

Pedagogical Alignment of LLMs requires Diverse Cognitively-Inspired Proxies

Suchir Salhan, Andrew Caines & Paula Buttery

sas245@cam.ac.uk

How can we improve LLMs in pedagogical settings?



Problem: Pedagogical Agents need improved Meta-Reasoning

Multi-Turn Interaction between Teachers and Students in pedagogical setting should be dynamically adjusted based on feedback about uncertainty and error based on student responses.



LLMs struggle to locate errors in student reasoning chains

Stepwise reasoning and self-correction are insufficient to improve LLM Teacher Agents' guidance and feedback actionability

Solution: Pedagogical Alignment via TEACHER-STUDENT AGENTIC FRAMEWORKS

Stage 1: Domain-Specific Error-Focussed Diverse Student Modelling

Generate Structured Error Trajectories (e.g., language learning or concept-based instruction) for interpretable signals for pedagogically-informed post-training.

Heterogeneous proxies can allow teacher strategies to generalise across learner populations.

Case Study: bilingual small language models as **cognitive proxies** for second language acquisition (SLA).

Variables include L1 and Age of Acquisition.



linguistic proxies (e.g., BabyLMs) trained on cognitively-plausible corpora

AoA ↑

- adult SLA
- child SLA
- late-state bilingualism
- simultaneous bilingualism



Post-train LLMs on structured proxy error distributions → the aim is to improve how Teacher models internalize realistic, interpretable learner mistakes.

Cognitive Motivation: **Bayesian models of Pedagogy** (for example) adaptively use structured heuristics derived from student behaviour.

- Minimally, we need to simulate how (1) teachers select examples by an initial **modeling of learner belief**, (2) learners (attempt) to infer the intended concept by **reasoning about teacher intentions** and (3) teacher chooses **maximally informative interventions (feedback)**.

Stage 3: Post-trained Teacher LLM Deployed + Evaluated on Pedagogical Multi-Turn Interactions

Multi-Turn Interaction

 $\mathcal{T}_1$ $\mathcal{S}_1$ $\mathcal{T}_2$ $\mathcal{S}_2$

Existing pedagogical evaluation frameworks include LearnLM (Gemini with pedagogical SFT).

- Does the teacher model adapt to student needs, proactively guide conversations, and does not withhold information unproductively?
- Does the model "deepen metacognition" of students (guiding mistake discovery, communicate session plans)?