ACM SIGCHI CHIGreece-Cyprus Chapters Symposium 2026
November 6–7, 2026 | Athens, Thessaloniki, Nicosia & Online



Pedagogical architectures of student–LLM interaction in L2 learning

A PRISMA systematic review and an action research design for Greek

Spartakos Anagnostaras, Iro Voulgari, & Stathis Selimis
Department of Speech and Language Therapy
University of the Peloponnese

Two parts, one argument



What 51 studies tell us

- The problem
- PRISMA 2020 review: method & corpus



An action research design for Greek as L2

- Research questions
- Two cohorts, two architectures

It is the architecture of student–LLM interaction, not the LLM per se, that promotes learning and shapes what pupils retain.

1 The problem: AI deployed in classrooms without sufficient empirical knowledge

AI in classrooms

Deployed without sufficient empirical knowledge, including L2 contexts

Higher-education bias

Initial research targets Higher Education; findings do not apply directly to schools

Surface vs. deep

LLMs boost fluency (grammar/vocabulary) but lack deep cognitive impact

Retention gap

L2 retention post-LLM withdrawal is rarely evaluated

Liu et al. (2026)

- Primary, Grade 6 (M age = 11.4 years); n = 65
- Randomised controlled trial (RCT) with control group and 1-month follow-up, 5 weeks (10 sessions)
- The differences in favour of the experimental group in written performance and across all four self-regulation strategies were not significant at the immediate follow-up and only became significant one month later, as the control group regressed to its initial levels.

2 It is the architecture, not the LLM *per se*

- **Interaction design is under-reported:** Research often neglects prompts, system rules, and interaction constraints. This lack of accurate reporting of the LLM version has immediate consequences since commercial LLMs are constantly updated without any warning.
- **Sung & Kang (2025):** Withholding ready-made answers lowered satisfaction but yielded higher text scores.
- **Pedagogical friction:** Desirable difficulty and productive struggle are key to deep learning.
- **Consequence:** The comparison worth making is between interaction architectures — not between 'with' and 'without' AI.

2 Five design principles the review points to

- Feedback must be specific (Zhuang et al., 2025)
- Division of labour between the LLM and the teacher (Kim & Chon, 2025; Ozan & Azap, 2026; Sung & Kang, 2025).
- For less proficient students, the support should be explicit, and for more proficient students, support should be opportunities for critical reflection (Woo et al., 2025)
- Prompts for LLMs must be taught and seen as a type of literacy (Duong & Chen, 2025; L. Liu et al., 2026; Woo et al., 2026).
- Students must learn to check a system that can be wrong.

3 Theoretical framework

Constructivism & discovery learning

Active knowledge co-construction through dialogue (Alanazi, 2016).

Feedback model

Hattie and Timperley (2007) describe levels of feedback as task, process, self-regulation, and person. Of these levels, feedback that focuses on process and self-regulation results in the greatest improvement in learning.

TPACK

The TPACK framework (Mishra & Koehler, 2006) indicates that the integration of pedagogical knowledge and content knowledge determines the value that technological knowledge holds in teaching. According to this viewpoint, the LLM is shifting the role of the teacher rather than replacing it.

4 Objective & research questions

Objective Examine how student-LLM interaction shapes feedback quality and learning retention.

Main question How do teacher-mediated vs. direct ChatGPT interaction architectures support L2 Greek?

Q1 How does ChatGPT feedback to the teacher align with Hattie & Timperley's (2007) feedback levels? — Study A

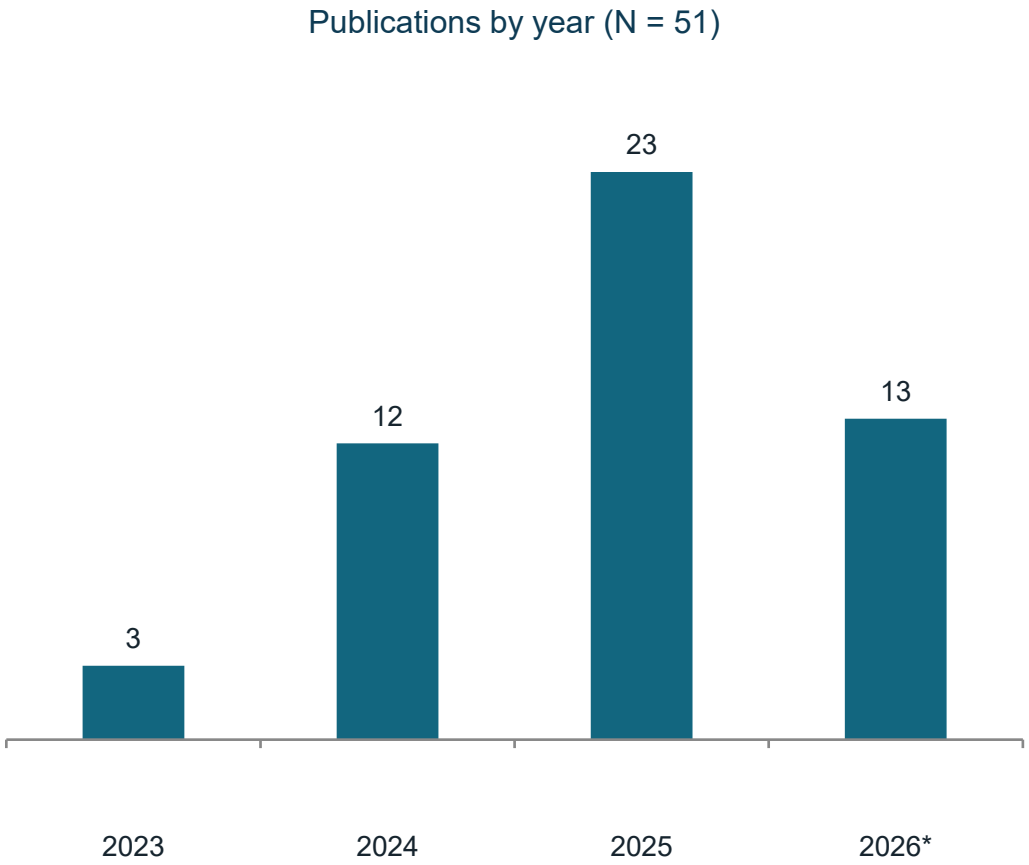
Q2 What is the quality of direct ChatGPT feedback to lower secondary learners? — Study B

Q3 How do L2 Greek skills develop and persist after AI support is withdrawn (retention)? — A + B

Q4 What interaction strategies do secondary school learners adopt with ChatGPT? — Study B

Q5 How does metalinguistic awareness evolve across L1–L2? — A + B

5 PRISMA 2020 REVIEW: An emerging field



No eligible studies before 2023 · *2026: first quarter only

56.9%

East Asia (n = 29): China, Hong Kong, South Korea, Taiwan

47.1%

Upper secondary (n = 24); primary education only n = 3

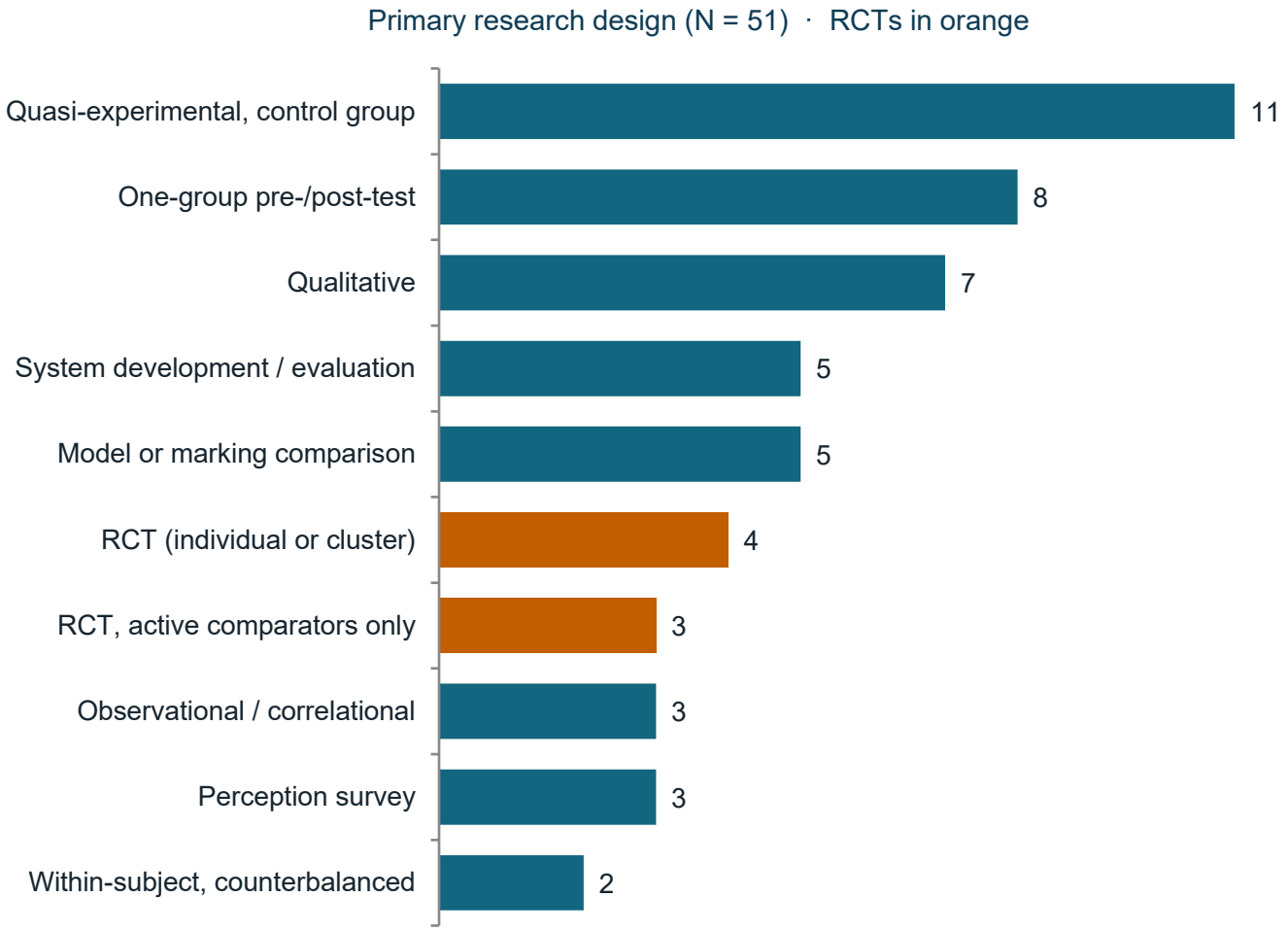
60.8%

Writing as the target skill (n = 31); listening n = 0

80.4%

GPT-family models

5 PRISMA 2020 REVIEW: How strong is the evidence?



35%

compare groups (18 / 51)

14%

randomise (7 / 51)

3–6 weeks

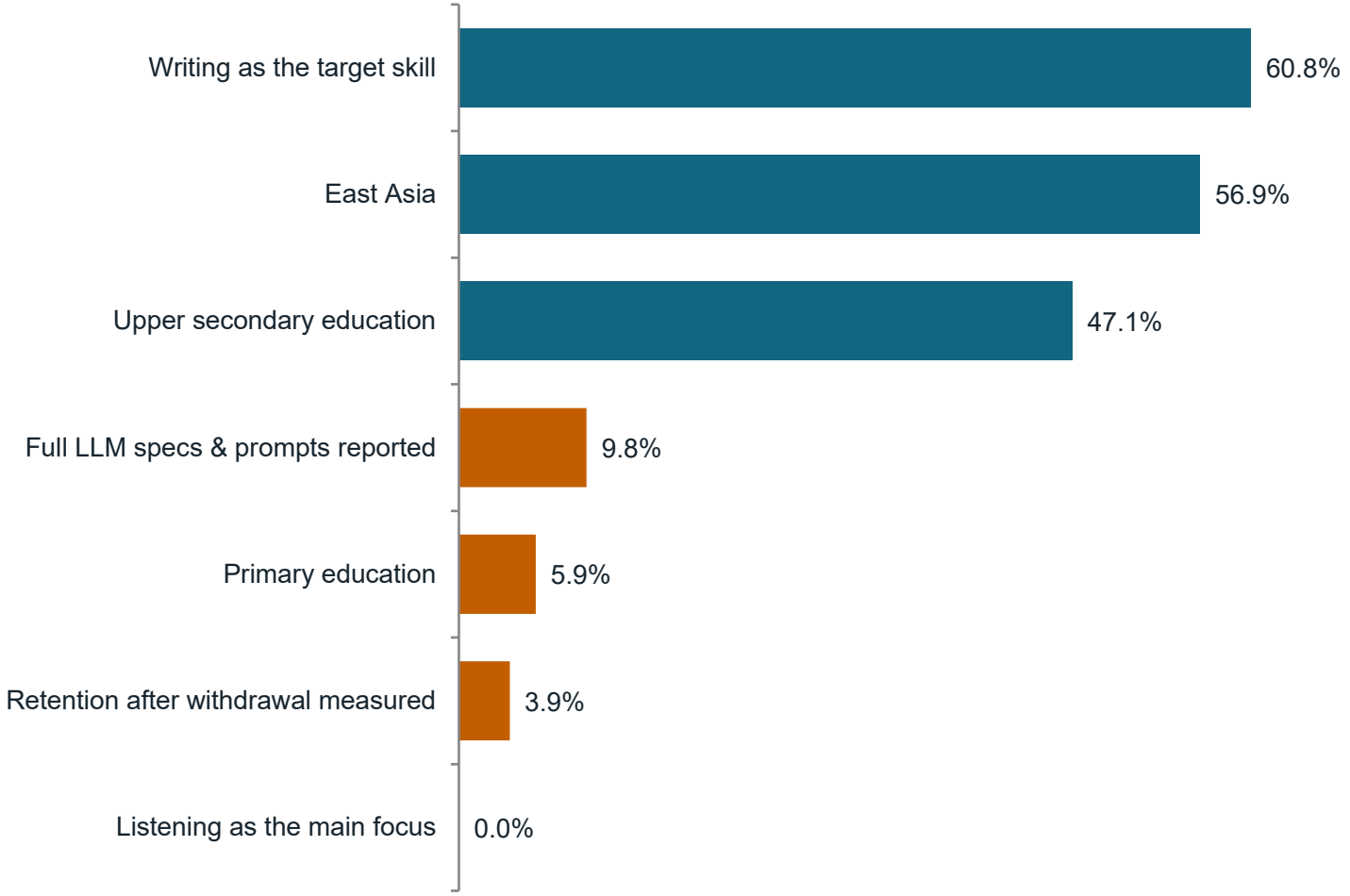
typical intervention length

2

studies with a delayed retention measure (Liu et al., 2026; Myung et al., 2026)

Quality appraisal (max 16): mean 12.8, median 14; 21.6% score ≤ 10.

6 Evidence breakdown & gap analysis



Where the evidence piles up

The gap this study targets

49 / 51

studies never measure what remains after the tool is withdrawn

7 Bridging the gap: This study

Gap in the evidence	This study
Primary education underrepresented (3 studies, 5.9%)	**Focus on primary (n = 16) & lower secondary (n = 6)**
English-language dominance	**Greek as L2 (low-resource, limited-exposure context)**
Short interventions (3–6 weeks)	**3-year prospective design (2026–2028)**
Focus on written output (60.8%)	**Dual focus: written texts & audio recordings**
Emphasis on motivation over linguistic gains	**Balanced tracking of language performance & retention**
Generic LLM feedback vs. teacher feedback	**Scaffolding quality via Hattie & Timperley feedback levels**
Omission of LLM specs & prompts	**Full reporting of LLM specs, versions & prompts**

8 Action Research Design

- **Methodology:** Prospective collective case study and practical action research
- **Setting & duration:** 'Ellada' school, Moldova — 3-year prospective study (2026–2028)
- **Participants:** N = 22 Romanian-speaking pupils of Greek origin

STUDY A — Primary (n = 16)

Teacher ⇄ ChatGPT → Pupils

STUDY B — Lower secondary (n = 6)

Pupils ⇄ ChatGPT

9 Research data, tools & ethics

Standardised tests

'Let's Speak Greek' I–III (Anastasiadi-Symeonidi et al., 2011)

DVIQ (Stavrakaki & Tsimpli, 2000)

Custom instruments

Mid-term & delayed retention tests

Observation forms

Feedback rubrics

Data sources

Student texts & audio recordings

ChatGPT interaction logs

Teacher's instruction log

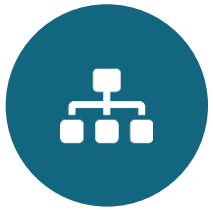
Reflective diary

10 Contribution



Comparative architecture

Mediated vs. direct LLM interaction models in school settings



Feedback mapping

Categorising LLM feedback levels (Hattie & Timperley, 2007) for L2 learners



Empirical baseline

First longitudinal dataset on LLM-assisted learning of Greek as L2

11 Open questions

? Scaffolding

How do we fade scaffolding in an AI tool that is designed never to stop helping?

? Evaluation

Learner satisfaction vs. actual learning outcome in LLM interaction?

? Validity

How do we account for LLM technology drift across a 3-year longitudinal study?

References (1/2)

Alanazi, A. (2016). A critical review of constructivist theory and the emergence of constructionism. *American Research Journal of Humanities and Social Sciences, 2*, 1–8.

Anastasiadi-Symeonidi, A., Papadopoulou, D., Tzevelekou, M., Varlokosta, S., Iakovou, M., Kantzou, V., Lytra, V., Sioupi, A., Stamouli, S., Terzi, A., & Chondrogianni, V. (2011). *Ας μιλήσουμε Ελληνικά I, II, III* [Let's speak Greek I, II, III]. Monochromia.

Duong, T.-N.-A., & Chen, H.-L. (2025). An AI chatbot for EFL writing: Students' usage tendencies, writing performance, and perceptions. *Journal of Educational Computing Research, 63*(2), 406–430. https://doi.org/10.1177/07356331241312363

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Kim, M., & Chon, Y. V. (2025). The impact of self-revision, machine translation, and ChatGPT on L2 writing: Raters' assessments, linguistic complexity, and error correction. *Assessing Writing, 65*, Article 100950. https://doi.org/10.1016/j.asw.2025.100950

Lavidas, K., & Aravantinos, S. (2026). Systematic literature review / Systematic reviews [Educational material for Educational Research Methodology]. Department of Educational Sciences and Early Childhood Education, University of Patras.

Liu, L., Daungcharone, K., & Thongkoo, K. (2026). Evaluating ChatGPT-assisted Chinese reading: Impacts on achievement, motivation, and technology acceptance. In *2026 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering (ECTI DAMT & NCON)*. IEEE. https://doi.org/10.1109/ECTIDAMTNCON67592.2026.11459975

Liu, Z.-M., Hwang, G.-J., Chen, C.-Q., Chen, X.-D., & Ye, X.-D. (2026). Integrating large language models into EFL writing instruction: Effects on performance, self-regulated learning strategies, and motivation. *Computer Assisted Language Learning, 39*(3), 466–490. https://doi.org/10.1080/09588221.2024.2389923

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054.

References (2/2)

Myung, J., Lim, H., Oh, H., Jin, H., Kang, N., Ahn, S. Y., Hong, H., Oh, A., & Kim, J. (2026). When scaffolding breaks: Investigating student interaction with LLM-based writing support in real-time K-12 EFL classrooms. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)* (Article 838). Association for Computing Machinery. https://doi.org/10.1145/3772318.3791517

Ozan, O., & Azap, S. (2026). Employing ChatGPT to improve high school students' writing skills by providing feedback on topic-specific writing tasks. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2026.2613219

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Stavrakaki, S., & Tsimpli, I.-M. (2000). Διαγνωστικό τεστ γλωσσικής νοημοσύνης για παιδιά σχολικής και προσχολικής ηλικίας: Στάθμιση, στατιστική ανάλυση, ψυχομετρικές ιδιότητες [DVIQ]. In *Πρακτικά του 8ου Συνεδρίου Λογοπεδικών* (pp. 95–106). Ellinika Grammata.

Sung, M.-C., & Kang, S. (2025). GPT API-based chatbots as adaptive and facilitative tutors for L2 English process writing. *Korean Journal of English Language and Linguistics, 25*, 448–471. https://doi.org/10.15738/kjell.25..202504.448

Woo, D. J., Guo, K., & Susanto, H. (2025). Exploring EFL students' prompt engineering in human–AI story writing: An activity theory perspective. *Interactive Learning Environments, 33*(1), 863–882. https://doi.org/10.1080/10494820.2024.2361381

Woo, D. J., Wang, D., & Guo, K. (2026). Comparing Hong Kong secondary school students' perceptions of ChatGPT-assisted EFL writing. *The Journal of Educational Research, 119*(2), 153–173. https://doi.org/10.1080/00220671.2025.2525243

Zhuang, Y., Zhao, R., Xie, Z., & Yu, P. L. H. (2025). Enhancing language learning through generative AI feedback on picture-cued writing tasks. *Computers and Education: Artificial Intelligence, 9*, Article 100450. https://doi.org/10.1016/j.caeai.2025.100450

.