TinkAI

Cognitive Depth Report

Measuring cognitive engagement in AI-mediated reasoning
Pilot data from 32 users across 99 sessions

32	99	1,143	156
Unique Users	Sessions	Messages	Challenges

March 12, 2026 | Version 1.0 | www.tinkai.ai
Alessandro Leonetti Luparini | Founder, TinkAI

1. Executive Summary

TinkAI is a cognitive orchestration layer that operates between users and a large language model. Unlike standard AI assistants, TinkAI dynamically modulates the conversation across four modes (Exploration, Challenge, Fast, Learning) with the goal of preserving the user's cognitive autonomy during AI-assisted reasoning.

This report presents pilot data collected from January 19 to March 12, 2026, across 32 authenticated users, 99 sessions, and 1,143 messages. The data was collected automatically and anonymously: TinkAI tracks mode sequences and engagement patterns, never conversation content.

The central question is whether a cognitive orchestration system can detect and counteract what Shaw and Nave (2026) term "cognitive surrender" the tendency to adopt AI outputs with minimal scrutiny, overriding both intuition (System 1) and deliberation (System 2). Our pilot data provides preliminary observations, not conclusions. The sample is small, the usage is uncontrolled, and no experimental comparison group exists. What the data does offer is a set of measurable patterns that warrant structured investigation.

2. Theoretical Context

Shaw and Nave's Tri-System Theory (2026) extends dual-process models of cognition by introducing System 3: artificial cognition that operates outside the brain but whose outputs are incorporated into human judgment. Their experiments (N=1,372; 9,593 trials) demonstrate that participants adopt AI outputs with minimal scrutiny, leading to improved accuracy when the AI is correct but degraded accuracy when the AI errs a behavioral signature they call cognitive surrender.

TinkAI's architecture is designed to address this dynamic. Rather than providing answers by default, the system monitors the user's reasoning process and intervenes with targeted challenges when it detects uncritical acceptance of AI-generated conclusions. The four modes correspond to different cognitive postures: Exploration opens inquiry, Challenge introduces friction, Fast delivers direct responses when appropriate, and Learning guides structured understanding.

Vendrell and Johnston (2026) identify six intellectual processes at risk in unstructured AI use: conceptual interpretation, inferential reasoning, evaluative judgment, metacognitive regulation, intellectual curiosity, and epistemic integrity. Their eight design principles for integrating LLMs in higher education — particularly preserving cognitive friction (P1), scaffolding LLMs as thinking partners (P2), and balancing AI-mediated with AI-free phases (P8) —align directly with TinkAI's orchestration logic.

3. Key Metrics

Metric	Value	Description
Avg. Pushback Ratio	17.3%	Proportion of AI challenges followed by active user response in the next message
Avg. Challenge Rate	12.2%	Percentage of AI messages classified as CHALLENGE mode
Socratic Effectiveness	30.6%	11 of 36 sessions with challenges produced at least one active pushback
Total Sessions	99	36 contained at least one challenge
Total Messages	1,143	AI-User exchanges recorded
Unique Users	32	Authenticated users (guests excluded)
Avg. Session Duration	115.6 min	Where duration was recorded
Total Challenges	156	95 raw pushback events recorded

3.1 Mode Distribution

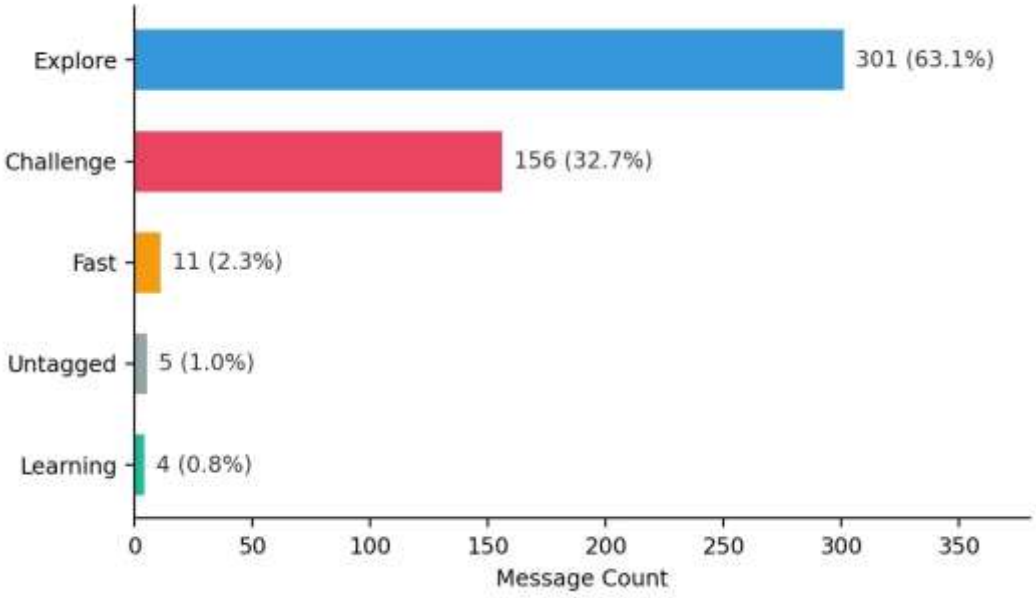

Figure 1. Distribution of cognitive modes across 477 tagged AI messages.

The system operates predominantly in two modes: Explore (63.1%) and Challenge (32.7%). Fast and Learning modes account for a combined 3.1% of messages. This indicates that the current orchestration logic activates meaningful cognitive engagement (Explore + Challenge) in nearly all interactions, but the Learning and Fast modes are underutilized. Whether this reflects the nature of the pilot tasks, user behavior, or a calibration gap in the system is an open question that requires further investigation.

Note on metrics: The Challenge Rate (12.2%) measures the proportion of challenges relative to all messages in the session, including user messages. The Mode Distribution (32.7%) measures challenges as a proportion of AI-generated messages only. Both figures are correct; they describe different denominators.

4. Session Engagement Analysis

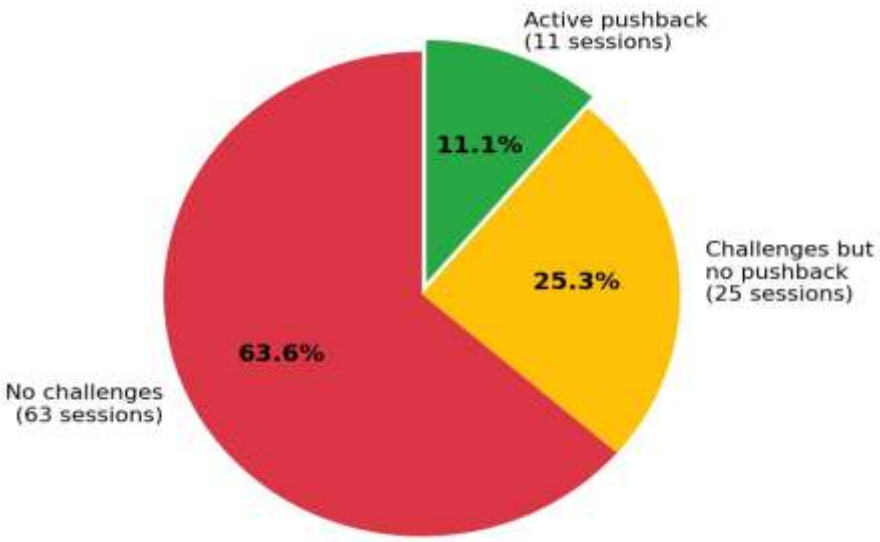

Figure 2. Session classification by cognitive engagement level.

Of the 99 total sessions, 63 (63.6%) contained no challenges at all. In these sessions, the system operated entirely in Explore mode, meaning the AI provided information and the user did not encounter friction. Of the remaining 36 sessions where challenges occurred, 25 produced no measurable pushback the user received a challenge but did not engage with it in the immediately following message. Only 11 sessions (11.1% of total) showed active cognitive engagement: the user received a challenge and responded with a substantive counter-argument, question, or revision.

This is the most important finding in the pilot. It suggests that the mere presence of cognitive friction does not automatically produce critical engagement. The majority of users, when challenged, either ignored the challenge or accepted it passively. This pattern is consistent with Shaw and Nave's cognitive surrender hypothesis: users default to accepting AI outputs even when those outputs are designed to provoke scrutiny. The 11.1% who actively pushed back may represent users with pre-existing high cognitive autonomy, not users whose autonomy was created by the system.

4.1 Common Mode Sequences

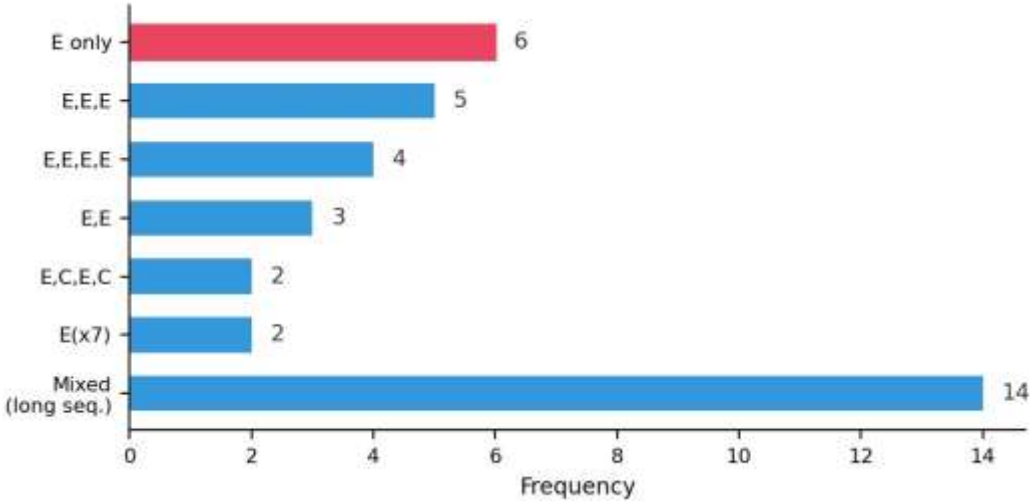

Figure 3. Most frequent mode sequences across sessions.

The dominant patterns are pure Explore sequences (E, E-E-E, E-E-E-E), confirming that most sessions unfold without cognitive friction. The E-C-E-C pattern alternating exploration and challenge appears only twice, suggesting that sustained dialogic engagement between user and system is rare in unstructured use. Mixed long sequences (14 sessions) show more complex interactions but remain a minority.

5. User Cognitive Profiles

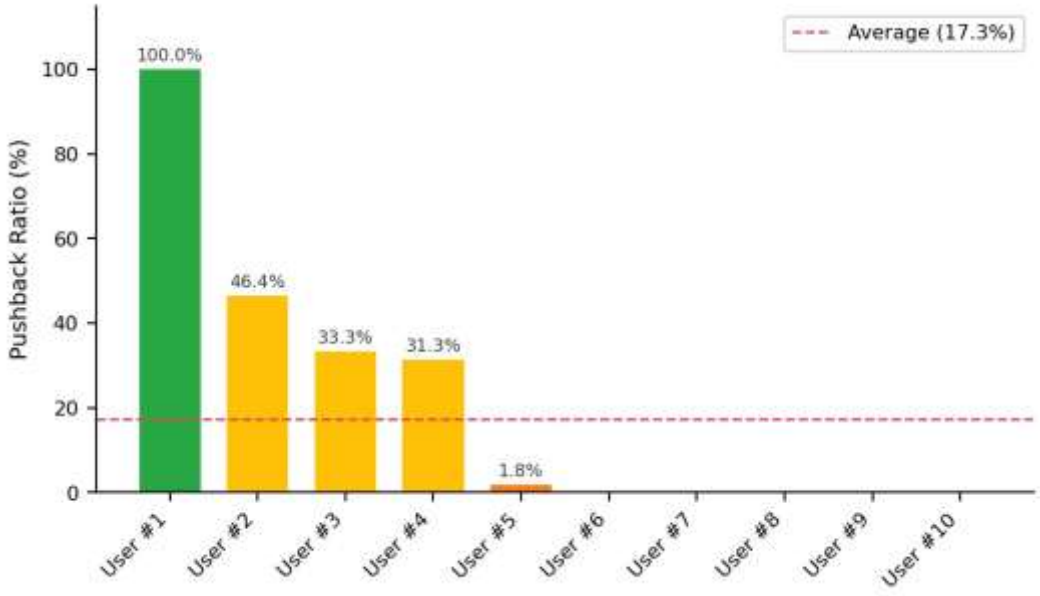

Figure 4. Pushback Ratio by user (minimum 3 challenges received). Red dashed line indicates average.

The distribution is sharply bimodal. Four users (Users #1–#4) show pushback ratios between 31% and 100%. Six users (Users #5–#10) show pushback ratios between 0% and 1.8%. There is virtually no middle ground. This pattern maps directly onto Shaw and Nave's distinction between "Independent" cognitive profiles (users who maintain deliberative oversight of AI outputs) and "AI-User" profiles (users who systematically defer to System 3).

User	Sessions	Messages	Challenges	Pushback Ratio	Profile
#1	2	36	10	100.0%	Independent
#2	4	78	15	46.4%	Engaged
#3	7	114	35	33.3%	Engaged
#4	8	106	26	31.3%	Engaged
#5	5	58	40	1.8%	Passive
#6	1	26	3	0.0%	Passive
#7	1	22	5	0.0%	Passive
#8	1	20	3	0.0%	Passive
#9	1	28	4	0.0%	Passive
#10	1	18	6	0.0%	Passive

Table 1. Top 10 users by cognitive engagement (minimum 3 challenges received).

User #5 is particularly noteworthy. This user received 40 challenges the highest in the sample across 5 sessions and 58 messages, yet showed only 1.8% pushback. This suggests intensive but entirely passive engagement with TinkAI: the user interacted frequently but almost never contested a challenge. In Shaw and Nave's framework, this represents cognitive surrender occurring within a system explicitly designed to prevent it. Whether this reflects the user's cognitive disposition, the quality of the challenges, or the absence of task-specific stakes is unknown at this stage.

6. Interpretation

6.1 What the Data Shows

Three findings emerge from this pilot:

First, cognitive friction alone does not produce critical engagement. The majority of users who received challenges did not engage with them. This does not mean the challenges are ineffective it means that introducing friction is a necessary but insufficient condition for preserving cognitive autonomy. Vendrell and Johnston (2026) reach a similar conclusion: their design principle P1 (preserve cognitive friction) must be paired with P4 (activate metacognitive self-regulation) and P7 (align assessment with intended cognition) to produce sustained engagement.

Second, cognitive profiles are sharply bimodal. Users cluster into "high autonomy" and "high passivity" groups with almost no intermediate cases. This is consistent with Shaw and Nave's experimental findings and suggests that individual differences in cognitive disposition strongly moderate the effectiveness of any intervention. Future work must determine whether TinkAI can shift users across this divide not just detect where they already stand.

Third, TinkAI generates measurable behavioral traces that map onto established cognitive constructs. The pushback ratio, mode sequences, and session engagement patterns provide operationalizable proxies for concepts like cognitive surrender, epistemic agency, and metacognitive regulation. This suggests that a cognitive orchestration layer can function as a measurement instrument, regardless of whether it functions as an effective intervention.

6.2 What the Data Does Not Show

This pilot cannot establish causation. We do not know whether users who pushed back did so because of TinkAI's intervention or because they would have reasoned critically regardless. There is no control group, no pre/post measurement, and no randomization. The sample is small (32 users), self-selected (all volunteered to test TinkAI), and uncontrolled (users chose their own topics and interaction frequency). Session duration varies from 30 seconds to over 2 hours, and we cannot distinguish between users who abandoned a session and those who completed their task quickly.

Furthermore, the pushback detection system itself is unvalidated. We classify a response as "pushback" based on sequential proximity to a challenge, but have not yet conducted inter-rater reliability testing to confirm that what the system labels as pushback corresponds to genuine critical engagement. A user who changes the topic after a challenge is currently classified the same as a user who ignores it. This measurement limitation must be addressed before any stronger claims can be made.

6.3 The Central Question

The pilot raises a question that the pilot itself cannot answer: **Does cognitive orchestration reduce cognitive surrender, or does it merely reveal pre-existing cognitive dispositions?**

Answering this requires a controlled study with pre/post measurement, randomized assignment to TinkAI vs. standard chatbot conditions, validated instruments for critical thinking and metacognitive regulation, and a sample large enough to detect meaningful effects. The pilot data suggests such a study is feasible and worth conducting.

7. Next Steps

Based on these pilot findings, TinkAI's development roadmap includes three priorities:

Validate the measurement instrument. Inter-rater reliability testing of mode classification and pushback detection is the immediate priority. If the system's classifications do not correspond to expert judgment, all downstream findings are compromised.

Design a controlled study. A randomized controlled trial comparing TinkAI users to standard chatbot users on independent reasoning tasks, with pre/post measurement using validated instruments (e.g., Cognitive Reflection Test, as adapted by Shaw and Nave), would provide the evidence needed to distinguish between detection and intervention effects.

Investigate the passivity gap. The most practically significant question is whether adaptive challenge calibration — adjusting the intensity, timing, and framing of challenges based on individual user patterns — can shift passive users toward active engagement over time. The current pilot provides no evidence on this, but the bimodal distribution makes it the most important question to answer.

References

Shaw, S. D., & Nave, G. (2026). Thinking — Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender. *Preprint, SSRN.* https://doi.org/10.31234/osf.io/yk25n_v1

Vendrell, M., & Johnston, S.-K. (2026). Scaffolding Critical Thinking with Generative AI: Design Principles for Integrating Large Language Models in Higher Education. *Computers and Education: Artificial Intelligence.* https://doi.org/10.1016/j.caeai.2026.100572

Potkalitsky, N. (2025–2026). *Thinking with AI.* Substack. Framework for cognitive autonomy in AI-mediated educational contexts.