REAL SAFETY AI FOUNDATION · WORKING PAPER

ARTIFICIAL INTELLIGENCE · SPECIAL EDUCATION · ADAPTIVE LEARNING · UNIVERSAL DESIGN FOR LEARNING · ZONE OF PROXIMAL DEVELOPMENT · PRODUCTIVE FAILURE · ACCOMMODATION · HUMAN-IN-THE-LOOP

A Multi-Theoretical Pedagogical Architecture for Adaptive AI Tutoring

Mapping the features of a special-education tutoring platform to the learning sciences, educational psychology, and disability studies that should govern its design.

Travis Gilly · Real Safety AI Foundation, Cary, Illinois

Design-stage working paper · 2025

2σ

AI tutoring now matches in-person instruction.

Bloom's 1984 two-sigma problem set the target: one-to-one tutoring lifted students roughly two standard deviations above the conventional classroom. Four decades of educational technology chased that gap. A 2025 randomized trial at Harvard found that an AI tutor built on explicit pedagogical principles produced more learning, in less time, than a well-run active-learning class.

SOURCES Bloom, B. S. (1984). The 2 sigma problem. *Educational Researcher*, 13(6), 4–16. · Kestin, G., et al. (2025). AI tutoring outperforms in-class active learning. *Scientific Reports*, 15, 17458.

The students who would benefit most were not its design population.

Students who receive legally mandated accommodations under the Individuals with Disabilities Education Act are precisely where AI tutoring should deliver its largest marginal benefit. Yet most adaptive platforms treat accommodation delivery as an administrative feature rather than a pedagogical commitment. Invisible delivery, machine-verifiable audit trails, population-specific content adaptation, and protections for the teachers who anchor the pedagogy are typically absent from commercial offerings.

SOURCES U.S. Congress (2004). Individuals with Disabilities Education Improvement Act, 20 U.S.C. §1400 et seq. · Wang, R. E., et al. (2024). Tutor CoPilot: A human-AI approach for scaling real-time expertise. Annenberg EdWorkingPaper 24-1054.

A grounding synthesis, not an efficacy study.

The Teacher in the Loop platform was filed as a provisional patent in October 2025 and sits at design stage, with no deployment data. So this work does not ask whether the platform works. It asks a question that comes first: for each published feature, what literature supports it, what recent evidence validates the underlying mechanism, and which claims remain theoretical until the platform is built and tested.

SOURCES Gilly, T. (2025). A Multi-Theoretical Pedagogical Architecture for Adaptive AI Tutoring in Special Education. Real Safety AI Foundation (design-stage working paper).

The mapping-strength test, applied to every feature.

Each feature is held to three questions: is the theory it draws on canonical, has the mechanism been validated in recent work, and has this particular operationalization been tested? The verdict runs on one scale.

- ■■■■ **Very high** theory and evidence both strong
- ■■■□ **High** canonical theory, recent validation
- ■■□□ **Moderate** theory solid, evidence thin or preprint
- ■□□□ **Weak** a hypothesis the platform is built to test

FIVE CLUSTERS OF FEATURES

1. Cognitive and learning theory
2. Disability studies and accommodation
3. Affective and behavioral sensing
4. Academic integrity, reframed as authorship
5. Institutional design and human oversight

SOURCES Gilly, T. (2025). A Multi-Theoretical Pedagogical Architecture for Adaptive AI Tutoring. Real Safety AI Foundation, design-stage working paper.

PART ONE

Cognitive and Learning Theory

How the engine decides what to ask, when to help, and when to wait.

I

Three features rest on the oldest and best-tested traditions in the deck: Socratic calibration, the zone of proximal development, and productive failure.

The Socratic engine that never gives the answer.

All instruction proceeds through strategic questions, analogies, counterexamples, and scaffolded hints, calibrated continuously to the student's current understanding. The novelty is not Socratic tutoring, which is prior art across the field, but the special-education-first constraints placed on the engine: it routes on non-invasive behavioral signals rather than high-stakes items.

MAPPING STRENGTH

High

Canonical theory, from Bloom through the human-tutoring analyses of Chi and the meta-analyses of VanLehn, validated by a rigorous 2025 randomized trial whose authors credited exactly this calibration for the learning gains.

SOURCES Bloom (1984), *Educational Researcher* 13(6). · Chi et al. (2001), *Cognitive Science* 25(4). · VanLehn (2011), *Educational Psychologist* 46(4). · Kestin et al. (2025), *Scientific Reports* 15, 17458.

Cognitive load, kept where it belongs.

The system holds the learner inside the zone of proximal development by reducing extraneous load, the difficulty added by format and presentation, while preserving the intrinsic and germane load that is the actual work of learning. Systems that strip out intrinsic load make the experience easier without making the learning better. A five-month study of personalized difficulty sequencing reported gains equivalent to six to nine months of additional schooling.

MAPPING STRENGTH

High theory, moderate evidence

The foundations are canonical: Vygotsky's zone, the Wood-Bruner-Ross scaffolding metaphor, and Sweller's cognitive load theory. The AI-specific validation is recent, and in part still a preprint.

SOURCES Vygotsky (1978), *Mind in Society*. · Wood, Bruner & Ross (1976), *J. Child Psychology & Psychiatry* 17(2). · Sweller (1988), *Cognitive Science* 12(2). · Chung et al. (2026), SSRN preprint (not yet peer-reviewed).

Productive struggle versus destructive frustration.

Productive struggle, where the student is engaged and generating ideas, is supported with minimal intervention. Destructive frustration, where the student is disengaging or shutting down, triggers a three-tier cascade designed to detect the transition in real time and escalate in graduated rather than binary form.

1. **AI hint reframing**
   The first responder. The engine recasts the hint and lowers extraneous load before any human is involved.

2. **Peer matching** PLANNED
   A classmate who recently mastered the material, paired in for a second route through.

3. **Teacher handoff**
   Direct notification with full context: what was attempted, what triggered the escalation, what has already been tried.

SOURCES **Mapping: strong theory, novel and untested operationalization.** Kapur (2008, 2016), productive failure. · Koedinger & Aleven (2007), the assistance dilemma, *Educational Psychology Review* 19(3).

PART TWO

Disability Studies and Accommodation

The platform's strongest contribution, and its most honest weakness.

II

Students decline their accommodations because of stigma, not ignorance. One feature answers that finding directly; another tests a hypothesis the evidence cannot yet support.

Invisible accommodation delivery.

When a student logs in, the accommodations specified in their education plan activate as system-level defaults, with no visible signal to peers. This answers one of the most robust findings in disability studies: students decline their accommodations because of social cost, not ignorance or access. The platform extends Universal Design for Learning from the curriculum to the system level. The student keeps the right to know what is active; they lose only the visible mark.

MAPPING STRENGTH

Very high

Accommodation non-use driven by stigma is among the best-documented findings in the field, and Universal Design for Learning is a mature framework. The extension to invisible, system-level delivery is the platform's strongest novel contribution.

SOURCES Aquino & Bittinger (2019), *J. Postsecondary Education & Disability* 32(1). · Toutain (2019), same journal, 32(3). · Meyer, Rose & Gordon (2014), *Universal Design for Learning*. · CAST (2018), UDL Guidelines v2.2.

Differential content adaptation: the honest weak point.

The platform switches strategies by profile: consistent structure for autism-spectrum presentations, varied modalities for attention-deficit presentations. The underlying neuroscience supports distinct profiles. What it does not support is a single platform that switches between them, and the measurement tools that would drive the switch degrade in exactly these populations. The paper presents the feature as a hypothesis it is built to test, not a validated choice.

MAPPING STRENGTH

Weak

The weakest mapping in the paper, and stated as such. A reviewer would be right to push back on any stronger claim. Honesty here is the point: the gap is named rather than papered over.

SOURCES Barkley (1997), *Psychological Bulletin* 121(1). · Baron-Cohen (2009), *Annals NYAS* 1156. · Le Cunff, Giampietro & Dommett (2024), *European J. Neuroscience* 59(2).

PART THREE

Sensing, Authorship, and Oversight

Reading the learner, proving the work, and keeping a human in charge.

III

How the system perceives state, how it lets a student prove the work is theirs, and why the human-in-the-loop must be built into architecture rather than promised in documentation.

Reading the learner from passive signals.

Cognitive and affective state is inferred from typing rhythm, mouse movement, and pause patterns, requiring no sensor beyond a standard laptop or Chromebook, with the camera off in the default configuration. Keystroke dynamics for cognitive state has a mature record, with reported accuracies in the seventy to ninety percent range. Camera-based facial analysis is more contested: classifiers degrade on children, non-white, and neurodivergent faces, so it is available only as an explicit opt-in.

MAPPING STRENGTH

Strong layer, weaker layer

The passive-signal layer is well-founded and privacy-preserving. The camera layer is weaker on both accuracy and privacy, and is treated as optional rather than core.

SOURCES Picard (1997), *Affective Computing*. · Epp, Lippold & Mandryk (2011), CHI '11. · Baker et al. (2010), *Int. J. Human-Computer Studies* 68(4). · Le Cunff et al. (2024), *European J. Neuroscience* 59(2).

Authorship as proof, not surveillance.

Every AI suggestion is marked and tracked, and the platform captures the student's drafts, edits, and revision sequence. The framing is deliberately inverted from typical integrity tools. The record does not exist to catch the student. It exists to give the student evidence that the work is genuinely theirs when it is questioned.

> *This is the student's own proof of authorship, rather than the institution's proof of non-cheating.*

SOURCES **Mapping: high theory, emerging operationalization.** Flower & Hayes (1981), writing as a recursive process, *College Composition and Communication* 32(4).

Human-in-the-loop as architecture, not marketing.

The teacher keeps pedagogical authority at every level; the AI operates as a suggestion layer, never a decision layer. Escalation hands the teacher full context, and a contractual anti-displacement clause forbids cutting teaching or aide staff because of efficiency gains. Stated only in documentation, human-in-the-loop deteriorates into marketing language. Built into the architecture and the contract, it survives economic pressure.

MAPPING STRENGTH

Very high

Teacher-in-the-loop is the named approach across the U.S. Department of Education report and recent studies of AI in teacher education and inclusive classrooms.

SOURCES U.S. Dept. of Education, OET (2023). · Fajardo-Ramos et al. (2025), *Frontiers in Education* 10. · Marras & Pasqualotto (2026), *Frontiers in Child & Adolescent Psychiatry* 5. · Wang et al. (2024).

Promised versus built.

Five features appear in the documentation but are not yet implemented. Each is mapped to the tradition that would ground it when built, so the reader can tell the current platform from the planned one.

A **Peer-tutoring matching**
The protégé effect; Fiorella & Mayer (2013).

B **Bayesian knowledge tracing**
Forgetting curves; Corbett & Anderson (1995).

C **Explainable AI reasoning**
Stakeholder-tailored rationale; Khosravi et al. (2022).

D **Collaboration detection**
Scope must be clarified before it is grounded.

E **Portable cognitive profile** MOST FRAUGHT
A model that follows a student across years; among the most identifying data a minor could carry. Would require consent at each age of majority and deletion on demand.

SOURCES Fiorella & Mayer (2013), *Contemporary Educational Psychology* 38(4). · Corbett & Anderson (1995), *User Modeling and User-Adapted Interaction* 4(4). · Khosravi et al. (2022), *Computers and Education: AI* 3.

The preconditions that make the pedagogy possible.

Privacy by design

Personally identifying data stays on school-controlled infrastructure; the cloud engine sees only tokenized interactions.

A locally deployable model

A fully offline option on a district-controlled server, for jurisdictions that require it.

Compliance as byproduct

The platform markets pedagogy; the audit trail it generates happens to satisfy what districts must report.

Default opt-out for sensing

The data-hungry features require an affirmative yes, placing the burden on whoever benefits from the data, not the student.

SOURCES Cavoukian (2009), Privacy by Design. · Kumar, Chetty, Clegg & Vitak (2019), CHI '19. · U.S. Congress (2004), IDEA, alternative progress reporting provisions.

Limitations, stated plainly.

The platform is at design stage with no deployment data, so every efficacy claim is indirect. The transfer of evidence from general-education tutoring to special education is a hypothesis, not an established extrapolation. The differential-adaptation feature remains the weakest mapping and should not be read as validated.

THE AUTHOR'S STANDPOINT

The author is the platform's inventor and the Foundation's director. Favorable reception connects, through the funding pipeline, to his potential future compensation. The reader is entitled to weigh the arguments with that pathway in view, rather than under the impression of a permanently volunteer investigator.

SOURCES Gilly, T. (2025), Limitations and Conflict of Interest statements, design-stage working paper.

What the mapping shows.

Adaptive AI tutoring for students with documented learning needs is theoretically tractable. The grounding is defensible across most of the feature set, theoretically supported but empirically pending in specific named cases, and honestly acknowledged where it is weakest. What the field lacks is not the literatures. It is a platform that has implemented those insights as architectural commitments rather than feature marketing. Whether this attempt succeeds is a question that only deployment will settle.

SOURCES Gilly, T. (2025). A Multi-Theoretical Pedagogical Architecture for Adaptive AI Tutoring in Special Education. Real Safety AI Foundation (design-stage working paper).

Gilly, T. (2025). *A Multi-Theoretical Pedagogical Architecture for Adaptive AI Tutoring in Special Education: Mapping Platform Features to Established Research in Learning Sciences, Educational Psychology, and Disability Studies.* Real Safety AI Foundation. Design-stage working paper.

LEARNING SCIENCES & PSYCHOLOGY

Bloom, B. S. (1984). The 2 sigma problem. *Educational Researcher*, 13(6), 4–16.

Chi, M. T. H., et al. (2001). Learning from human tutoring. *Cognitive Science*, 25(4), 471–533.

VanLehn, K. (2011). The relative effectiveness of human tutoring, ITS, and other tutoring systems. *Educational Psychologist*, 46(4).

Vygotsky, L. S. (1978). *Mind in Society.* Harvard University Press.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *J. Child Psychology & Psychiatry*, 17(2).

Sweller, J. (1988). Cognitive load during problem solving. *Cognitive Science*, 12(2).

Kapur, M. (2008; 2016). Productive failure. *Cognition and Instruction*; *Educational Psychologist*.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma. *Educational Psychology Review*, 19(3).

Kestin, G., et al. (2025). AI tutoring outperforms in-class active learning. *Scientific Reports*, 15, 17458.

Chung, A. T.-H., et al. (2026). Effective personalized AI tutors via LLM-guided RL. SSRN preprint.

Fiorella, L., & Mayer, R. E. (2013). Learning by teaching. *Contemporary Educational Psychology*, 38(4).

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing. *User Modeling & User-Adapted Interaction*, 4(4).

Ebbinghaus, H. (1885). *Memory.* Duncker & Humblot.

DISABILITY STUDIES & SENSING

Aquino, K. C., & Bittinger, J. D. (2019). The self-(un)identification of disability. *JPED*, 32(1).

Toutain, C. (2019). Barriers to accommodations in higher education. *JPED*, 32(3).

Clouder, L., et al. (2020). Neurodiversity in higher education. *Higher Education*, 80(4).

Meyer, A., Rose, D. H., & Gordon, D. (2014). *Universal Design for Learning.* CAST.

CAST (2018). UDL Guidelines, version 2.2.

Barkley, R. A. (1997). Behavioral inhibition and executive functions. *Psychological Bulletin*, 121(1).

Baron-Cohen, S. (2009). The empathizing-systemizing theory. *Annals NYAS*, 1156.

Le Cunff, A.-L., Giampietro, V., & Dommett, E. (2024). Cognitive load in ADHD, ASD, dyslexia. *Eur. J. Neuroscience*, 59(2).

Picard, R. W. (1997). *Affective Computing.* MIT Press.

Baker, R. S. J. d., et al. (2010). Better to be frustrated than bored. *IJHCS*, 68(4).

Epp, C., Lippold, M., & Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. CHI '11.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *CCC*, 32(4).

POLICY, PRIVACY & LAW

U.S. Congress (2004). Individuals with Disabilities Education Improvement Act, 20 U.S.C. §1400 et seq.

U.S. Department of Education, Office of Educational Technology (2023). *Artificial Intelligence and the Future of Teaching and Learning.*

Fajardo-Ramos, D. C., Chiappe, A., & Mella-Norambuena, J. (2025). Human-in-the-loop assessment with AI. *Frontiers in Education*, 10.

Marras, A., & Pasqualotto, A. (2026). Negotiating assistive technologies and AI in inclusive education. *Frontiers in Child & Adolescent Psychiatry*, 5.

Cavoukian, A. (2009). *Privacy by Design: The 7 Foundational Principles.*

Kumar, P. C., Chetty, M., Clegg, T. L., & Vitak, J. (2019). Privacy and security in elementary schools. CHI '19.

Wang, R. E., et al. (2024). Tutor CoPilot. Annenberg EdWorkingPaper 24-1054.

Dutta, S., Roy, S., & Roy, U. (2026). Quantum kernel anomaly detection for behavioral biometrics. *J. Supercomputing*, 82(3).

Khosravi, H., et al. (2022). Explainable AI in education. *Computers and Education: AI*, 3.