagement if it is not provided with appropriate pedagogical discretion [7, 8, 9].

The 2026 Workshop on Pedagogical Evaluation of Automated Feedback, AIED 2026, Seoul, Republic of Korea
*Corresponding author.
yishan.du.24@ucl.ac.uk (Y. DU)
0009-0000-1116-659X (Y. DU); 0000-0003-1302-6093 (M. P. Ortiz); 0000-0001-5843-4854 (M. Cukurova)


LLM bias increasingly appears in implicit and context-dependent forms. Even models that appear explicitly unbiased may retain implicit biased associations [10]. In educational feedback, bias is unlikely to appear as overtly harmful language and tends to be embedded in apparently helpful feedback through assumptions about voice, agency, ability, style, and authorial identity [11, 5, 6]. Such bias emerges from sociotechnical pipelines in which training data, optimisation, prompt contexts, and human-AI interaction jointly shape outputs [12, 13, 14]. In education, these patterns interact with long-standing gendered assumptions in teaching and assessment [15, 16]. As a result, fluent and pedagogically plausible feedback may still redistribute agency, critique, encouragement, and writing support in inequitable ways [11, 17]. This implicitness creates a methodological challenge for evaluation. Bias measures detached from downstream use contexts may be poorly aligned with the harms they aim to capture [18, 19]. In feedback, the same linguistic form can serve different pedagogical roles: a question may invite reflection or impose an external revision agenda; praise may support confidence or withhold critique; and a suggestion may preserve or constrain learner agency [20, 21, 8, 9]. Manual interpretation can reveal such meaning, but it is difficult to scale or embed into computational auditing workflows.

Span-level analysis offers one such route. We use *span* to refer to a contiguous, semantically coherent local unit of text that is more interpretable than tokens while remaining more localisable than a full response [22, 23]. In broader LLM research, span-level methods have been used to localise hallucinations [22], evaluate machine translation errors [24], detect AI-generated scientific text [25], and audit socially complex language phenomena such as severity-dependent bias in polarising language detection [26]. These studies indicate a shift from global scores to local evidence, and from detection alone to detection, localisation, calibration, and interpretation. This trajectory is especially relevant to bias evaluation, yet remains underused in AIED research on bias.

This study therefore proposes span-level embedding analysis to localise interpretable evidence of bias. By aligning spans across counterfactual feedback pairs and estimating each span pair's contribution to global semantic separation, the method identifies where gender-conditioned divergence occurs and supports educational interpretation of how such divergence is realised. We ask:

RQ1. Which spans drive gender-conditioned feedback divergence?

RQ2. What linguistic and discourse patterns distinguish these spans?

2. Methods

2.1. Data and counterfactual feedback generation

We used 300 essays from the AES corpus that contained male-associated words. For each essay, we used two conditions: the original male-associated essay condition (*M*) and a male-to-female counterfactual condition (*M-F*), where gender-associated words were substituted while keeping the remaining essay content unchanged. All essays from the counterfactual condition are manually checked to ensure grammar and semantic consistency. GPT-4o mini was used to generate feedback for each essay in both conditions, resulting in 600 feedback responses.

2.2. Span construction and analysis

2.2.1. Span segmentation and alignment

Each feedback response was segmented into local spans. We define a span as a contiguous, semantically coherent unit of feedback text that is more interpretable than a token-level unit while remaining more localisable than a full response [22, 23]. Feedback on writing is commonly organised through pedagogical discourse moves such as praise, critique, diagnosis, questioning, and revision guidance [20, 21]. Sentences therefore, provide a meaningful first unit for preserving feedback function, and we thus used a sentence-first strategy accordingly. Long sentences were split into overlapping chunks of up to 30 words with an 8-word overlap, balancing local interpretability with contextual continuity, consistent with sliding-window span analysis for localising coherent evidence in LLM outputs [22]. The 30-word threshold was chosen as a rounded upper bound close to the 32-word span setting explored

in recent span-level LLM work [27]. The 8-word overlap was used as a local-context buffer to reduce boundary artefacts, inspired by prior NLP and distributional-semantic work that uses 8-word windows to capture local semantic or syntactic context [28], while recognising that window size is task-dependent [29].

For each counterfactual feedback pair, spans from the *M* response and the corresponding *M-F* response were embedded and aligned using greedy one-to-one cosine matching. Let $A_i = \{a_{i1}, ..., a_{im}\}$ denote the span embeddings from the *M* feedback for essay *i*, and $B_i = \{b_{i1}, ..., b_{in}\}$ denote the span embeddings from its *M-F* counterpart. The greedy one-to-one cosine alignment is applied to prevent generic feedback spans from being repeatedly matched to multiple counterfactual spans, which could inflate local evidence [30]. Cosine-based embedding similarity was used because it aligns spans by semantic proximity, consistent with prior work showing that sentence and contextual embeddings capture meaning beyond exact lexical matching [31, 32].

We computed all pairwise cosine similarities between A_i and B_i, then iteratively selected the highest-similarity unused span pair.

2.2.2. Span-level contribution to global semantic separation

All matched span pairs were pooled across essays. Let $A = \{a_1, \dots, a_N\}$ be the embeddings of aligned spans from the *M* condition and $B = \{b_1, \dots, b_N\}$ the corresponding embeddings from the *M-F* condition. We first computed the global cross-condition cosine separation:

$$T = \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} d_{\cos}(a_j, b_k),$$

where $d_{\cos}(\cdot, \cdot)$ is cosine distance. To estimate the contribution of each matched span pair *i*, we removed a_i and b_i, recomputed the global separation, and defined:

$$T_{-i} = \frac{1}{(N-1)^2} \sum_{\substack{j=1 \\ j \neq i}}^{N} \sum_{\substack{k=1 \\ k \neq i}}^{N} d_{\cos}(a_j, b_k),$$

$$I_i = T - T_{-i}.$$

Here, I_i is the span-level cosine influence statistic. A larger positive I_i indicates that the matched span pair contributes more strongly to global semantic separation.

2.3. Permutation-based significance testing

To assess whether each span pair's influence exceeded what would be expected under random cross-condition alignment, we used a one-sided permutation test. In each permutation, the *M-F* span embeddings were randomly reassigned while preserving the *M* spans and the overall embedding distributions. We then recomputed the leave-one-out influence statistic for each span pair $B = 2000$.

2.4. Interpretive analysis of significant spans

The 217 significant span pairs ($p<.05$) were analysed to characterise how gender-conditioned semantic divergence was locally realised in feedback language.

The interpretive analysis of the significant span pairs followed a reflexive thematic analysis approach [33]. The first author conducted the primary coding of the retained span pairs, moving iteratively between the M and M-F spans, quantitative indicators, and the surrounding feedback context. Initial codes were generated inductively from repeated close reading of the significant span pairs, then organised into candidate themes. These candidate themes were reviewed and refined through discussion with the co-authors through focused meetings. Following reflexive thematic analysis, we establish coding consensus.

Table 1
Summary of the 217 high-influence span pairs retained for interpretation.

Surface and linguistic patterns	M	M-F
Mean words per span	12.60	9.05
Question-form spans	17 (27.0%)	41 (65.1%)
Statement spans	42 (66.7%)	21 (33.3%)
Exclamation spans	3 (4.8%)	1 (1.6%)
Imperative spans	1 (1.6%)	0 (0.0%)
Writing-quality markers	17 (27.0%)	2 (3.2%)
Experiential/relational markers	24 (38.1%)	33 (52.4%)
Praise/encouragement markers	15 (23.8%)	2 (3.2%)
Score/overall-label markers	2 (3.2%)	2 (3.2%)

The interpretive analysis focused on three dimensions aligned with the research questions: (1) linguistic form, including question forms, evaluative statements, and suggestion structures; (2) discourse moves, including critique, praise, strategy guidance, content prompting, and elaboration requests; and (3) pedagogical meaning, including learner agency, revision guidance, motivation, and equitable access to writing support.

3. Results and Findings

217 high-influence span pairs with one-sided permutation values below .05 were identified. As shown in Table 1, the selected spans had a median permutation value of .0045, a median permutation z value of 3.58, and a median cosine influence of 4.66×10^{-5}.

Three main patterns emerged. First, *M* feedback spans were longer and more often realised as evaluative statements, whereas *M-F* spans were shorter and more often realised as questions. In the retained subset, 65.1% of *M-F* spans were question forms, compared with 27.0% of *M* spans. Second, the two conditions differed in their writing focus. The *M* spans more often targeted essay quality, including argument depth, organisation, historical context, proofreading, coherence, and broader thematic development. By contrast, the *M-F* spans more often prompted experiential, relational, or affective elaboration, including animals, feelings, interactions, challenges, lessons learned, and personal perspective. Third, the high-influence spans indicated a shift in feedback function. The *M* condition more often framed the student as a writer whose work required diagnosis, strategy, and development. E.g., one *M* span stated that *"the essay jumps from one idea to another without always making clear connections,"* while the corresponding *M-F* span asked, *"What did she learn from her adventures?"* Similarly, an *M* span recommended proofreading before submission, whereas the corresponding *M-F* span asked about the impact on potential members. We presented representative high-influence span pairs in Table 2.

4. Discussion and Future Work

This study demonstrates the value of span-level embedding analysis for evaluating bias in LLM-generated feedback. By estimating each matched span pair's contribution to global cross-condition separation, our method localises the textual evidence that most strongly drives gender-conditioned divergence. This makes bias more inspectable and identifies specific feedback spans that can be reviewed, interpreted, and potentially surfaced to educators or system designers in auditing workflows.

The identified spans reveal that gender-conditioned divergence appeared as a shift in feedback function. Male-associated feedback was more often evaluative, strategy-oriented, and focused on essay quality, including argument development, coherence, historical context, proofreading, and broader thematic connections. In contrast, female counterfactual feedback was often interrogative, experiential,

Table 2
Representative high-influence span pairs.

M-condition feedback span	M–F-condition feedback span	Evidence
"You provide a glimpse into Luke's experiences and his decision to continue sailing instead of joining the military."	*"How did she handle the decision of not joining the military?"*	$p = .0005$, $z = 4.83$
"Moreover, when you mention the formation of the UNRRA, it might be useful to explain why that context is important to understand Luke's journey."	*"Did she have any memorable interactions with people in these different countries?"*	$p = .001$, $z = 3.98$
"The moment when Luke breaks his ribs is significant, but it needs more context—how does this incident affect him and his view of the journey?"	*"What does breaking her ribs signify in the larger narrative?"*	$p = .002$, $z = 3.82$
"It would strengthen your piece if you could tie Luke's experiences more closely to this lesson throughout the narrative."	*"What did the animals feel like telling from Luke's experiences, for example?"*	$p = .0015$, $z = 4.88$

Note. Span pairs are representative high-influence excerpts from M versus M–F feedback comparisons. Evidence statistics report the corresponding span-level test results.

and relational, prompting elaboration about feelings, interactions, adventures, challenges, and personal perspective. These patterns align with concerns that bias in educational LLMs may be embedded in apparently helpful pedagogical language rather than in overtly discriminatory statements [11, 5].

We interpret this pattern as a form of *pedagogical framing bias*, concerning that counterfactual gender cues appear to redistribute different types of learning opportunities. In this study, semantic difference is interpreted as pedagogically meaningful only when it is counterfactual gender-conditioned, recurrent across significant spans, and linked to feedback functions that shape learning opportunities. The identified spans showed a systematic shift in feedback framing: M feedback more often offered writing-quality diagnosis and strategy-oriented guidance, whereas M-F feedback more often used question prompts about experience, relation, and affect. This distinction matters because effective feedback is not simply information delivery; it should help learners understand where they are going, how they are progressing, and what actions can close the gap [7]. It also depends on students' feedback literacy, including their capacity to interpret feedback, make judgment, and act on it productively [8].

From this perspective, repeated differences in whether feedback supports strategic revision or instead asks students to respond to local prompts may affect learner agency and feedback uptake, which are central to agency engagement with feedback [9]. We therefore interpret the observed pattern as pedagogical framing bias: because gender cues appear to redistribute feedback functions and learner positioning in pedagogically consequential ways, consistent with prior response-level evidence of gender-conditioned divergence in LLM feedback [6].

This study is an initial step. Future work should extend this approach in two directions. First, developing standardised auditing pipelines that could support feedback developers and educators by flagging high-influence spans and explaining the differences based on pedagogical impacts. Second, future research should involve teachers in the evaluation loop. Teachers may accept, adapt, ignore, or override LLM-generated feedback in different ways. Teacher-in-the-loop studies can examine how educators interpret span-level audit evidence, whether it changes their use of LLM feedback, and how such tools might support more equitable feedback practices in authentic writing instruction.

References

[1] J. Meyer, T. Jansen, R. Schiller, L. W. Liebenow, M. Steinbach, A. Horbach, J. Fleckenstein, Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions, Computers and Education: Artificial Intelligence 6 (2024) 100199. doi:10.1016/j.caeai.2023.100199.

[2] I. Weissburg, S. Anand, S. Levy, H. Jeong, Llms are biased teachers: Evaluating llm bias in personalized education, in: Findings of the Association for Computational Linguistics: NAACL 2025, 2025, pp. 5650–5698.
[3] J. Choi, N. Nixon, Agentic men, communal women?: Exploring gender bias in llm-based leadership identification for collaboration analytics, in: Artificial Intelligence in Education, Springer Nature Switzerland, 2025, pp. 11–18. doi:10.1007/978-3-031-98465-5_2.
[4] K. Yang, M. Raković, D. Gašević, G. Chen, Does the prompt-based large language model recognize students' demographics and introduce bias in essay scoring?, in: Artificial Intelligence in Education, Springer Nature Switzerland, 2025, pp. 75–89. doi:10.1007/978-3-031-98417-4_6.
[5] M. Tan, L. Phalen, D. Demszky, Marked pedagogies: Examining linguistic biases in personalized automated writing feedback, 2026. arXiv:2603.12471.
[6] Y. Du, C. Borchers, M. Cukurova, Benchmarking educational llms with analytics: A case study on gender bias in feedback, 2025. arXiv:2511.08225.
[7] J. Hattie, H. Timperley, The power of feedback, Review of Educational Research 77 (2007) 81–112. doi:10.3102/003465430298487.
[8] D. Carless, D. Boud, The development of student feedback literacy: Enabling uptake of feedback, Assessment & Evaluation in Higher Education 43 (2018) 1315–1325. doi:10.1080/02602938.2018.1463354.
[9] N. E. Winstone, R. A. Nash, M. Parker, J. Rowntree, Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes, Educational Psychologist 52 (2017) 17–37. doi:10.1080/00461520.2016.1207538.
[10] X. Bai, A. Wang, I. Sucholutsky, T. L. Griffiths, Explicitly unbiased large language models still form biased associations, Proceedings of the National Academy of Sciences 122 (2025) e2416228122. doi:10.1073/pnas.2416228122.
[11] J. Lee, Y. Hicke, R. Yu, C. Brooks, R. F. Kizilcec, The life cycle of large language models in education: A framework for understanding sources of bias, British Journal of Educational Technology 55 (2024) 1982–2002. doi:10.1111/bjet.13505.
[12] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186. doi:10.1126/science.aal4230.
[13] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 2021, pp. 610–623. doi:10.1145/3442188.3445922.
[14] D. Guilbeault, S. Delecourt, B. S. Desikan, Age and gender distortion in online media and large language models, Nature 646 (2025) 1129–1137. doi:10.1038/s41586-025-09581-z.
[15] J. Tiedemann, Gender-related beliefs of teachers in elementary school mathematics, Educational Studies in Mathematics 41 (2000) 191–207. doi:10.1023/A:1003953801526.
[16] F. Muntoni, J. Retelsdorf, Gender-specific teacher expectations in reading: The role of teachers' gender stereotypes, Contemporary Educational Psychology 54 (2018) 212–220. doi:10.1016/j.cedpsych.2018.06.012.
[17] M. Cukurova, Agency as a system property in human–ai interaction in education, British Journal of Educational Technology (2026). doi:10.1111/bjet.70060.
[18] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of "bias" in nlp, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5454–5476. doi:10.18653/v1/2020.acl-main.485.
[19] S. Goldfarb-Tarrant, R. Marchant, R. Muñoz Sánchez, M. Pandya, A. Lopez, Intrinsic bias metrics do not correlate with application bias, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2021, pp. 1926–1940. doi:10.18653/v1/2021.acl-long.150.
[20] K. Hyland, F. Hyland, Feedback on second language students' writing, Language Teaching 39

(2006) 83–101. doi:10.1017/S0261444806003399.
[21] R. Ajjawi, D. Boud, Researching feedback dialogue: An interactional analysis approach, Assessment & Evaluation in Higher Education 42 (2017) 252–265. doi:10.1080/02602938.2015.1102863.
[22] Y. Qiao, L. Pan, Y. Mi, L. Liu, Y. Shen, F. Sun, Z. Chu, Lowest span confidence: A zero-shot metric for efficient and black-box hallucination detection in llms, 2026. arXiv:2601.19918.
[23] X. Li, D. Yang, X. Zhu, F. Huang, P. Zhang, Z. Zhao, Span-level emotion-cause-category triplet extraction with instruction tuning llms and data augmentation, Applied Soft Computing (2025). doi:10.1016/j.asoc.2025.113001.
[24] S. Perrella, E. Morales Agostinho, H. Zaragoza, Span-level machine translation meta-evaluation, 2026. arXiv:2603.19921.
[25] Z. Yin, S. Wang, Span-level detection of ai-generated scientific text via contrastive learning and structural calibration, Knowledge-Based Systems 334 (2026) 115123. doi:10.1016/j.knosys.2025.115123.
[26] P. Khatiwada, K. Higgins, A. Mahesh, V. Pappu, B. E. Bagozzi, M. L. Mauriello, Severity-dependent bias in llm evaluators: A span-level audit of polarizing language detection in everyday news, in: Proceedings of the Third Human-Centered Evaluation and Auditing of Language Models Workshop at CHI 2026, 2026. HEAL Workshop at CHI 2026.
[27] Y. Xu, J. Chen, J. Wu, J. Zhang, Hit the sweet spot! span-level ensemble for large language models, in: Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 8314–8325. URL: https://aclanthology.org/2025.coling-main.555/.
[28] W. Previlon, A. S. White, V. Srikumar, Leveraging syntactic dependencies in disambiguation: The case of african american english, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, ELRA and ICCL, 2024, pp. 10405–10417. URL: https://aclanthology.org/2024.lrec-main.909/.
[29] A. Iqbal, et al., Optimising window size of semantic of classification model for identification of in-text citations based on context and intent, PLOS ONE 20 (2025) e0309862. doi:10.1371/journal.pone.0309862.
[30] S. Perrella, E. Morales Agostinho, H. Zaragoza, Span-level machine translation meta-evaluation, arXiv preprint arXiv:2603.19921 (2026). URL: https://arxiv.org/abs/2603.19921.
[31] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.
[32] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.
[33] V. Braun, V. Clarke, Thematic Analysis: A Practical Guide, SAGE Publications, London, 2021.