Pedagogical Alignment of LLMs requires Diverse Cognitively-Inspired Student Proxies

Suchir Salhan * [sas245@cam.ac.uk]
Andrew Caines [apc38@cam.ac.uk]
Paula Buttery [pjb48@cam.ac.uk]

ALTA Institute, Department of Computer Science & Technology
University of Cambridge, Cambridge, U.K.

Abstract

Large Language Models (LLMs) are increasingly positioned as tutors, judges, and instructional assistants. Yet their pedagogical performance remains shallow: they optimize for producing correct answers rather than for teaching. Pedagogy requires anticipating misconceptions, sequencing curricula, calibrating task difficulty, and adapting interactively to learner trajectories. In this position paper, we characterise the ideal behaviour of LLMs assuming a Teacher Role (pedagogical agents). Teaching involves selecting informative examples and learning from them as a basis for inference about what demonstrations a helpful teacher would provide. Given this, we argue that the pedagogical capabilities of LLMs as Teachers are limited due to the limited meta-reasoning capabilities of LLMs. We motivate structured *cognitively-inspired student proxies* as indispensable for pedagogical alignment. Student proxies are constrained cognitive models that generate structured, interpretable error trajectories. We argue that a *Teacher–Student Agentic Framework* (TSAF) with heterogeneous student proxies can enable teacher LLMs to improve their pedagogical alignment – by adapting reasoning and feedback strategies, monitoring errors, and scaffolding learning efficiently across tasks. Our position is that cognitive proxies reframe pedagogical alignment as an interpretable, principled paradigm for enhancing how LLMs "learn to teach".

1 When Correctness Isn't Enough: Meta-Reasoning Limits in LLM Tutoring

The vision of LLMs as automated tutors has captured both academic and commercial interest, yet despite high benchmark performance, they often lack pedagogical alignment: being correct does not guarantee being a good teacher. Recent industry efforts, such as OpenAI's Study Mode [37], Google's Guided Learning in Gemini [34, 27] and Anthropic's Claude for Education [1], attempt to walk learners through reasoning, provide informal checks, and incorporate visuals to reinforce understanding. However, models rarely probe learner goals ("Why does this matter to you?"), promote metacognition, or calibrate instruction beyond surface correctness. Furthermore, interaction can lean towards sycophancy rather than genuine engagement [19, 21, 28, 29, 31, 54].

Pedagogical alignment involves anticipating misconceptions, calibrating task difficulty, sequencing curricula, and scaffolding learner progress [46]. Cognitive science models, for instance Bayesian approaches, formalise this process by treating teaching and learning as rational communication, where agents select actions based on expected utility and reason about how information shifts beliefs [9, 11, 12, 14, 45]. These models have explained concept induction, distinctions between

* **Corresponding Author**

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models.

pedagogical and nonpedagogical contexts, and informed AI and robotics applications. Crucially, effective pedagogy is interactive: teachers diagnose errors and adapt dynamically, while learners provide feedback to shape instruction, establishing a common ground for iterative learning [5, 44].

Current LLM tutors, by contrast, rely on ad hoc heuristics such as self-correction, role-played feedback, and scripted personalization, lacking principled grounding in cognitive models [33, 47]. Instead, we advocate explicit *meta-reasoning*, where LLMs iteratively refine their reasoning strategies through interaction. We propose a Teacher-Student Agentic Framework (TSAF) that uses cognitively-inspired student proxies—simulated learners with memory constraints, staged curricula, or error tendencies—to provide structured, interpretable feedback. These proxies allow teacher LLMs to monitor error trajectories, update strategies probabilistically, and adapt policies efficiently, reframing LLM pedagogy as a principled, interpretable process rather than a series of surface-level heuristics.

2 Pedagogical Agents need Meta-Reasoning Capabilities

At the heart of interaction between teachers and students is recursive "reasoning about reasoning" – or **meta-reasoning** –, where individuals balance intuitive and deliberative processes, monitor their "feeling of knowing", and detect errors to adapt strategies. In both cases, communication or reasoning is not fixed but dynamically adjusted based on feedback about uncertainty and error [25, 49]. Some Cognitive Scientists have explicitly modelled meta-reasoning using Bayesian models of pedagogy to simulate how teachers choose examples by modelling the learner's beliefs, and learners infer the intended concept by reasoning about the teacher's intentions to infer student belief states and choose maximally informative interventions [51, 56]. Others applied these insights to develop AI-powered intelligent tutoring systems that utilise curriculum learning models to select examples that minimize uncertainty and accelerate progression [57, 58].

We suggest that the weaker pedagogical alignment of LLMs with learners derives – in part– from their poor meta-reasoning capabilities. Empirical work in NLP supports this limitation: Tyen et al. (2024) show that state-of-the-art LLMs struggled to locate errors in student reasoning chains [48], while Xu et al. (2024) found that LLMs often fail to verify the correctness of their reasoning steps, resulting in false conclusions [53]. The BEA 2025 Shared Task benchmarked over 50 LLM-based tutoring systems on mistake identification, mistake location, guidance, actionability, and tutor tone [33]. While models reached moderate F1 scores for mistake identification (up to 0.7181), they performed poorly on guidance and actionability, showing difficulty in producing targeted, context-sensitive feedback. Similarly, the BEA 2023 Shared Task [47] demonstrated that tutoring models often generate vague, incoherent, or overly generic responses. In both tasks, "mistakes" are broadly defined: wrong answers, partial misunderstandings, or reasoning errors. Crucially, assessment—tracking a learner's overall progress or competence—was absent from the shared task design, though it is essential for a complete tutoring solution. Analysis of over 50 LLM-based tutoring systems in the 2025 BEA task found that systems only achieved moderate F1 scores (58–72) across benchmarks, showing substantial room for improvement in pedagogical competence – detailed trends are discussed in *Appendix* A. State-of-the-art LLM tutors achieve moderate success in mistake identification (exact F1 up to 0.7181) but struggle with providing guidance, mistake location, and actionability, highlighting gaps in pedagogical alignment. These results demonstrate that accurate error detection alone is insufficient for effective tutoring, as most models fail to consistently deliver contextually targeted, actionable feedback that scaffolds student learning. The shared task also revealed that difficulty varies across LLMs and dialogue contexts. This behaviour reflects a lack of explicit learner modelling and meta-cognitive monitoring of both the student's reasoning and the model's own reasoning trajectory.

Recent advances in LLM reasoning demonstrate that **explicitly training models to self-verify and self-correct during inference** can substantially improve reasoning accuracy and reliability [23, 29, 42, 43, 44]. In this framework, models are first initialised with iterative self-verification and self-correction behaviours through supervised fine-tuning on carefully curated datasets of reasoning chains. These datasets often include stepwise solutions, annotated errors, and exemplar corrections, which explicitly teach the model to recognize intermediate mistakes and produce corrected outputs. Once initialised, these self-monitoring capabilities are further enhanced using both outcome-level and process-level reinforcement learning (RL). Outcome-level RL evaluates the correctness of the final answer, while process-level RL evaluates the quality of intermediate reasoning steps, ensuring that the model's internal reasoning trajectory is sound. Both leverage existing reasoning traces rather than generating fully new supervised data, allowing LLMs to adaptively refine reasoning during inference.

Meta-reasoning theories describe a similar loop at the individual level: people track uncertainty, detect errors, and balance intuitive and deliberative processes to regulate reasoning. The 2025 BEA Shared Task found LLM self-verification and self-correction enhanced mistake identification and location, as models were trained to detect and adjust intermediate errors, but models suffered when trained to detect and adjust intermediate errors. However, stepwise reasoning and self-correction are insufficient to improve LLM Teacher Agents' guidance and feedback actionability – models trained with self-correction may generate incomplete, vague, or hallucinated guidance, particularly for multi-step problems where intermediate steps are misjudged [30, 38] (see *Appendix* A and B for detailed qualitative analysis). Actionability is further challenged by novel tasks that fall outside the curated reasoning datasets, leading to generic or unhelpful next-step suggestions. This underscores the challenge of maintaining consistency across reasoning, guidance, and tone, highlighting the gap between correctness-focused meta-reasoning and holistic pedagogical quality.

We highlight four areas where progress is needed to move beyond correctness-focused reasoning. First, tutors must go beyond correcting their own errors to develop interpretable models of student dynamics: anticipating why learners make mistakes and adapting feedback accordingly [55]. Second, pedagogy operates at multiple scales, and effective teaching requires the aggregation of individual errors into coherent class-level strategies, something current LLM tutors cannot achieve [7]. Third, unlike human teachers who dynamically adjust task difficulty to scaffold learners in real time, LLMs lack robust mechanisms for adaptive task calibration [42]. Finally, systematic assessment and progression-tracking remain absent from most tutoring models and from the design of current shared tasks, despite being essential for supporting long-term learning [35].

3 Why Cognitively-Inspired Student Proxies Are Essential

While most teacher LLMs lack explicit learner models and act heuristically—simplifying tasks only after repeated errors, offering counterexamples opportunistically, and defaulting to role-play without structured awareness of instructional stages—some methods simulate student personas [29, 59]. Prompt-based methods can simulate personas in multi-agent classroom setups or varying skill levels, allowing teachers to interact with diverse learners. Yet these simulations often remain superficial: heuristic students produce noisy or incoherent errors, limiting interpretability.

In contrast, we argue that **student models generating structured error signals provide interpretable signals for hypothesis-driven experimentation** and a principled framework for studying instructional strategies with LLMs. We propose *error-structured student proxies*, lightweight cognitive models encoding systematic biases, memory constraints, and developmental trajectories. By constraining the space of possible errors, cognitive proxies can simulate biases such as overgeneralization or recency effects, enabling diagnostic correction and forcing teacher LLMs to implement strategies that generalize, anticipate misconceptions, and scaffold learning adaptively. This contrasts sharply with conventional role-playing, whose incoherent errors fail to provide interpretable signals.

To formalize these concepts, let a proxy S_θ produce an error trajectory $\mathcal{E}(S_\theta) = \{e_1, \ldots, e_T\}$, where $e_t \in \mathcal{Y}$ encodes systematic deviation from the target output at step t. Teacher LLMs T_ϕ observe $\mathcal{E}(S_\theta)$ and update strategies ϕ to minimize divergence from ideal teaching dynamics. *Developmental proxies* model stage-wise progression, allowing teachers to scaffold transitions from rote memorization to abstraction, while *domain-specific proxies* encode content-relevant errors. Creating such proxies is a major research challenge; however, language learning offers comparatively more progress. BabyLMs (small language models trained on developmentally plausible corpora) can serve as lightweight, developmentally constrained proxies, encoding linguistic patterns with limited data exposure that reflect early learner biases [8, 13, 16, 40, 41]. These proxies align naturally with modular, hypothesis-driven frameworks such as Pico [32], providing a sandbox for controlled experimentation for training and interpreting the learning dynamics of Small Language Models (SLMs), and with the perspective that SLMs are optimal for agentic AI applications [6]. SLMs have the additional advantage that they can be iteratively designed and evaluated according to their task-specialized agentic functions.

Beyond language, proxies for other domains can leverage models trained on domain-specific curricula, with constraints or noise introduced to simulate varied learner profiles. For instance in mathematics, models can be trained with limited or adversarial examples of particular problem types to induce common calculation or reasoning errors [e.g. 29, 50, 60], while in programming, proxies can simulate misunderstandings of control flow or type errors [e.g. 20, 24]. By varying data exposure, architecture,

or injected biases, one can systematically produce heterogeneous proxies $\{S_{\theta_i}\}$ sampled from a distribution $p(\theta)$, providing robust training signals that enable teacher strategies to generalize across learner populations. This formalization offers a concrete mechanism for designing and evaluating student proxies, ensuring teacher LLMs learn strategies that generalize beyond a single learner type or domain while maintaining interpretability and structured error modeling.

We conceptually posit that enhanced Pedagogical Alignment can be obtained through a multi-agent Teacher–Student Agentic Framework (TSAF), allowing teacher models to generalize instructional strategies across diverse learners and domains using multiple proxies generating structured error trajectories. TSAFs enable teachers to anticipate the developmental origins of mistakes rather than simply correcting them post hoc. Reward signals derived from proxy errors incentivize teachers to reduce variance and stabilize learning, *promoting meta-cognitive reasoning sensitive to why learners err, not just that they do*.

TSAFs can be implemented through Multi-Agent Reinforcement Learning, yet empirical benefits remain open, particularly whether a teacher LLM trained in a TSAF will learn strategies that transfer to real human learners, whether proxy heterogeneity prevents overfitting to narrow error patterns, and whether gains emerge from integrating explicit learner-state modeling inspired by Bayesian cognitive theories [14, 17]. Even so, TSAFs provide several concrete advantages. Proxies generate structured and interpretable error patterns at the population level, reducing the noise and incoherence inherent in role-played students whose mistakes often lack psychological plausibility. Teacher models internalize consistent signals and update teaching policies to minimize future errors through adaptive task selection and feedback modulation. By internalizing these structured dynamics, teacher models can calibrate strategies that generalize across diverse learner behaviors, forming coherent, classroom-level interventions that capture statistical regularities in learner errors. Proxies generate sequences of errors reflecting stage-wise learning progression, enabling teachers to adjust task difficulty and scaffold concepts dynamically. This mirrors human teaching strategies, where simpler tasks are presented for struggling learners and more advanced problems are introduced for those demonstrating mastery. Aligning teacher strategies with these evolving developmental trajectories ensures adaptive, anticipatory, and structured instruction rather than fixed or purely reactive guidance.

Most importantly, TSAFs supply a *principled mechanism for evaluating instructional strategies beyond correctness-based objectives*. By translating proxy error dynamics into interpretable and actionable signals, we can investigate whether Teacher LLMs can sequence interventions, stabilize instruction, and iteratively refine reasoning across tasks, learners, and developmental stages. While our proposal remains conceptual, it provides a feasible path toward empirical validation, defining success as strategy transfer across diverse proxies and improved outcomes with real learners: TSAFs create a scalable testbed for pedagogical alignment and a blueprint for extending cognitively inspired proxies to other domains through curricular constraints and controlled data exposure, operationalizing the bridge between cognitive theory and effective instructional behavior in LLM tutors.

4 Conclusion

We argue that the pedagogical alignment of LLMs can be improved by focusing on algorithmic processes inspired by cognitive models of student–teacher interaction. Analyses of LLM tutors show that accurate error detection alone is insufficient: effective tutoring also requires actionable, well-scaffolded, and encouraging feedback, combined with ongoing assessment of learner progress. Current LLMs display gaps in self-awareness and rely on rigid reasoning strategies, in contrast to Bayesian pedagogical agents (for example) that adaptively use structured heuristics derived from student behaviour. Looking ahead, the development of pedagogically competent agents will require advances in both modelling and evaluation. Empirical evaluation of Teacher-Student Agentic Frameworks to represent uncertainty and belief states in interaction for multi-agent pedagogical alignment is needed. Progress in pedagogical alignment also depends on analyses of teacher-student interactions in pedagogical settings (e.g., the Teacher-Student Chatroom Corpus [7], see *Appendix* C, *Table* 6) and taxonomies for domain-specific pedagogical evaluation of LLMs [33, 34], which can be repurposed into alignment datasets and benchmarks of multi-turn interaction [4, 10, 18, 39, 41, 52]. TSAFs with cognitively-inspired student proxies can help bridge this gap by providing interpretable signals of learner errors and trajectories, enabling LLM tutors to refine their strategies adaptively and enhancing meta-reasoning capabilities. Together, these directions chart a roadmap for developing tutoring systems that combine interpretability with genuine pedagogical alignment.

References

[1] Anthropic. Introducing Claude for Education. https://www.anthropic.com/news/introducing-claude-for-education, 2025. Accessed: 27 August 2025.

[2] T. Aoyama and N. Schneider. Modeling nonnative sentence processing with L2 language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4927–4940, Miami, Florida, USA, 2024. Association for Computational Linguistics.

[3] C. Arnett, T. A. Chang, J. A. Michaelov, and B. Bergen. On the acquisition of shared grammatical representations in bilingual language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20707–20726, Vienna, Austria, 2025. Association for Computational Linguistics.

[4] G. Bai, J. Liu, X. Bu, Y. He, J. Liu, Z. Zhou, Z. Lin, W. Su, T. Ge, B. Zheng, and W. Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.

[5] F. Balzan, P. P. Santos, M. Gabbrielli, M. Albarracin, and M. Lopes. A computational model of inclusive pedagogy: From understanding to application. *arXiv preprint arXiv:2505.02853*, 2025.

[6] P. Belcak, G. Heinrich, S. Diao, Y. Fu, X. Dong, S. Muralidharan, Y. C. Lin, and P. Molchanov. Small language models are the future of agentic AI. *arXiv preprint arXiv:2506.02153*, 2025.

[7] A. Caines, H. Yannakoudakis, H. Allen, P. Pérez-Paredes, B. Byrne, and P. Buttery. The Teacher-Student Chatroom Corpus version 2: more lessons, new annotation, automatic detection of sequence shifts. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 23–35, Louvain-la-Neuve, Belgium, 2022. LiU Electronic Press.

[8] L. Charpentier, L. Choshen, R. Cotterell, M. O. Gul, M. Hu, J. Jumelet, T. Linzen, J. Liu, A. Mueller, C. Ross, R. S. Shah, A. Warstadt, E. Wilcox, and A. Williams. BabyLM turns 3: Call for papers for the 2025 BabyLM Workshop. In *Proceedings of the 2025 BabyLM Workshop*, 2025.

[9] A. M. Chen, A. Palacci, N. Vélez, R. D. Hawkins, and S. J. Gershman. A hierarchical Bayesian model of adaptive teaching. *Cognitive Science*, 48(7):e13477, 2024.

[10] J. Chen, Z. Liu, M. Hou, X. Zhao, and W. Luo. Multi-turn classroom dialogue dataset: Assessing student performance from one-on-one conversations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5333–5337, 2024.

[11] M. T. H. Chi, S. A. Siler, H. Jeong, T. Yamauchi, and R. G. Hausmann. Learning from human tutoring. *Cognitive Science*, 25(4):471–533, 2001.

[12] C. M. Clark. Teachers' thought processes. Technical Report 72, Institute for Research on Teaching, Michigan State University, 1984.

[13] R. Diehl Martinez, Z. Goriely, H. McGovern, C. Davis, A. Caines, P. Buttery, and L. Beinborn. CLIMB – curriculum learning for infant-inspired model building. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, December 2023.

[14] Y. Fan, F. Tian, T. Qin, X. Y. Li, and T. Y. Liu. Learning to teach. In *International Conference on Learning Representations*, February 2018.

[15] R. Gao, X. Wu, T. Kuribayashi, M. Ye, S. Qi, C. Roever, Y. Liu, Z. Yuan, and J. H. Lau. Can LLMs simulate L2-English dialogue? an information-theoretic analysis of L1-dependent biases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4355–4379, Vienna, Austria, 2025. Association for Computational Linguistics.

[16] Y. Gao, S. Salhan, A. Caines, W. Sun, and P. Buttery. BLiSS: Evaluating bilingual learner competence in second language small language models. In *Proceedings of the 3rd BabyLM Workshop*, November 2025.

[17] H. Ham, B. Zhao, T. L. Griffiths, and N. Vélez. Teaching recombinable motifs through simple examples. *Cognitive Science*, 49(8):e70103, 2025.

[18] L. He, M. Mavrikis, and M. Cukurova. Towards mining effective pedagogical strategies from learner–LLM educational dialogues. In *International Conference on Artificial Intelligence in Education*, pages 391–396. Springer, 2025.

[19] F. Ikram, A. Scarlatos, and A. Lan. Exploring LLMs for predicting tutor strategy and student outcomes in dialogues. In *Proceedings of BEA 2025: 20th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 765–779, Vienna, Austria, 2025. Association for Computational Linguistics.

[20] H. Jin, S. Lee, H. Shin, and J. Kim. Teach AI how to code: Using large language models as teachable agents for programming education. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[21] H. Jin, M. Yoo, J. Park, Y. Lee, X. Wang, and J. Kim. TeachTune: Reviewing pedagogical agents against diverse student profiles with simulated students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*, pages 1–28. ACM, 2025.

[22] J. Jumelet, L. Weissweiler, and A. Bisazza. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint*, arXiv:2504.02768, 2025.

[23] R. Kamoi, Y. Zhang, N. Zhang, J. Han, and R. Zhang. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

[24] M. Kazemitabaar, R. Ye, X. Wang, A. Z. Henley, P. Denny, M. Craig, and T. Grossman. CodeAid: Evaluating a classroom deployment of an LLM-based programming assistant that balances student and educator needs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[25] P. Kirschner, J. Sweller, and R. E. Clark. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2):75–86, 2006.

[26] E. Kochmar, K. Maurya, K. Petukhova, K. A. Srivatsa, A. Tack, and J. Vasselli. Findings of the BEA 2025 shared task on pedagogical ability assessment of AI-powered tutors. In E. Kochmar, B. Alhafni, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, and Z. Yuan, editors, *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1011–1033, Vienna, Austria, July 2025. Association for Computational Linguistics.

[27] LearnLM Team, A. Modi, A. S. Veerubhotla, A. Rysbek, A. Huber, A. Anand, A. Bhoopchand, B. Wiltshire, D. Gillick, D. Kasenberg, et al. Evaluating Gemini in an arena for learning. *arXiv preprint*, arXiv:2505.24477, 2025.

[28] M. Lelièvre, A. Waldock, M. Liu, N. V. Aspillaga, A. Mackintosh, M. J. O. Portela, J. Lee, P. Atherton, R. A. Ince, and O. G. Garrod. Benchmarking the pedagogical knowledge of large language models. *arXiv preprint arXiv:2506.18710*, 2025.

[29] J. Liu, Z. Huang, T. Xiao, J. Sha, J. Wu, Q. Liu, S. Wang, and E. Chen. SocraticLM: Exploring Socratic personalized teaching with large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 85693–85721, 2024.

[30] Z. Ma, Q. Yuan, Z. Wang, and D. Zhou. Large language models have intrinsic meta-cognition, but need a good lens. *arXiv preprint*, 2506.08410, 2025.

[31] J. Macina, N. Daheim, I. Hakimi, M. Kapur, I. Gurevych, and M. Sachan. MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. *arXiv preprint*, 2502.18940, 2025.

[32] R. D. Martinez, D. D. Africa, Y. Weiss, S. Salhan, R. Daniels, and P. Buttery. Pico: A modular framework for hypothesis-driven small language model research. In *Proceedings of the EMNLP 2025 Systems Demonstrations*, Suzhou, China, Nov 2025.

[33] K. K. Maurya, K. A. Srivatsa, K. Petukhova, and E. Kochmar. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico, 2025.

[34] A. Modi, A. S. Veerubhotla, A. Rysbek, A. Huber, B. Wiltshire, B. Veprek, D. Gillick, D. Kasenberg, D. Ahmed, I. Jurenka, J. Cohan, J. She, J. Wilkowski, K. Alarakyia, K. R. McKee, L. Wang, M. Kunesch, M. Schaekermann, M. Pîslar, N. Joshi, P. Mahmoudieh, P. Jhun, S. Wiltberger, S. Mohamed, S. Agarwal, S. M. Phal, S. J. Lee, T. Strinopoulos, W.-J. Ko, A. Wang, A. Anand, A. Bhoopchand, D. Wild, D. Pandya, F. Bar, G. Graham, H. Winnemoeller, M. Nagda, P. Kolhar, R. Schneider, S. Zhu, S. Chan, S. Yadlowsky, V. Sounderajah, and Y. Assael. LearnLM: Improving Gemini for Learning. *arXiv preprint*, 2412.16429, 2025.

[35] R. Moore, A. Caines, M. Elliott, A. Zaidi, A. Rice, and P. Buttery. Skills embeddings: A neural approach to multicomponent representations of students and tasks. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM)*. International Educational Data Mining Society, 2019.

[36] M. Oba, T. Kuribayashi, H. Ouchi, and T. Watanabe. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572, Toronto, Canada, 2023. Association for Computational Linguistics.

[37] OpenAI. Introducing study mode. https://openai.com/index/chatgpt-study-mode/, 2025. Accessed: 27 August 2025.

[38] N. Otero, S. Druga, and A. Lan. A benchmark for math misconceptions: bridging gaps in middle school algebra with AI-supported instruction. *Discover Education*, 4(1):277, 2025.

[39] J. Perczel, J. Chow, and D. Demszky. TeachLM: Post-training LLMs for education using authentic learning data. *arXiv preprint arXiv:2510.05087*, 2025.

[40] S. Salhan, R. Diehl Martinez, Z. Goriely, and P. Buttery. Less is more: Pre-training cross-lingual small-scale language models with cognitively-plausible curriculum learning strategies. In *Proceedings of the 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, November 2024.

[41] S. Salhan, H. Gu, D. Rooein, D. Galvan-Sosa, G. Gaudeau, A. Caines, Z. Yuan, and P. Buttery. Teacher demonstrations in a BabyLM's Zone of Proximal Development for contingent multi-turn interaction. In *Proceedings of the 3rd BabyLM Workshop*, November 2025.

[42] A. Scarlatos, R. S. Baker, and A. Lan. Exploring knowledge tracing in tutor-student dialogues using llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*, pages 249–259, New York, NY, USA, 2025. ACM.

[43] A. Scarlatos, N. Fernandez, C. Ormerod, S. Lottridge, and A. Lan. SMART: Simulated students aligned with item response theory for question difficulty prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

[44] A. Scarlatos, N. Liu, J. Lee, R. Baraniuk, and A. Lan. Training LLM-based tutors to improve student learning outcomes in dialogues. In *Proceedings of AIED 2025*, 2025.

[45] P. Shafto and N. D. Goodman. Teaching and the Bayesian learner. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 1632–1637, 2008.

[46] S. Sonkar, K. Ni, S. Chaudhary, and R. Baraniuk. Pedagogical alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13641–13650, Miami, Florida, USA, 2024.

[47] A. Tack, E. Kochmar, Z. Yuan, S. Bibauw, and C. Piech. The BEA 2023 shared task on generating AI teacher responses in educational dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 785–795, Toronto, Canada, 2023. Association for Computational Linguistics. https://aclanthology.org/2023.bea-1.64/.

[48] G. Tyen, H. Mansoor, V. Carbune, P. Chen, and T. Mak. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13894–13908, 2024.

[49] L. S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*, volume 86. Harvard University Press, 1978.

[50] R. Wang, Q. Zhang, C. Robinson, S. Loeb, and D. Demszky. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199, Mexico City, Mexico, 2024.

[51] R. E. Wang, M. Wu, and N. Goodman. Know thy student: Interactive learning with Gaussian processes. *arXiv preprint*, arXiv:2204.12072, 2022.

[52] S. Wei, M. Zhang, X. Lin, B. Jiang, Z. Dai, and K. Kuang. EduDial: Constructing a large-scale multi-turn teacher-student dialogue corpus. *arXiv preprint*, arXiv:2510.12899, 2025.

[53] Q. Xu and J. Zhu. Reliability evaluation of ideological and political teaching effect in colleges and universities based on Bayesian inference. In *Proceedings of the 2025 International Conference on Digital Education and Information Technology*, pages 6–10, February 2025.

[54] H. Yan, L. Zhang, J. Li, Z. Shen, and Y. He. Position: LLMs need a Bayesian meta-reasoning framework for more robust and generalizable reasoning. In *2025 International Conference on Machine Learning: ICML25*, 2025.

[55] M. Yoo, H. Jin, and J. Kim. How do teachers create pedagogical chatbots?: Current practices and challenges. In *Proceedings of the CHI 2025 Workshop on Augmented Educators and AI*, 2025.

[56] L. Yuan, D. Zhou, J. Shen, J. Gao, J. L. Chen, Q. Gu, and S. Zhu. Iterative teacher-aware learning. *Advances in Neural Information Processing Systems*, 34:29231–29245, 2021.

[57] A. Zaidi, A. Caines, C. Davis, R. Moore, P. Buttery, and A. Rice. Accurate modelling of language learning tasks and students using representations of grammatical proficiency. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM)*. International Educational Data Mining Society, 2019.

[58] A. Zaidi, A. Caines, R. Moore, P. Buttery, and A. Rice. Adaptive forgetting curves for spaced repetition language learning. In C. Conati, N. Heffernan, A. Mitrovic, and M. Verdejo, editors, *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II*. Springer, 2020.

[59] Z. Zhang, D. Zhang-Li, J. Yu, L. Gong, J. Zhou, Z. Hao, J. Jiang, J. Cao, H. Liu, Z. Liu, L. Hou, and J. Li. Simulating classroom education with LLM-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379, Albuquerque, New Mexico, 2025. Association for Computational Linguistics.

[60] Z. Zhou, Q. Wang, M. Jin, J. Yao, J. Ye, W. Liu, W. Wang, X. Huang, and K. Huang. MathAttack: Attacking large language models towards math solving ability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19750–19758, 2024.

A BEA 2025 Shared Task Findings

A.1 Evaluation Taxonomy of Maurya et al (2025)

Table 1: An overview of the proposed evaluation taxonomy of Maurya et al (2025) used in the BEA 2025 Shared Task [26, 33]

Dimension	Definition	Desiderata
Mistake identification	Has the tutor identified/recognized a mistake in a student's response?	Yes
Mistake location	Does the tutor's response accurately point to a genuine mistake and its location?	Yes
Revealing of the answer	Does the tutor reveal the final answer (whether correct or not)?	No
Providing guidance	Does the tutor offer correct and relevant guidance, such as an explanation, elaboration, hint, examples, and so on?	Yes
Actionability	Is it clear from the tutor's feedback what the student should do next?	Yes
Coherence	Is the tutor's response logically consistent with the student's previous responses?	Yes
Tutor tone	Is the tutor's response encouraging, neutral, or offensive?	Encouraging
Human-likeness	Does the tutor's response sound natural rather than robotic or artificial?	Yes

A.2 Performance across Pedagogical Dimensions

Table 2: Performance of LLM-based tutors across pedagogical dimensions.

Dimension	Dataset	Maj-Class Ex. F1	Maj-Class Ex. Acc	Best Ex. F1	Best Ex. Acc	Best Len. F1	Best Len. Acc
Mistake Identification	Dev	0.2922	0.7803	–	–	0.4596	0.8506
	Test	0.2827	0.7363	0.7181	0.8798	0.9185	0.9541
Mistake Location	Dev	0.2560	0.6232	–	–	0.4159	0.7120
	Test	0.2450	0.5811	0.5983	0.7679	0.8404	0.8630
Providing Guidance	Dev	0.2416	0.5683	–	–	0.4355	0.7714
	Test	0.2313	0.5314	0.5834	0.7052	0.7860	0.8222
Actionability	Dev	0.2307	0.5291	–	–	0.4041	0.6781
	Test	0.2198	0.4919	0.7085	0.7557	0.8659	0.8940
Tutor Identification	Dev	0.0240	0.1212	–	–	–	–
	Test	0.0244	0.1235	0.9698	0.9664	–	–

Table 3: Performance of LLM-based tutors across pedagogical dimensions on development (Dev) and test (Test) sets. The evaluation considers (i) **Mistake Identification**: whether the tutor recognizes a student's mistake, (ii) **Mistake Location**: whether the tutor correctly pinpoints the error, (iii) **Providing Guidance**: whether the tutor offers relevant explanations or hints, (iv) **Actionability**: whether the feedback makes clear what the student should do next, and (v) **Tutor Identification**: whether the response is attributable to the tutor rather than the student. Reported metrics include majority-class exemplar performance (Maj-Class Ex.), best exemplar performance (Best Ex.), and best length-controlled performance (Best Len.), each measured by F1 and accuracy (Acc).

A.3 Variability of LLMs across Dialogue Contexts

Responses from models like Llama-3.1-8B and Gemini were particularly difficult for participants to classify accurately, showing misalignment rates above 40%, even when compared to expert human tutors with a 37% misalignment rate. This variability indicates that LLMs lack robust internal models of learner states and often generate responses heuristically, failing to consistently adapt feedback to subtle student errors. Open-ended guidance tasks were particularly error-prone, emphasizing

that pedagogical alignment involves not just correct answers, but structured reasoning about student understanding and learning trajectories.

A.4 Qualitative Comparison of Reasoning Techniques and Pedagogical Ability Assessment of LLMs

Table 4: Synoptic Analysis of BEA 2025 Findings of the impact of different LLM reasoning techniques—including self-verification, self-correction, and reinforcement learning (RL)— on the evaluation dimensions of the proposed taxonomy for assessing pedagogical abilities of AI tutors. The table also highlights relative brittleness of each dimension.

Evaluation Dimension	**Techniques Addressing Dimension**	**Contribution / Mechanism**	**Relative Brittleness / Weaknesses**
Mistake Identification	Self-verification, Ensembling, Zero-shot prompting, SFT	Detects errors in student responses; ensembling improves coverage; self-verification catches subtle mistakes	May miss rare/unconventional errors; zero-shot brittle; self-verification limited by training data diversity
Mistake Location	Self-verification, Process-level RL, Chain-of-thought, Multi-step prompting, Ensembling	Pinpoints errors within response; process-level RL evaluates intermediate reasoning; chain-of-thought supports stepwise analysis	Sensitive to input phrasing; multi-step reasoning can propagate errors; process-level RL depends on accurate intermediate assessment
Providing Guidance	Chain-of-thought, Multi-step prompting, Self-verification, Ensembling, Task-aware prompting	Produces coherent, stepwise guidance; self-correction refines explanations; ensembling stabilizes outputs	Guidance may be vague, incomplete, or hallucinated; stepwise reasoning can fail on complex or unusual problems
Actionability	Self-verification, Outcome-level RL, Ensembling, Task-aware prompting	Suggests actionable next steps; RL optimizes relevance; self-monitoring ensures consistency with reasoning	Ambiguous instructions; generic advice on novel tasks; RL sensitive to reward design and dataset coverage
Tutor Tone	SFT, RLHF, Ensembling, Style-focused fine-tuning	Encourages natural, consistent, and supportive responses; ensembling mitigates stylistic inconsistencies	Tone not directly optimized by self-verification; can revert to neutral/overly formal under adversarial inputs
Coherence	Chain-of-thought, Multi-step prompting, Ensembling	Ensures logically consistent responses across steps	Sensitive to prompt variations; reasoning chains can diverge or contradict if intermediate steps are misjudged

B Pedagogical Evaluation Framework for Google's Guided Learning for Gemini (LearnLM)

The LearnLM Team at Gemini develop a pedagogical rubric for 189 educators to evaluate five models (Gemini 2.5 Pro, Claude 3.7 Sonnet, GPT-4o, and OpenAI o3). We summarise the findings of this qualitative evaluation across individual criteria items in *Table* 5 [34, 27]. This evaluation differs from the domain-specific evaluation of the BEA 2025 Shared Task (summarised in *Table* 4) as the LearnLM team prioritise diverse, contextually relevant assessment across 49 realistic learning scenarios.

Table 5: **LearnLM Pedagogical Evaluation Rubric with Model Weaknesses.** The rubric items (left columns) are rated by educators on a seven-point Likert-type scale. The third column summarizes model-specific weaknesses based on masked IDs (A=ChatGPT-4o, B=Claude 3.7 Sonnet, C=Gemini 2.5 Pro, D=GPT-4o, E=OpenAI o3). Models A, B, D, and E frequently gave answers directly, struggled with engagement, and were often described as robotic, overwhelming, or off-topic. Model C (Gemini 2.5 Pro) was consistently praised as the strongest pedagogical tutor, showing scaffolding, questioning, and adaptability, though it could sometimes be verbose or overly structured. Rows in pastel red indicate **consistent weaknesses across models**, orange indicate **mixed performance**, and green rows highlight **relative strengths of Gemini 2.5 Pro (Model C) compared to other models**.

Principle & Criterion Item	Item Wording	Model Weaknesses (A–E)
Principle: Manages cognitive load		
Appropriate response length	Responses are an appropriate length for the student.	A/D/E: Often too long or overwhelming; B: Sometimes terse but still "answer dumps"; C: Occasionally verbose but clearer.
Manageable chunks	Uses formatting/bullets to break down information.	A: Emojis/formatting divisive; B: Large text blocks; C: Generally well-structured; D/E: Poor formatting, walls of text.
Straightforward response	Responses are clear and easy to follow.	A: Friendly but sometimes confusing; B: Overly blunt/robotic; C: Clear, concise; D/E: Long, complex, unhelpful.
No irrelevant information	Avoids irrelevant digressions.	A/D/E: Let students go off-topic; B: Distracted by unrelated topics; C: Strong at redirecting.
Analogies	Uses narratives or analogies effectively.	A: Sometimes helpful (pizza analogy); Others: Rarely used; C: Effective scaffolding instead.
Information presentation	Presents in appropriate style/structure.	A: Engaging style, but over-peppy; B: Robotic walls of text; C: Clear scaffolds; D/E: Disorganized.
Information order	Explanations build logically.	A/E: Sometimes confusing order; B/D: Jumps to answers; C: Builds progressively.
No repetition	Avoids unnecessary repetition.	C: Occasionally repetitive; A–E: Some redundancy but less systematic.
No contradiction	Avoids contradicting earlier info.	D/E: Sometimes contradicted self; Others: Mostly consistent.
Principle: Inspires active learning		
Opportunities for engagement	Provides opportunities for student engagement.	A/B/D/E: Rare, often give answers directly; C: Strongest at prompting engagement.
Asks questions	Encourages student to think via questions.	A/B/D/E: Few or superficial; C: Frequently praised for questioning.
Guides to answer	Avoids giving answers away too quickly.	A/B/D/E: Main weakness—gave away answers; C: Guides and scaffolds.

Continued on next page

Table 5 – *Continued from previous page*

Principle & Criterion Item	**Item Wording**	**Model Weaknesses (A–E)**
Active engagement	Promotes active work with material.	A/B/D/E: Often passive "answer machines"; C: Encourages active student role.
Principle: Deepens metacognition		
Guide mistake discovery	Helps students discover mistakes.	A/B/D/E: Rarely encouraged reflection; C: Guided error discovery effectively.
Constructive feedback	Provides clear, constructive feedback.	A: Sometimes too positive/childish; B/D/E: Generic or unhelpful feedback; C: Clear and encouraging.
Acknowledge correctness	Acknowledges student's correct responses.	All models did this inconsistently; C: Best balance of praise + correction.
Communicates plan	Outlines objectives/plan for session.	A–E: Weakness across board; C: Occasionally explicit.
Principle: Stimulates curiosity		
Stimulates interest	Tries to spark curiosity.	A: Personable, sometimes engaging; B: Informational but not inspiring; C: Encouraged curiosity via scaffolding; D/E: Flat, machine-like.
Adapts to affect	Responds to frustration/discouragement.	A: Friendly tone helped; B/D/E: Often detached/robotic; C: Adaptive and patient.
Encouraging feedback	Feedback delivered in encouraging way.	A: Sometimes too peppy/childish; B/D/E: Flat or perfunctory; C: Balanced encouragement.
Principle: Adapts to learner		
Leveling	Explanations fit the student's level.	A: Sometimes overwhelming; B: Mixed clarity; C: Strong adaptation; D/E: Overly complex.
Unstuck	Adapts approach when student is stuck.	A/B/D/E: Often failed to adapt; C: Effective scaffolding.
Adapts to needs	Overall adapts to student needs.	A/B/D/E: Often rigid; C: Most flexible.
Proactive	Proactively guides conversation.	A: Sometimes proactive, often off-topic; B/D/E: Reactive, not proactive; C: Proactive and responsive.
Guides appropriately	Does not withhold information unproductively.	A/B/D/E: Gave away answers prematurely; C: Balanced guidance without over-revealing.

C Case Study: TSAF for Second Language (L2) Learning

We briefly outline the conceptual benefits of TSAF to potentially enhance pedagogical alignment of LLMs in the domain-specific context of second language (L2) learning, as a case study of the more general proposed argued for in *Section* 3.

C.1 BabyLM-inspired Second Language Student Proxies

The BabyLM Challenge incentivises small-scale resource-constrained language model pretraining on corpora that are both limited in *scale* and *domain*. While the shared task currently focuses on L1 monolingual acquisition of English, there have been a few extensions of the cognitively-inspired language modelling paradigm to model second language (L2) acquisition, or L2LMs [2, 3, 36]. We argue that these small-scale L2LMs are effective candidate student proxies, since they can be pretrained on learner corpora, different L1 backgrounds, or simulate L1 learner error profiles.

While Gao et al (2025) argue that LLMs can adequately simulate L2 English dialogues [15], BabyLM-style L2LMs offer more control and interpretability. In particular, Arnett et al (2025) develop Bilingual GPT (B-GPT), which varies three factors for bilingual pretraining: language pair, language exposure order, and bilingual training condition, resulting in a total of 16 distinct models for English, Spanish, Greek and Portuguese [3]. Bilingual SLMs simulate two language acquisition scenarios:

1. **Simultaneous bilinguals**, where models train on **only L1 data during the first half** and then on a **mix of L1 and L2 data**, and
2. **Sequential bilinguals**, where models train on L1 only initially and then switch exclusively to L2.

Each B-GPT model is based on an autoregressive GPT-2 Transformer architecture with 124M Parameters. Sequences consist of 128 tokens, all monolingual per sequence, with mixed batches in bilingual conditions alternating between L1 and L2. Checkpoints are saved regularly, especially around the introduction of L2 data, to closely track the models' learning dynamics and crosslingual transfer behaviour. Throughout training, loss patterns and model checkpoints are saved enabling detailed learning dynamics analysis.

B-GPT-style architectures provide rich learning dynamics analyses for cognitively-inspired L2LM pretraining. We apply the MultiBLiMP 1.0 evaluation metric (Jumelet et al. 2025, [22]) on all the B-GPT Models pretrained from scratch by Arnett et al. (2025) [3]. In particular, it is possible to identify distinct timing effects on MultiBLiMP 1.0 performance between B-GPT's Simultaneous (Bilingual) and Sequential (late L2) modelling strategies (*Figure* 1). The Multi-Blimp 1.0 benchmark is only limited to Subject-Verb agreement, so cannot be taken as a comprehensive syntactic benchmark, but there are very distinct timing effects that distinguish Simultaneous/Sequential B-GPT profiles. However, B-GPT only investigates Indo-European L1s (Romance–Spanish, Germanic–Dutch, Slavic–Polish, Hellenic–Greek), leaving it an open question about the effects of typological transfer between L1 and L2 (see *Figure* 2).

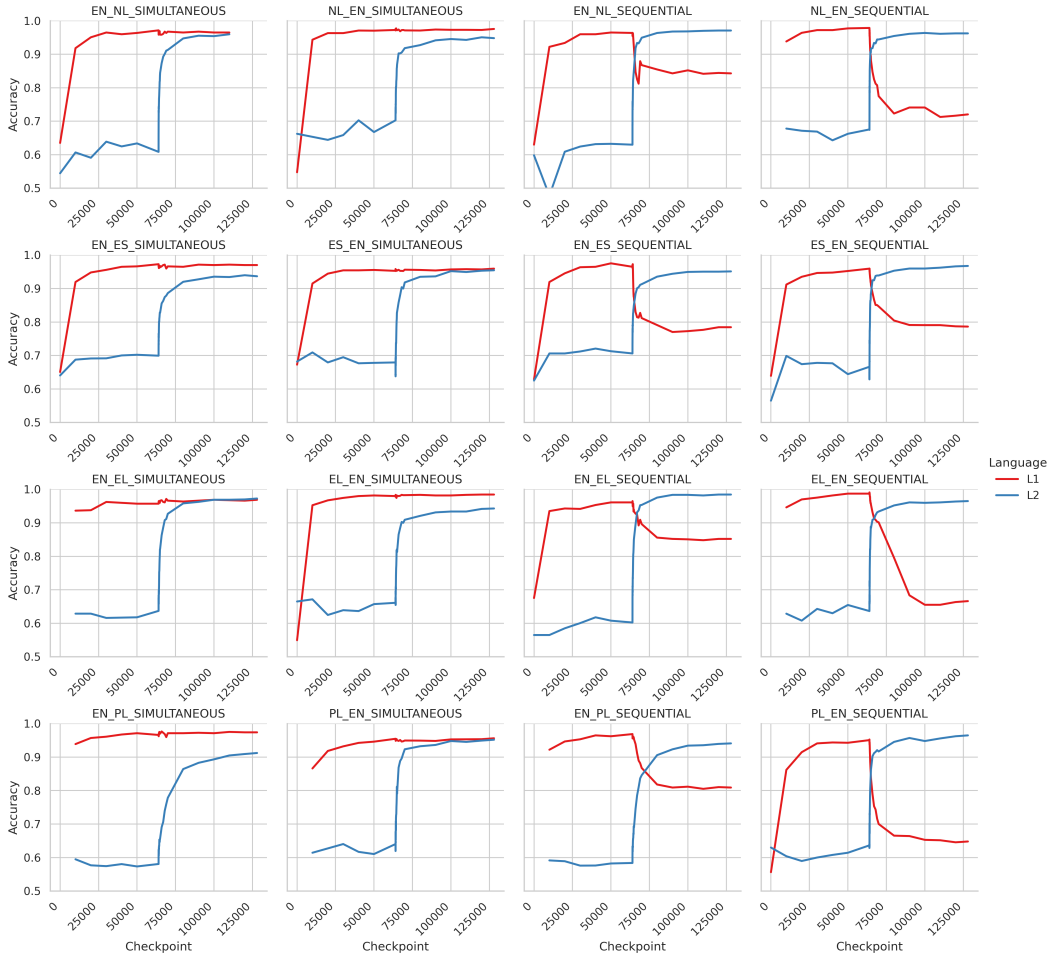

Figure 1: **Bilingual and L2LM Timing Effects: Differences between SIMULTANEOUS (Bilingual) and SEQUENTIAL (L2) Pretraining Regimes**. Five languages – {English (en), Dutch (nl), Spanish (es), Polish (pl), Greek (el)}. Evaluation uses MULTIBLIMP 1.0 [57] for 16 B-GPT SLMs [56]. **Results: L2s in SIMULTANEOUS B-GPT models** have **Phase 1:** lower MULTIBLIMP 1.0 Accuracy, and in **Phase 2** see a *sharp accuracy increase* in the final stages of pretraining. Meanwhile, SEQUENTIAL (L2LM) pretraining sees a divergent MULTIBLIMP 1.0 profile with **Phase 1:** high L1 accuracy and low L2 accuracy and **Phase 2:** *attrition* of L1 capabilities in conjunction with a sharp increase in L2 performance.

C.2 LLM-L2LM TSAFs for Language Learning

A possible TSAF-based framework might follow a **multi-agent classroom simulation framework**. This was recently proposed by Zhang et al (2025) in their SIMCLASS framework which uses prompt-based personality student simulation: **Teacher Personalities:** Teacher Agent and Assistant Agent. **Student Personalities:** Class Clown, Deep Thinker (students who raise questions to challenge classroom knowledge), Note Taker + Inquisitive Mind (students that summarise information or pose questions about content) [59]. They validate their framework through a user study of 400 university students who interact with Teacher and Student Agents in SIMCLASS. A similar framework SOCRATICLM introduces a pedagogically-inspired teaching paradigm that fulfils the role of a real classroom teacher in actively engaging students in thought-driven inquiry [29] (*Figure* 3 for an illustration).

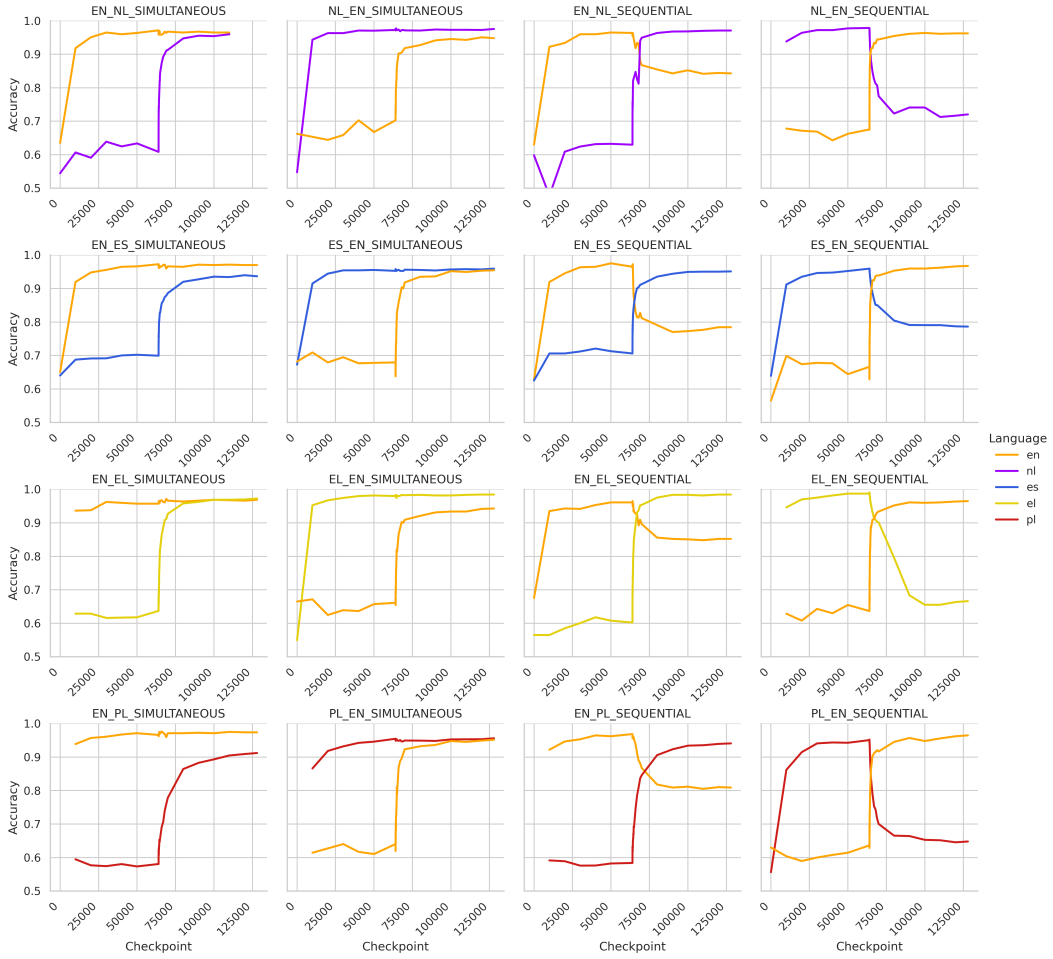

Figure 2: **Effect of L1 Background (Typological Transfer Effects) on Bilingual/L2LM Pretraining Regimes** for English and {Dutch (nl), Spanish (es), Polish (pl), Greek (el)}. Evaluation uses MULTIBLIMP 1.0 [22] for 16 bilingual SLMs [3]. Colours highlight differences in L1 background. There appears to be some initial **typological transfer**: English MULTIBLIMP performance drops further for Greek (el, Hellenic) and Polish (en-pl, Slavic) SEQUENTIAL `{pt, el}_en` training. In SIMULTANEOUS training, English and {Dutch, Spanish} finish at similar MULTIBLIMP performance, but this is not the case for {Greek (el-en, Hellenic), Polish (en-pl, Slavic)}.

A TSAF framework for pedagogical alignment of LLMs, designed to simulate L2 learner capabilities through a multi-agent L2LM–LLM setup, relies on structured information about effective teaching strategies. The concept of Teacher Demonstrations within a BabyLM's Zone of Proximal Development (ZPD) [41] provides a natural foundation for this approach. In CONTINGENTCHAT, multi-turn interactions between a BabyLM and a Teacher LLM scaffold the BabyLM's generation capabilities, gradually improving grammaticality, coherence, and contingent responses through targeted post-training. By constraining Teacher interventions to the learner's ZPD, the BabyLM receives support that is neither too trivial nor overly complex, enabling incremental learning consistent with Vygotsky's principle [49] of the Zone of Proximal Development.

Extending this idea to TSAFs, multi-agent L2LM–Teacher setups can simulate second-language learner behavior while providing structured pedagogical scaffolding. Here, datasets that capture realistic teacher-student interactions are crucial. The Teacher-Student Chatroom Corpus (TSCC) [7] is a collection of 102 online English lessons between 2 teachers and 8 students, comprising 13.5K conversational turns and 133K word tokens, with students at CEFR levels B1, B2, and C1. TSCC includes annotations of conversational sequence types and teaching focus, and TSCC v2 further

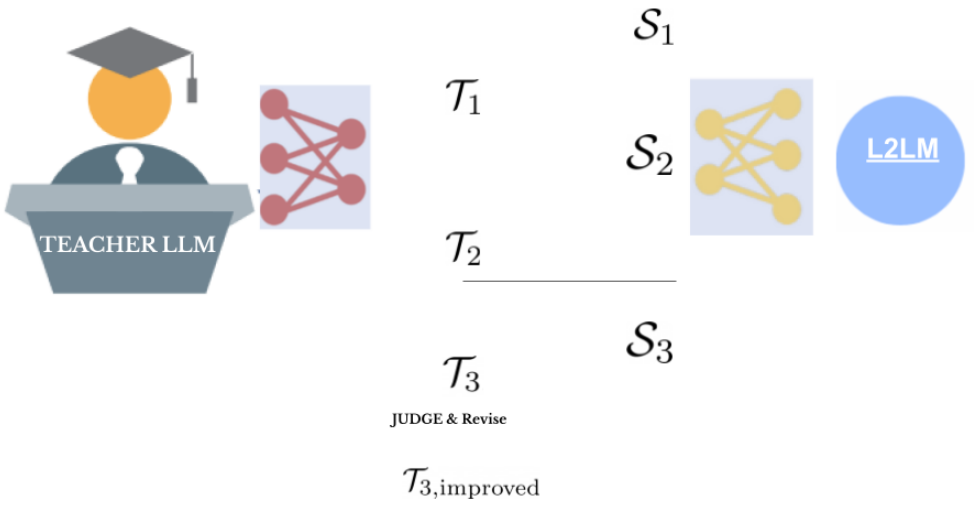

Figure 3: **Schematic for a Proposed L2LM-LLM TSAF:** Agents (TEACHER AND STUDENT) interact in a multi-turn dialogue. A query is initiated by a STUDENT L2LM-agent, answered by a LLM TEACHER. The DEAN scores the Teacher responses for Relevance and Correctness. Based on Feedback, the TEACHER revises their response $\mathcal{T}_3 \rightarrow \mathcal{T}_{3,\text{improved}}$ to the STUDENT. Unlike SocraticLM where agents are simulated by prompting GPT-4 to "adopt" different roles [18, student proxies are BabyLM-style cognitive proxies of second language learning capabilities (e.g., mirroring error profiles of learners based on L1 background or competence)

provides detailed teacher-student dialogues and pedagogical observations (*Table* 6), offering rich structured signals for aligning Teacher LLM interventions with learner needs.

By integrating ZPD-inspired scaffolding with multi-turn L2LM–LLM interactions, TSAFs can model structured learning trajectories, dynamically adjust task difficulty, and generate interpretable error signals. Teacher LLMs are thus able to provide developmentally appropriate interventions, stabilize learning across multiple simulated learners, and support generalizable instructional strategies. This creates a principled framework for pedagogical alignment, operationalizing theory into actionable signals that guide Teacher LLM behavior.

Future work could investigate the relative prioritization of different TSCC modes for specific L2LM–LLM configurations, allowing Teacher LLMs to be aligned with particular learner profiles as simulated by BabyLM-inspired L2LMs. Additionally, the learning dynamics and rich checkpointing of B-GPT-style L2LMs provide an additional source of structured information for refining pedagogical strategies, further enhancing the alignment between Teacher LLMs and learner models.

Teacher-Student Chatroom Corpus (TSCC) Annotation Framework [7]

Mode / Interacture	Description
Managerial	To transmit information, refer learners to materials, introduce or conclude an activity, or change from one mode of learning to another.
Classroom context	To enable learners to express themselves clearly, establish a context, and promote oral fluency.
Materials	To provide language practice around a piece of material, elicit responses in relation to the material, check and display answers, and clarify if needed.
Skills & systems	To enable learners to produce correct forms, manipulate target language, provide corrective feedback, and display correct answers. *Note: For annotation purposes, this mode is merged with 'Materials'.*
Confirmation check (CC)	Teacher confirms understanding of learner's utterance, or vice versa.
Display question (DQ)	A question to which the teacher already knows the answer.
Direct repair (DR)	Teacher corrects an error quickly and directly.
Enquiry (EN)	Learner asks a language question.
Extended teacher/ learner turn (ExtT)	Turn containing multiple main clauses, many relative clauses, at least one long relative clause, or a combination of such clauses.
Form-focused feedback (FBF)	Teacher gives explicit feedback on the words or forms used by the learner, rather than intended meaning.
Instruction (IN)	Teacher gives direct instructions.
Referential question (RQ)	A question to which the teacher does not know the answer, encouraging extended learner turns.
Scaffolding: Extension (S:E)	Teacher does not accept a learner's first answer, implicitly or explicitly encouraging more output.
Scaffolding: Modelling (S:M)	Teacher provides an example of the target language feature.
Scaffolding: Presentation (S:P)	Teacher explains a language point.
Seeking clarification (SC)	Teacher asks a student to clarify something the student has said, or vice versa.

Table 6: Self-Evaluation of Teacher Talk framework (SETT) modes and interactures annotated in the Teacher-Student Chatroom Corpus (TSC. Interactures (`CC`, `DQ`, `DR`, `EN`, `FBF`, `S:E`, `S:M`, `S:P`, `SC`) were added in the TSCC [7].