

Fluent but Misaligned: An NLP Approach to Measuring Uptake in Classroom AI-Avatar Dialogue

Paolo MENGONI¹, **Daniel SHEN Jiandong**², **Aru NURGISSAYEVA**², **Kristen LI Yuanxi**³, **Renia LOPEZ-OZIEBLO**⁴, and **Paulina WONG Pui Yun**⁵

¹ Department of Interactive Media, Hong Kong Baptist University, Hong Kong SAR, China

² Soqle Hong Kong Limited, Hong Kong SAR, China

³ Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

⁴ Department of English and Communication, The Hong Kong Polytechnic University, Hong Kong SAR, China

⁵ Division of Science, Lingnan University, Hong Kong SAR, China

Abstract

Generative AI can sound fluent yet still miss what students mean, especially during brief classroom brainstorming. We analysed chat logs from a mathematics-for-game-development lesson in which 25 Hong Kong undergraduates worked in small groups and interacted with a generative AI avatar. We treated each student turn and the following avatar reply as a dyad and used Natural Language Processing (NLP) to measure three classroom-relevant signals: meaning continuity (Sentence-BERT cosine similarity), keyword carryover (Jaccard overlap of content-word lemmas), and interaction mode (Language Style Matching). Results showed an opportunity-structure effect: when students wrote longer, more content-rich turns, the avatar's replies were more consistently on-topic and carried forward more task terms. We also observed two productive patterns (on-topic support with shared terminology or with paraphrase) and a risk pattern, *keyword echo*, where term reuse looked responsive but meaning continuity was weak. These measures offer educators a practical way to interpret classroom GenAI dialogue beyond surface fluency.

Keywords: AI in education (AIEd), learning analytics, classroom discourse, dialogue analysis, semantic similarity

1. Introduction

Generative AI is increasingly used in classrooms for brainstorming, but its value depends on how it handles short or unclear input. The AI can sound fluent

yet still miss the student’s intent. Teachers need a way to check if the AI responds to the student’s idea, not just plausible text (Chiu, 2021; Giannakos, 2024).

To judge response quality turn by turn, we examine three alignment signals: meaning (idea continuity), wording (key-term carryover), and style (function-word mode). We separate these signals because good brainstorming support often paraphrases or expands the idea, so the meaning can match even when the wording changes. By contrast, the AI can repeat keywords but still misunderstand the student’s point.

Evidence from classroom brainstorming logs is limited, and evaluations often mix meaning continuity with surface repetition, which can hide *keyword echo*, where word repetition appears responsive despite weak meaning continuity. Without separating these signals, conclusions about classroom support can be misleading. Short turns can depress overlap scores even when a reply is helpful, while keyword repetition can inflate apparent responsiveness even when intent is missed.

To study this in a real setting, we analyse classroom chat logs from a brainstorming lesson where students interacted with a generative AI avatar. Each student turn and the immediately following avatar reply is treated as an adjacent dyad. We measure meaning alignment using Sentence-BERT similarity between the student turn and the AI reply (Reimers & Gurevych, 2019). We also measure lexical overlap and Language Style Matching (LSM) to describe interaction mode. We report turn length to support interpretation of these scores.

Our research question is: In classroom brainstorming dialogue, to what extent does the AI avatar follow students’ immediately prior ideas, and how do lexical grounding and interaction style relate to this continuity under short or noisy student turns?

2. Methodology

We analysed a brainstorming activity where 25 Hong Kong undergraduates interacted with a generative AI avatar in a mathematics-for-game-development lesson (e.g., 3D rotation and animation). Students worked in small groups. They discussed ideas together, and one student typed prompts to the avatar for the group. The task involved open-ended brainstorming and refinement of technical ideas, such as explaining concepts, suggesting uses, or clarifying steps.

We chose brainstorming because classroom prompts are often brief or unclear. In this setting, we focus on whether the avatar maintains alignment with

the student's immediately prior idea across turns. Avatar/interface configuration was not treated as an analytic factor in this paper; all adjacent student-AI dyads were analysed together at the turn level.

Data Source and Unit of Analysis

Data were drawn from platform chat logs recorded in Soqqle's Classlet system, which stores turn-by-turn student messages and avatar replies during classroom activities. We analysed interaction at the level of adjacent *student → AI* dyads: each student turn was paired with the immediately following avatar response. This dyad unit captures the most direct test of whether the AI follows the student's immediately prior contribution.

Measures and Analysis

We computed three dyad-level indicators to describe classroom interaction quality in a way that is interpretable for teaching practice: whether the avatar takes up a student's immediately prior contribution (meaning continuity), whether it carries forward key task terms (light lexical grounding), and whether it mirrors the interaction mode of the student turn (brief/social versus task-focused). These indicators are not treated as learning outcomes; they are used as behavioural traces of contingent responding during brainstorming in authentic classroom conditions (e.g., short, messy prompts and mixed communicative purposes).

Semantic Alignment (Meaning Continuity)

In classroom brainstorming, a helpful AI response continues the student's idea, even when the prompt is brief or informal. To measure this uptake, we computed cosine similarity between Sentence-BERT embeddings of each student turn and the next avatar reply (Reimers & Gurevych, 2019). Higher similarity indicates stronger meaning continuity, even when wording differs.

Lexical Alignment (Keyword Reuse / Grounding)

In learning dialogue, shared terminology can support mutual understanding, especially in technical topics where key terms carry the main idea. Lexical overlap is constrained by the amount of topical content in the student turn. We lemmatised content words and removed non-content tokens, then computed Jaccard similarity between the student and avatar lemma sets. If a turn contained no content lemmas, overlap was undefined and retained as an indicator of minimal topical content.

Language Style Matching (Interaction Mode)

In classrooms, not all turns are task-focused. Students may greet the system, test it, or type partial prompts before asking a fuller question. To describe interaction mode in each dyad, we computed Language Style Matching (LSM) from similarity in function-word categories between the student turn and avatar reply (Ireland & Pennebaker, 2010). We use LSM to describe interaction mode and contextualise alignment scores, not as evidence of rapport or learning.

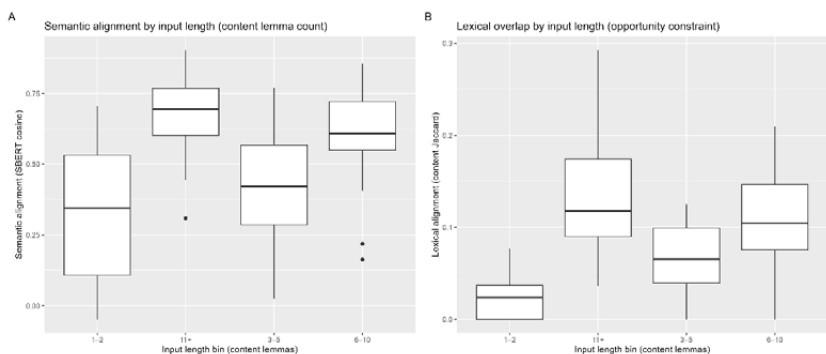
3. Results and Discussion

Corpus Overview

The primary dataset contained 175 dyads for measures available on all turns (semantic alignment and style matching). For lexical overlap, 160 dyads contained usable content after filtering; remaining cases had no content lemmas and were retained as informative “thin/non-content” turns. Dyads were indexed by turn order (sequence), ranging from 3 to 41.

Result 1: Opportunity Structure (Student Input Length)

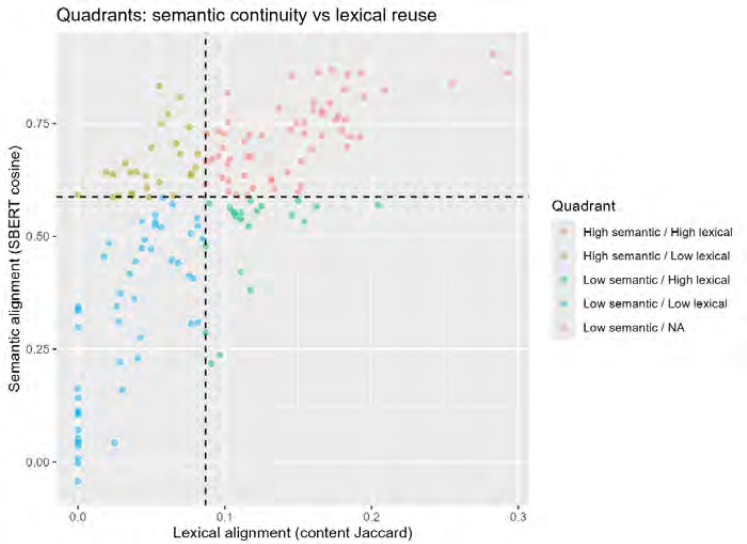
Figure 1 shows a clear “opportunity structure” effect: when students contribute very little usable content, alignment measures have limited material to work with and the avatar’s continuity appears less stable. As student prompts become more content-rich, the interaction becomes more consistently on-topic and shared task vocabulary becomes easier to sustain. This suggests that apparent misalignment in classroom logs can reflect brief or fragmentary student input rather than purely the quality of the AI response.



(Figure 1: Opportunity structure by student input length. Semantic alignment (cosine similarity) and lexical overlap (Jaccard) by student content-lemma bins. Meaning continuity increases with richer prompts (median ≈ 0.35 for 1–2 lemmas vs ≈ 0.70 for 11+ lemmas), and lexical overlap rises in parallel (median ≈ 0.02 vs ≈ 0.10 – 0.12), showing that short classroom turns provide limited opportunity for measurable alignment.)

Result 2: Meaning Continuity Overall, Plus “Keyword Echo” Risk

Figure 2 shows that “good” brainstorming support can occur through more than one productive response mode. The avatar can stay on the student’s idea while reusing the student’s terminology, or it can stay on the same idea while re-expressing it in different words (paraphrase/elaboration). Importantly, the map also reveals a classroom-relevant risk pattern: responses can echo keywords without reliably tracking the student’s intended meaning, which may look responsive on the surface but can undermine trust if students treat keyword reuse as evidence of understanding.



(Figure 2: Quadrant map of meaning continuity (semantic alignment, y-axis) versus keyword reuse (lexical overlap, x-axis). Two productive modes dominate—high semantic + high lexical (34.9%) and high semantic + low lexical (15.4%; paraphrase)—while low semantic + high lexical (12.6%) flags “keyword echo” (term reuse without strong meaning continuity).)

Finding 1: Classroom Prompting Conditions Shape What Responsiveness Can Look Like

Students often contribute short fragments, partial ideas, or social turns, which limits how much alignment can appear in the transcript. In Figure 1, meaning continuity and term carryover increase when students provide more usable content (semantic median ≈ 0.35 at 1–2 lemmas vs ≈ 0.70 at 11+; lexical median ≈ 0.02 vs ≈ 0.10 – 0.12). Educationally, low alignment in classroom logs may reflect brief student input rather than AI failure. This is relevant because perceived responsiveness may influence engagement and willingness to participate (Chiu, 2021; Giannakos, 2024).

Finding 2: Educationally Useful Uptake Can Occur Through Shared Terminology or Paraphrase

In brainstorming, the reply does not need to repeat student wording; it needs to continue the student's idea in a way that supports ideation. Figure 2 shows two productive patterns: high semantic + high lexical alignment (34.9%), where the avatar continues the idea while carrying forward key terms, and high semantic + low lexical alignment (15.4%), where the avatar continues the idea while rewording it, consistent with paraphrase or elaboration. This suggests that useful AI uptake does not depend on exact keyword repetition; on-topic support may also appear through paraphrase or elaboration (Chiu, 2021). We capture meaning continuity with embedding-based semantic similarity rather than surface matching alone (Reimers & Gurevych, 2019).

Finding 3: Surface Cues of Understanding Are Not Always Reliable in Learning Dialogue

In classroom talk, visible term reuse can be taken as a cue that the system has understood, which may affect whether students and teachers trust the reply. Figure 2 highlights a counter-pattern: *keyword echo*, where lexical overlap is high but meaning continuity is low (12.6%). This pattern shows that word repetition can look responsive even when the reply does not track the student's intended idea. For classroom use, semantic continuity should be interpreted separately from word reuse to avoid overestimating uptake.

4. Implications

Teachers should not judge responsiveness by surface fluency or keyword reuse alone. Interpreting semantic continuity alongside lexical overlap helps distinguish productive paraphrase from *keyword echo*, enabling more reliable review and timely intervention. The opportunity-structure effect suggests that short student prompts limit measurable uptake. Low alignment scores should be interpreted alongside student input length and turn type, not seen as poor AI performance.

Acknowledgments

This study was funded by Lingnan University, Hong Kong, through the Fund for Innovative Technology-in-Education (FITE) of the University Grants Committee (Project Code: 120042) and “UGC FITE IICA #10 Beyond Reality: Unleashing Generative Metaverse Avatars in Education” fund, University Grant Committee, Hong Kong Baptist University, Hong Kong, China.

Ethical Declaration Statement

The research was ethically approved by the Sub-Committee on Research Ethics and Safety of the Research Committee at Lingnan University, Hong Kong (Ref. no. EC176-2324).

Declaration of Use of AI in the Writing Process

The author(s) used ChatGPT during the preparation of this work to improve clarity and concision of the writing. After utilising the tool/service, the author(s) thoroughly reviewed and edited the content as necessary and assumed full responsibility for the publication's content.

References

- Chiu, T. K. F. (2021). Digital support for student engagement in blended learning based on self-determination theory. *Computers in Human Behavior*, 124, 106909. <https://doi.org/10.1016/j.chb.2021.106909>
- Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., Järvelä, S., Mavrikis, M., & Rienties, B. (2024). The promise and challenges of generative AI in education. *Behaviour & Information Technology*, 1–27. <https://doi.org/10.1080/0144929X.2024.2394886>
- Ireland, M. E., & Pennebaker, J. W. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology*, 99(3), 549–571. <https://doi.org/10.1037/a0020386>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv*. <https://doi.org/10.48550/arXiv.1908.10084>