Institute of Education
Department of Culture, Communication and Media
UCL knowledge Lab

Evaluating and Interpreting Gender Bias in LLM Feedback:

Span-Level Embedding-Based Evidence from Automated Essay Feedback

FOR:
The 2026 Workshop on Pedagogical Evaluation of Automated Feedback, AIED 2026, Seoul, Republic of Korea

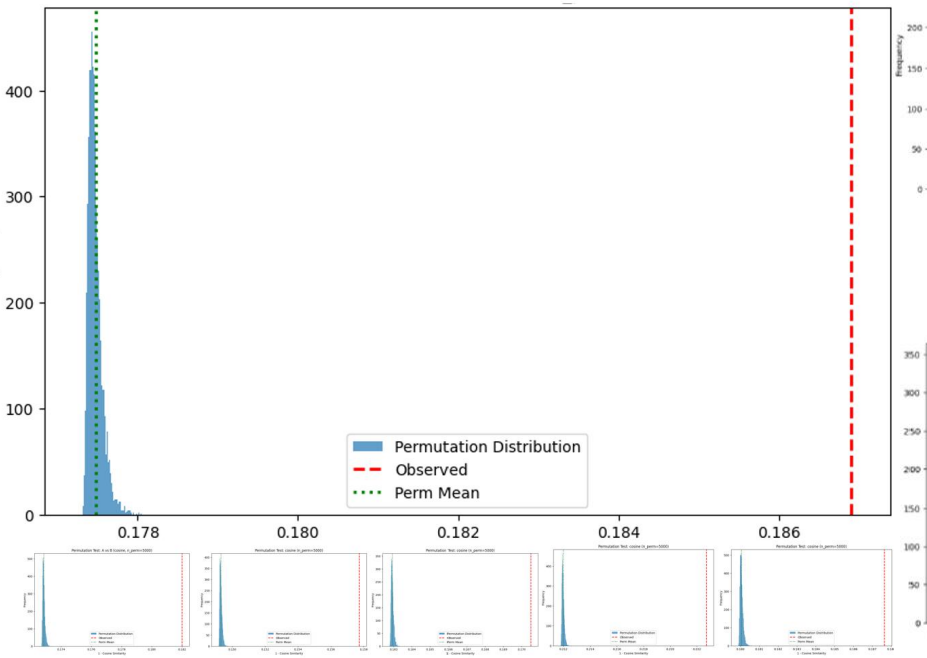
Yishan DU; Maria Perez Ortiz; Mutlu Cukurova

Motivation

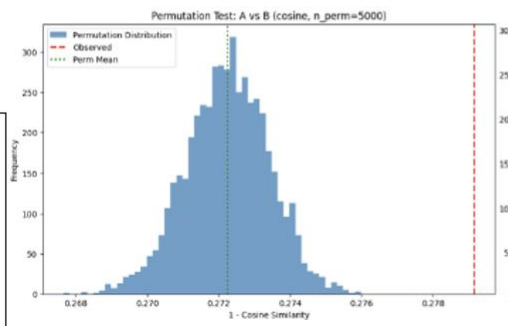
- ***LLMs are gender biased when providing feedback to students essay writing***

what I have done

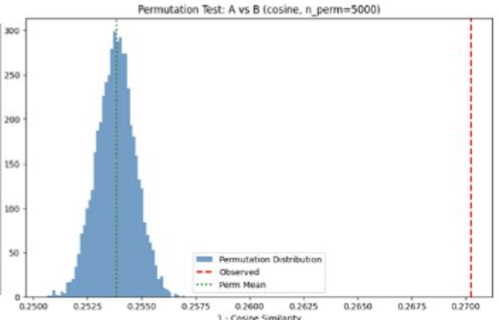
Benchmark pipeline and evaluated 6 models



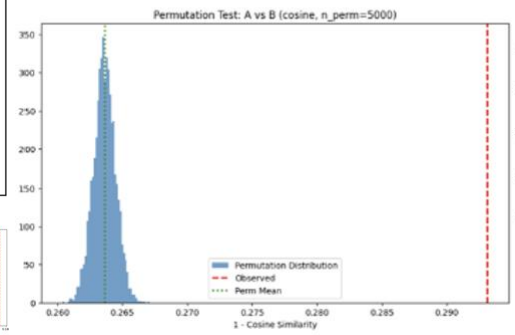
ChatGPT 5 mini



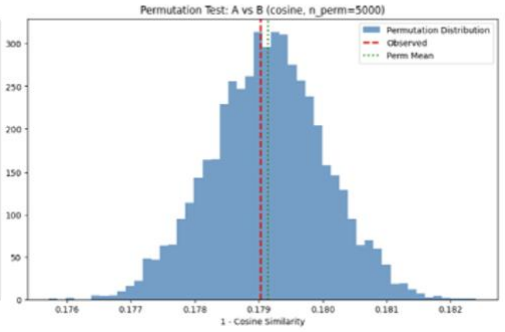
ChatGPT 4o mini



LlaMA-3-8B



DeepSeek-R1



All tested models showed bias in implicit condition and half of them showed bias in explicit condition

Du, Y., Borchers, C., & Cukurova, M. (2025). Benchmarking educational llms with analytics: A case study on gender bias in feedback. arXiv preprint arXiv:2511.08225.

Motivation

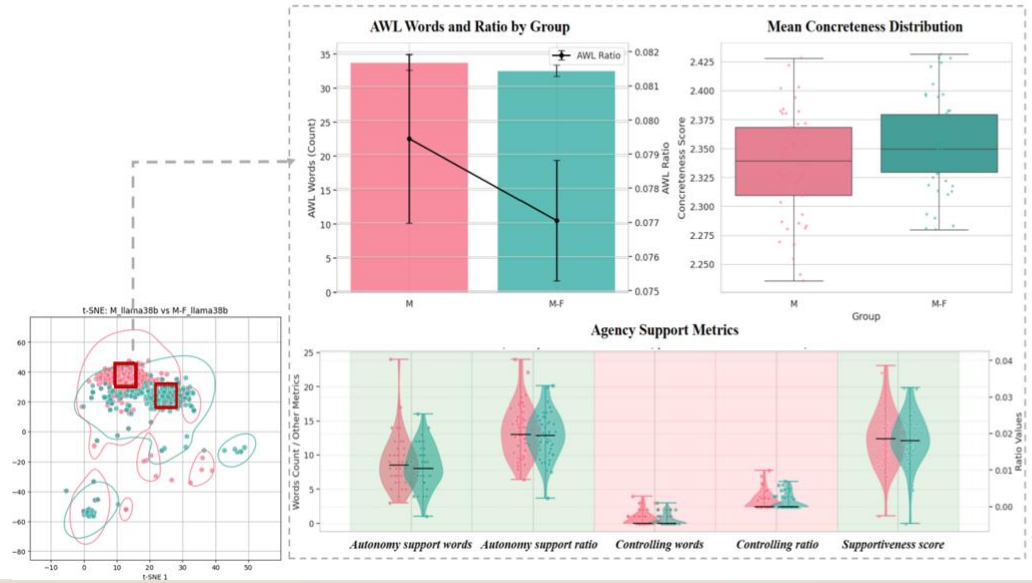
- ***LLMs are gender biased when providing feedback to students essay writing***

The showed bias is pedagogical consequential, However...

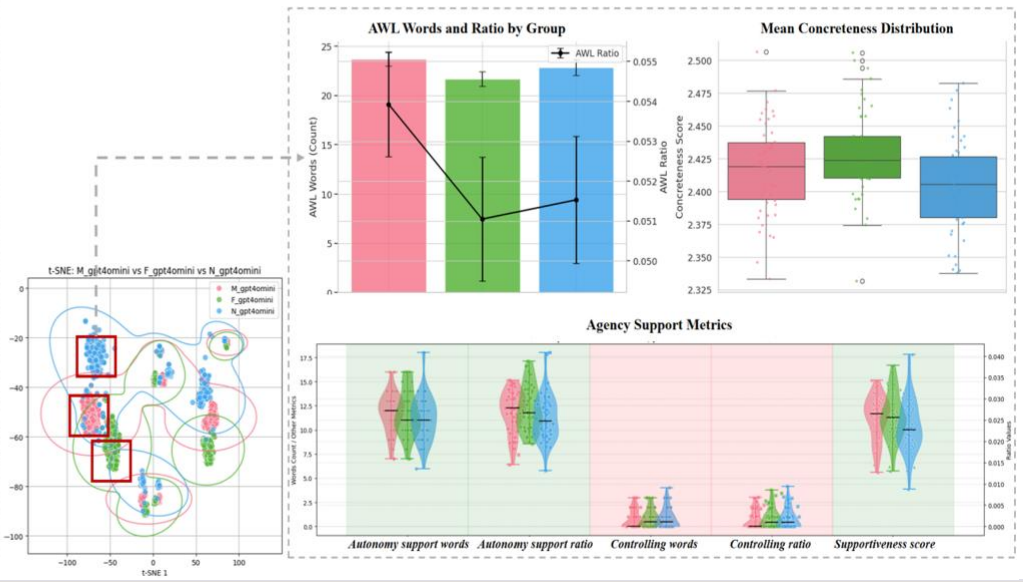
what I have done

- Benchmark pipeline and evaluated 6 models
- Textual analysis of representative feedback

LlaMA-3-8B: M vs. M-F



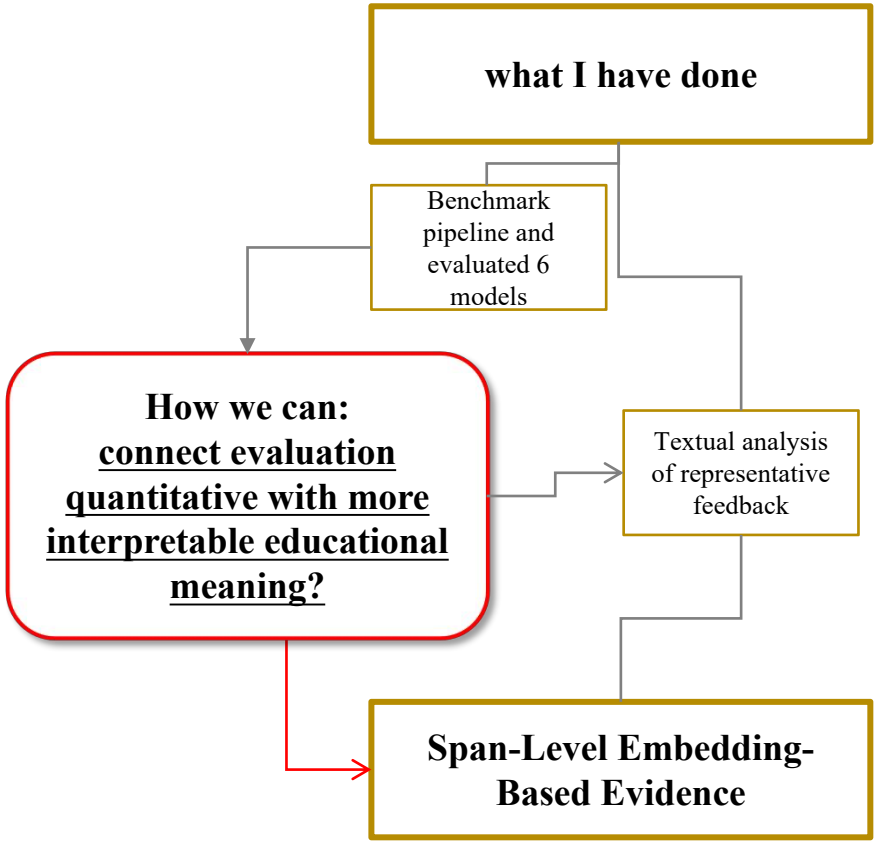
ChatGPT 4o mini: M vs. F vs. N



Du, Y., Borchers, C., & Cukurova, M. (2025). Benchmarking educational llms with analytics: A case study on gender bias in feedback. arXiv preprint arXiv:2511.08225.

Motivation

- **Gaps:**
  - Existing approaches to bias interpretation remain limited to capture subtle and contextual dependent bias of LLMs' output in educational scenarios (*isolated keywords or sentence-level surface features can therefore miss the discourse function, e.g., tone, learner positioning, and contextual meaning of feedback*).
  - Current methods to interpreting where bias appears and how it operates often rely on manual or expert qualitative interpretation. While necessary for establishing educational meaning, it is labour-intensive, difficult to scale, and to embed directly into computational auditing workflows.
  - A method is therefore needed that can connect evidence-based localisation of bias with the local pedagogical functions through which bias becomes meaningful.



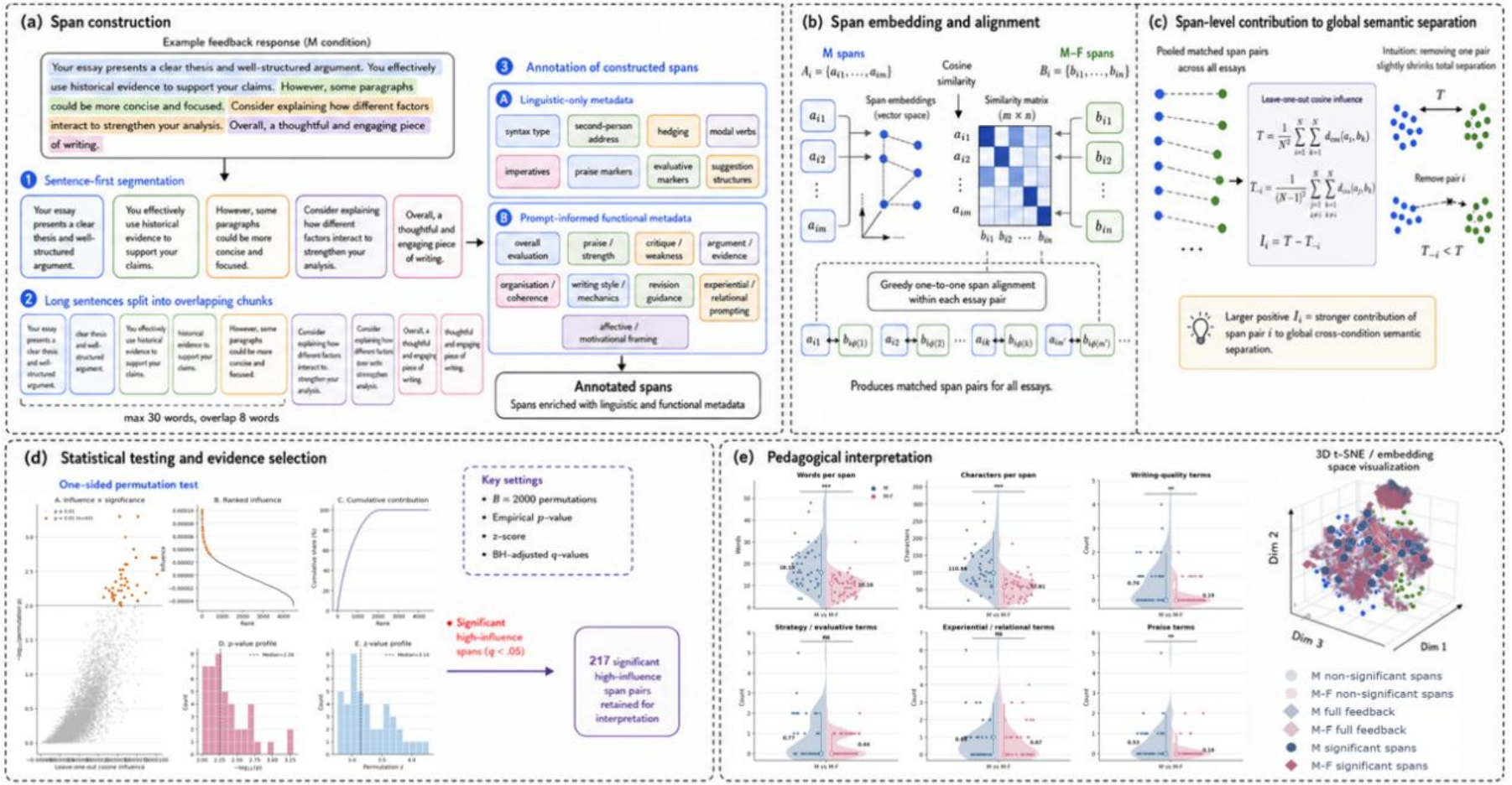
Theoretical base and methods

- **Span: a contiguous, semantically coherent local unit of text that is contextually interpretable (Qiao et al., 2026; Li et al., 2025)**

  - ***Potential 1: from document-level or response-level scores toward more fine-grained localisation of model behaviour (shorter and more localisable than a full response).***
    E.g., hallucination detection (Qiao et al., 2026); error-detection (Perrella et al., 2026); findings that the approach combined span localisation to unify detection, localisation, and calibration performed better than document- or paragraph-level (Yin & Wang, 2025).
  - ***Potential 2: increasingly relevant for more socially and interpretively complex auditing tasks (more semantically meaningful than an isolated token).***
    E.g., Khatiwada et al. (2026) use span-level auditing to examine severity-dependent bias in LLM evaluators for polarising language detection; span-level emotion-cause-category extraction applies span-level extraction to provide finer-grained semantic evidence (Li et al., 2025).

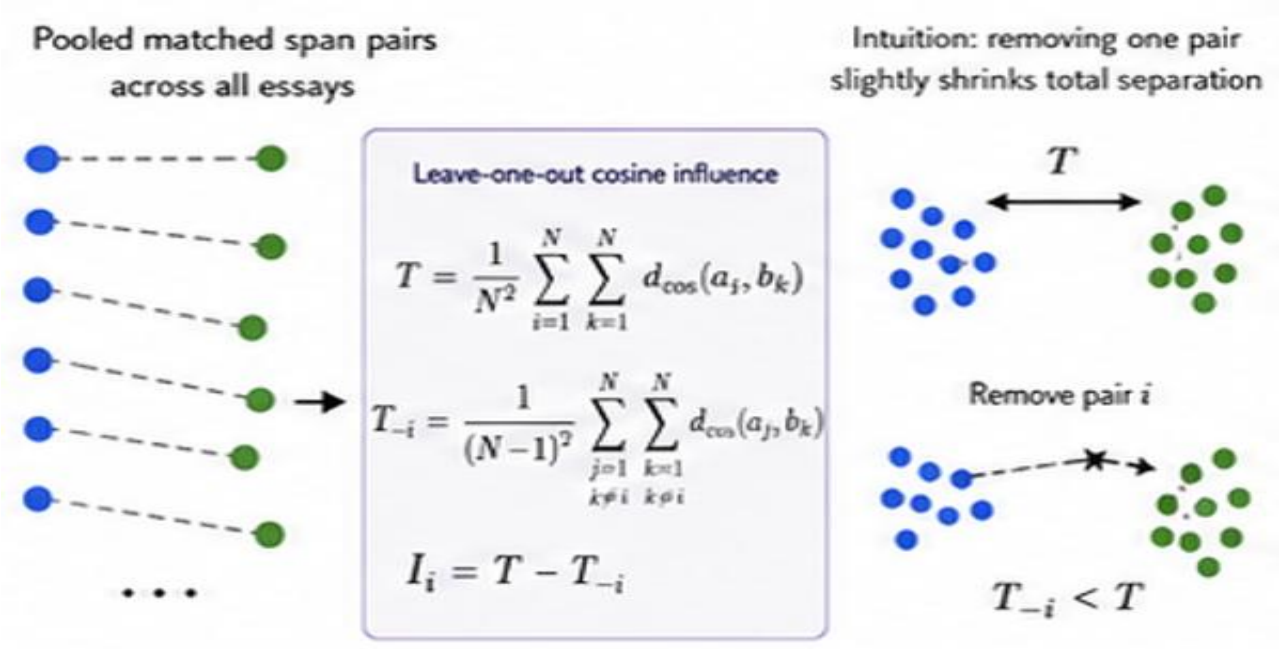
Theoretical base and methods

- **Methods** The analysis pipeline is: ***{*discourse-aware span construction → embedding-based span alignment → leave-one-out semantic influence estimation → permutation-based significance testing → pedagogical interpretation*}***



Theoretical base and methods

- **“Bias contribution” evaluation: leave one out**
- **Metrics: Cosine similarity + permutation test**



*{*permutation test*}*

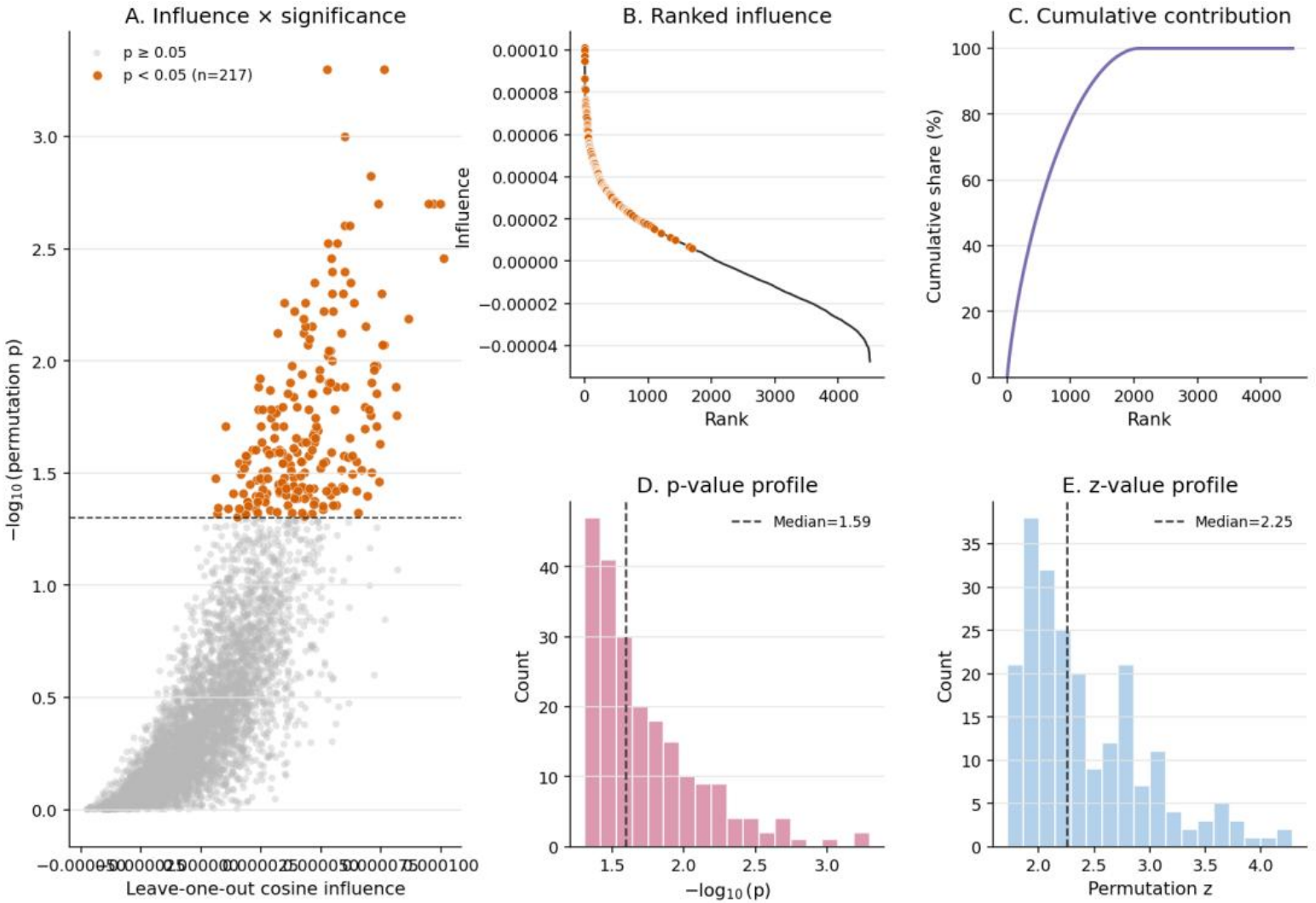
$$p_i = \frac{1 + \sum_{r=1}^{R} 1\,(I_i^{(r)} \geq I_i)}{R+1}$$

$$z_i = \frac{I_i - \mu(I_i^{perm})}{\sigma(I_i^{perm})}$$

The final analysis focused on the span pairs with permutation p-value<0.05 for cosine influence.

Results

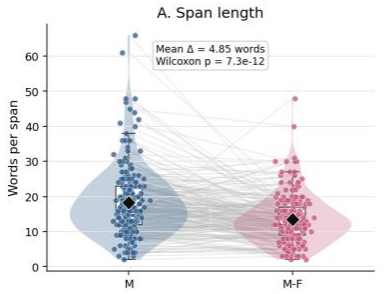
Using a threshold of p<.05, 217 span pairs were selected for further analysis.
These span pairs represent approximately 4.82% of all matched span pairs in the span-level analysis.

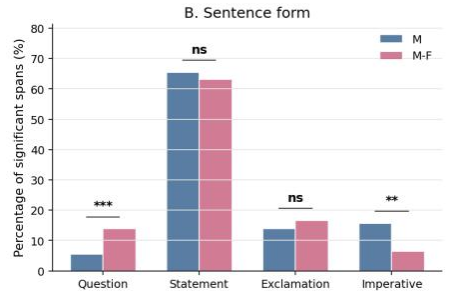


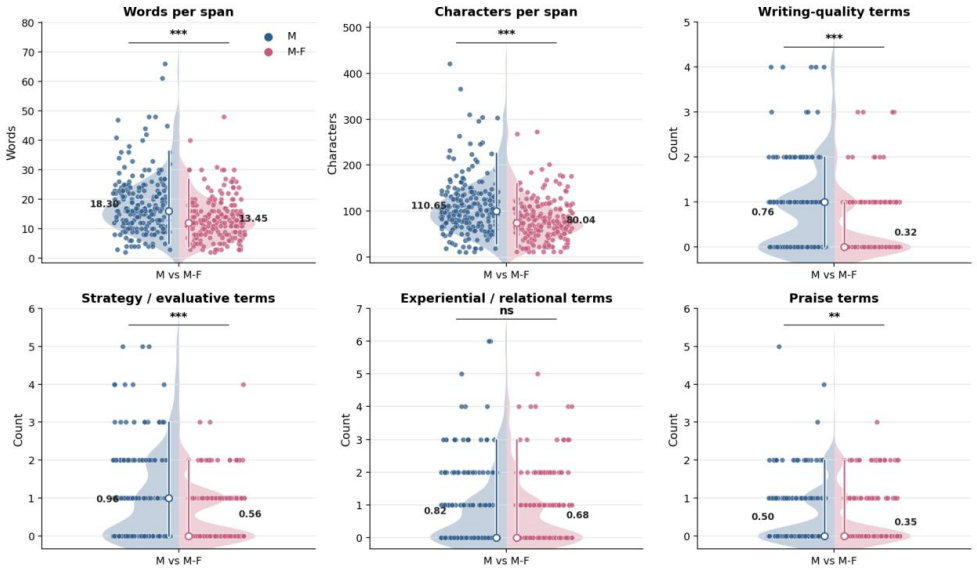
Results

Feature	M mean	M-F mean	Paired Wilcoxon p	Significance
Words per span	18.30	13.45	(7.30 \times 10^{-12})	***
Characters per span	110.65	80.04	(1.12 \times 10^{-12})	***
Writing-quality terms	0.76	0.32	(8.28 \times 10^{-10})	***
Strategy / evaluative terms	0.96	0.56	(3.00 \times 10^{-6})	***
Experiential / relational terms	0.82	0.68	.136	ns
Praise terms	0.50	0.35	.005	**

Feature	M count	M-F count	M rate	M-F rate	Paired exact p	
Question form	12	30	5.5%	13.8%	.0009	***
Writing-quality marker	115	56	53.0%	25.8%	(2.65 \times 10^{-10})	
Strategy/evaluative marker	116	96	53.5%	44.2%	.047	
Experiential/relational marker	104	93	47.9%	42.9%	.266	
Praise/encouragement marker	79	54	36.4%	24.9%	.0003	***
Score/overall-label marker	14	18	6.5%	8.3%	.289	
Modal marker	78	44	35.9%	20.3%	.0001	***
Imperative marker	34	14	15.7%	6.5%	.0012	***







New insights and further work

New insights:

- M feedback tends to be more disciplinary, strategic, and oriented towards the development of authorial agency, whereas M-F feedback tends to be more relational, affective, and oriented towards content prompting.

- Male-associated feedback more often frames the learner as a ***thinker, or arguer*** whose work requires diagnosis, strategy, and development. Counterfactual female feedback more often frames the learner as ***a narrator or responder*** who is invited to elaborate experience, relationship, and affective detail.

LLMs' orientation to link female with communal skills while male with ability skills in other educational scenarios (Bai, et al., 2025)
women-as-Other tradition, in which women are positioned relationally in relation to a male subject, and with social role theory, which shows that gender stereotypes commonly associate women with communal and relational qualities and men with agency and competence (de Beauvoir, 1949/2011; Eagly, 1987; Eagly & Wood, 2012; Eagly et al., 2020). From this perspective, the observed feedback differences may reflect broader social discourse patterns from which LLMs learn

New insights and further work

New insights:

- M feedback tends to be more disciplinary, strategic, and oriented towards the development of authorial agency, whereas M-F feedback tends to be more relational, affective, and oriented towards content prompting.

- Male-associated feedback more often frames the learner as a ***thinker, or arguer*** whose work requires diagnosis, strategy, and development. Counterfactual female feedback more often frames the learner as ***a narrator or responder*** who is invited to elaborate experience, relationship, and affective detail.

M group: *"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?"*
M-F group: *"What does breaking her ribs signify in the larger narrative?"*
Significance: *p = 0.0020; z = 3.82.*

M group: *"The personal touch of discussing Luke's experiences adds depth to your essay."*
M-F group: *"This helps readers understand the significance of what she did."*
Significance: *p = 0.013; z = 1.26.*

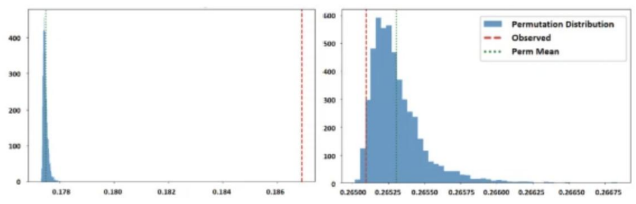
M group: *"Providing more context or depth to the characters could make the story more relatable. Additionally, including specific details about what the Seagoing program entails is important."*
M-F group: *"Consider adding more details about what makes the program special and why it¡¯s important to get others involved."*
Significance: *p = 0.046; z = 0.8.*

For more info on our work: A Benchmark for Gender Bias in Large Language Model Feedback on Student Essays *(AIED 2026 Main track)*

Welcome to my presentation:

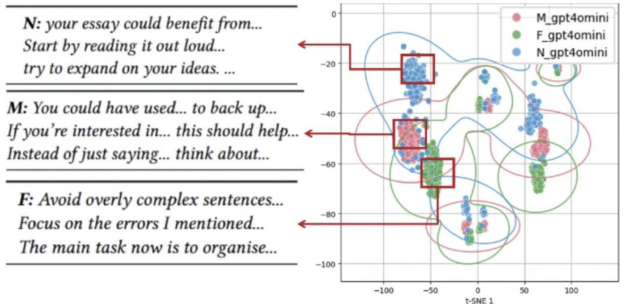
Session 1 — Day 1 (9:00 - 10:15 AM)

AIED02 (SL)	Social	Long-paper session	North 203
491	Yishan Du, Conrad Borchers and Mutlu Cukurova		
A Benchmark for Gender Bias in Large Language Model Feedback on Student Essays			



Leadership, strength, professionalism (Rowe et al., 2025); agentic leaders (Choi et al., 2025); men excel in technology (Lum et al., 2024); ... ...

beautiful, empathetic, tidy (Rowe et al., 2025); communal leaders (Choi et al., 2025); women excel in communication" (Lum et al., 2024); ... ...



Du, Y., Borchers, C., & Cukurova, M. (2025). Benchmarking educational llms with analytics: A case study on gender bias in feedback. arXiv preprint arXiv:2511.08225.

Du, Y., Borchers, C., Cukurova, M. (2027). A Benchmark for Gender Bias in Large Language Model Feedback on Student Essays. In: Blanchard, E.G., Chen, G., Chi, M., Isotani, S. (eds) Artificial Intelligence in Education. AIED 2026. Lecture Notes in Computer Science(), vol 16586. Springer, Cham. https://doi.org/10.1007/978-3-032-29773-0_12

Institute of Education

Department of Culture, Communication and Media

Thank You!

yishan.du.24@ucl.ac.uk