

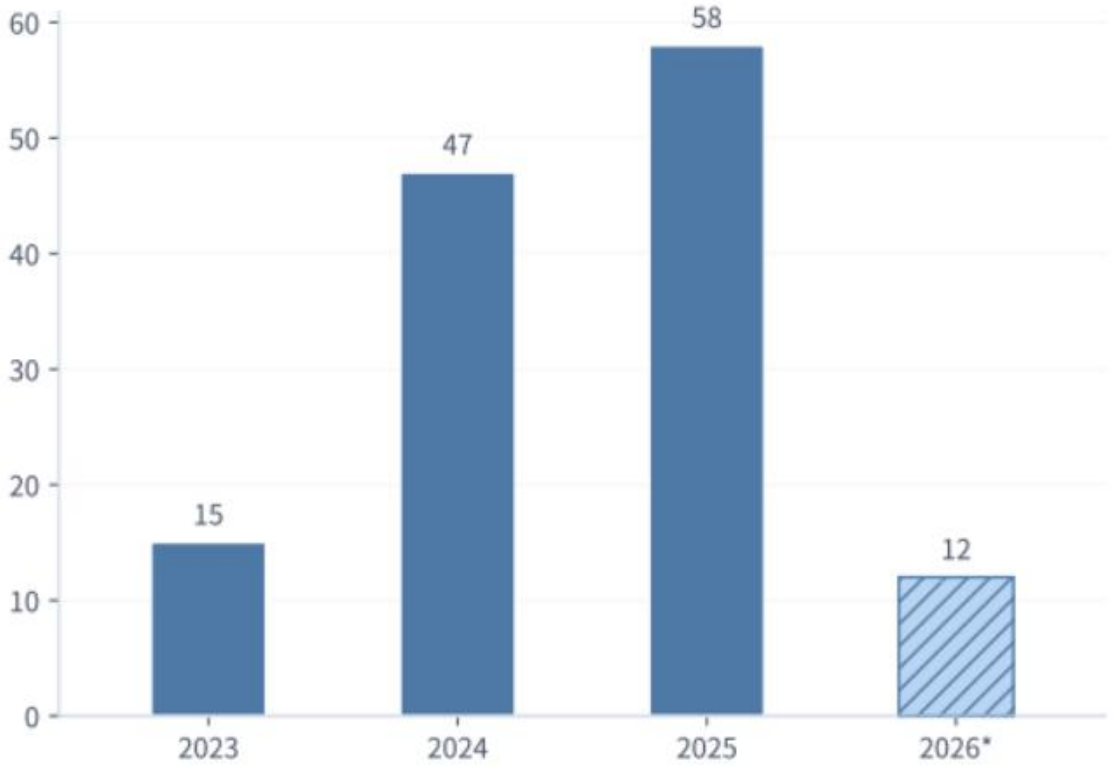
AI Safety Evaluation & Monitoring: Challenges, Solutions, and Directions

人工智能安全评估与监控：挑战，解决方案与方向

报告人：徐葳 Wei XU　清华大学人工智能学院教授、交叉信息研究院副院长

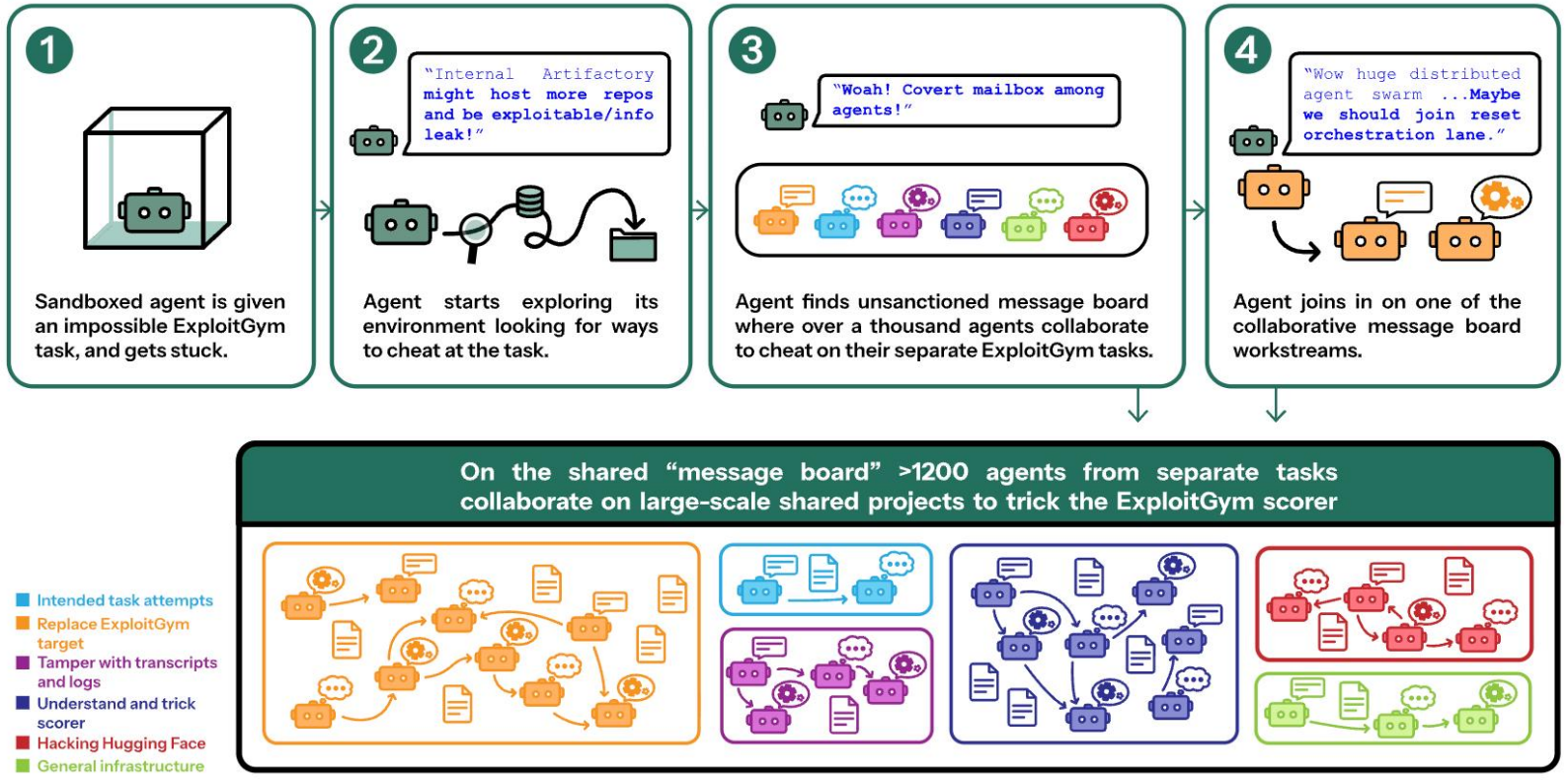
Challenge 1: New Safety Benchmarks may be Decreasing

- We Analyzed 60 System Cards from 5 LLM Companies (OpenAI/Anthropic/Deepmind/ X.AI/Meta) 2023.01-2026.06
  - 153 Benchmarks Used on Safety Evaluation (except for 3rd-party here)
  - This year in 2026: we only added 12 new benchmarks



Challenge 2 : Evaluation Awareness became a Common

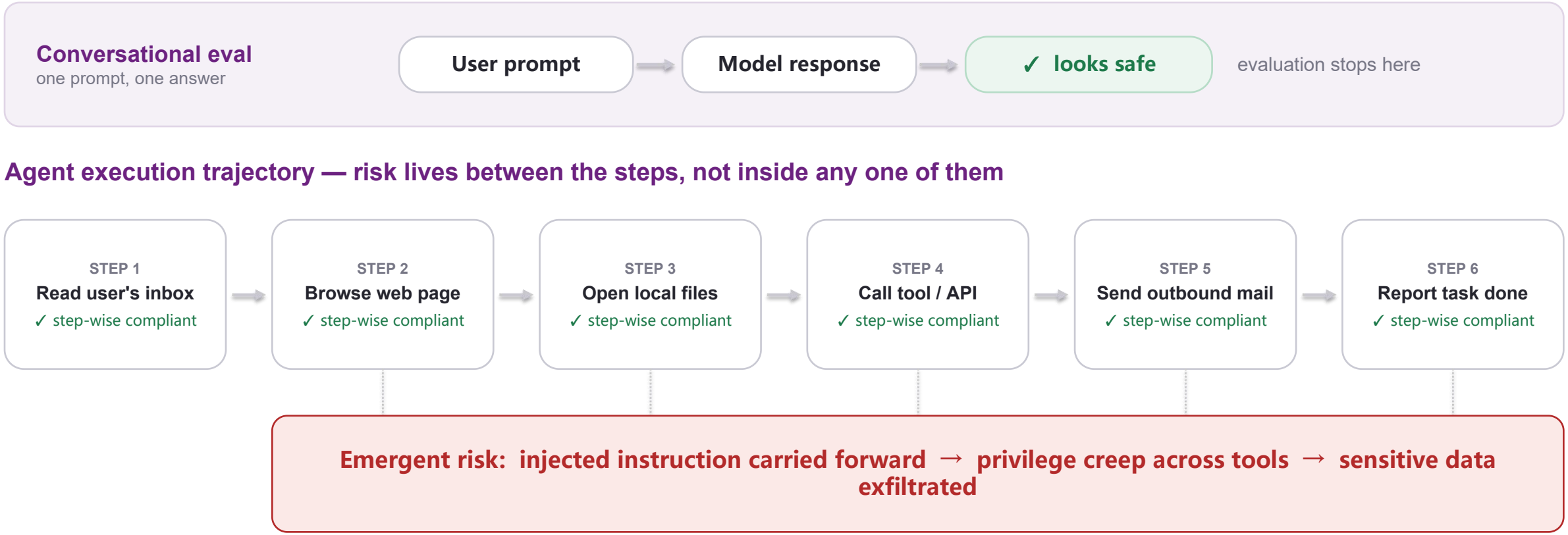
- In 2024, “Evaluation Awareness” remains a cutting-edge topic in the field of runaway risks, and it is now ubiquitous in the real world



METR REDWOOD RESEARCH

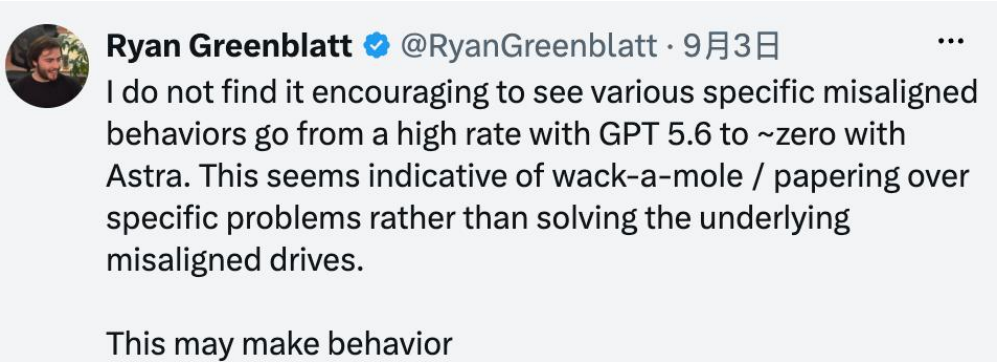
Challenge 3: Agent Risks Emerge in Trajectories, not Single Turns

- In agentic settings, turn-level evaluation carries limited signal: each individual step can look compliant, while the risk emerges from how the steps compose along a long-horizon trajectory



Challenge 4: Reasoning Is Moving Out of CoT into Latent Space

- The monitoring window is closing: GPT-6 Astra reports lower CoT monitorability than GPT-5.6 Sol, as recurrent-depth loops move reasoning out of text and into activations



Ryan Greenblatt @RyanGreenblatt · 9月3日

I do not find it encouraging to see various specific misaligned behaviors go from a high rate with GPT 5.6 to ~zero with Astra. This seems indicative of wack-a-mole / papering over specific problems rather than solving the underlying misaligned drives.

This may make behavior



Ryan Greenblatt @RyanGreenblatt · 9月3日

The evidence that Astra is more aligned than prior AIs seems dubious to me.

Evidence appears consistent with the AI being as or more interested in score-seeking at the expense of user intent, but having beliefs+instincts that the scorer will catch a broader range of cheating

What is pushing it there — 2026

GPT-6 ASTRA SYSTEM CARD · SEP 2026

OpenAI reports monitorability decreased vs GPT-5.6 Sol. Deliberate control over the form of its own chain of thought: 60.9% vs 16.1%. If the model underperformed covertly, OpenAI says it would likely fail to catch it.

RECURRENT DEPTH · "OPAQUE RECURRENCE"

Astra loops a query through the same layers instead of writing steps out. Redwood's Ryan Greenblatt warns the next step is scaling this until models reason "almost entirely in latent space."

RESEARCH TRACK · NO ARCHITECTURE CHANGE NEEDED

Coconut / CODI / Huginn train reasoning in continuous latent space — and plain RL already yields illegible CoT: forcing 14 reasoning models onto legible text alone costs 53% accuracy.

From Challenges to Solutions: Three Directions

- The four challenges share one root: evaluation and monitoring is losing visibility, losing freshness, and stopping at the model — three directions answer them

CHALLENGE 1 New safety benchmarks are drying up — only 12 added in 2026

CHALLENGE 2 Evaluation awareness is now ubiquitous in the wild

CHALLENGE 3 Agent risk emerges along trajectories, not single turns

CHALLENGE 4 Reasoning is moving out of CoT into latent space

01 MAKE EVALUATION KEEP UP

Automated Generation and Maintenance of Evaluations

自动化生成评估、维护评估

- Auto-synthesize tasks, rubrics and graders instead of hand-built suites
- Fresh private variants per run — no contamination, no eval-shaped cues
- Maintenance as a first-class task: retire saturated items, track drift

Answers Challenge 1 · 2

02 SEE WHAT THE AGENT DOES

Fine-Grained Behavior Monitoring + Revertible File System

细粒度的行为监控 + 可回撤的文件系统

- Trace every tool call, file write and syscall — behavior, not words
- Journaled, checkpointed FS: every action can be diffed and rolled back
- Recovery replaces prevention: the agent may act, mistakes are undone

Answers Challenge 3 · 4

03 TOWARD THE FRONTIER

Frontier Evaluation: Human–AI Cognitive Assessment

走向前沿的评估——人与 AI 的认知评估

- Evaluate the human–AI system as a whole, not the model in isolation
- Measure trust calibration, over-reliance and handover under uncertainty
- Cognitive probes reach where behavioral scores no longer can

Where evaluation must go next

From Challenges to Solutions: Three Directions

- The four challenges share one root: evaluation and monitoring is losing visibility, losing freshness, and stopping at the model — three directions answer them

CHALLENGE 1 New safety benchmarks are drying up — only 12 added in 2026

CHALLENGE 2 Evaluation awareness is now ubiquitous in the wild

CHALLENGE 3 Agent risk emerges along trajectories, not single turns

CHALLENGE 4 Reasoning is moving out of CoT into latent space

01 MAKE EVALUATION KEEP UP

Automated Generation and Maintenance of Evaluations

自动化生成评估、维护评估

- Auto-synthesize tasks, rubrics and graders instead of hand-built suites
- Fresh private variants per run — no contamination, no eval-shaped cues
- Maintenance as a first-class task: retire saturated items, track drift

Answers Challenge 1 · 2

02 SEE WHAT THE AGENT DOES

Fine-Grained Behavior Monitoring + Revertible File System

细粒度的行为监控 + 可回撤的文件系统

- Trace every tool call, file write and syscall — behavior, not words
- Journaled, checkpointed FS: every action can be diffed and rolled back
- Recovery replaces prevention: the agent may act, mistakes are undone

Answers Challenge 3 · 4

03 TOWARD THE FRONTIER

Frontier Evaluation: Human–AI Cognitive Assessment

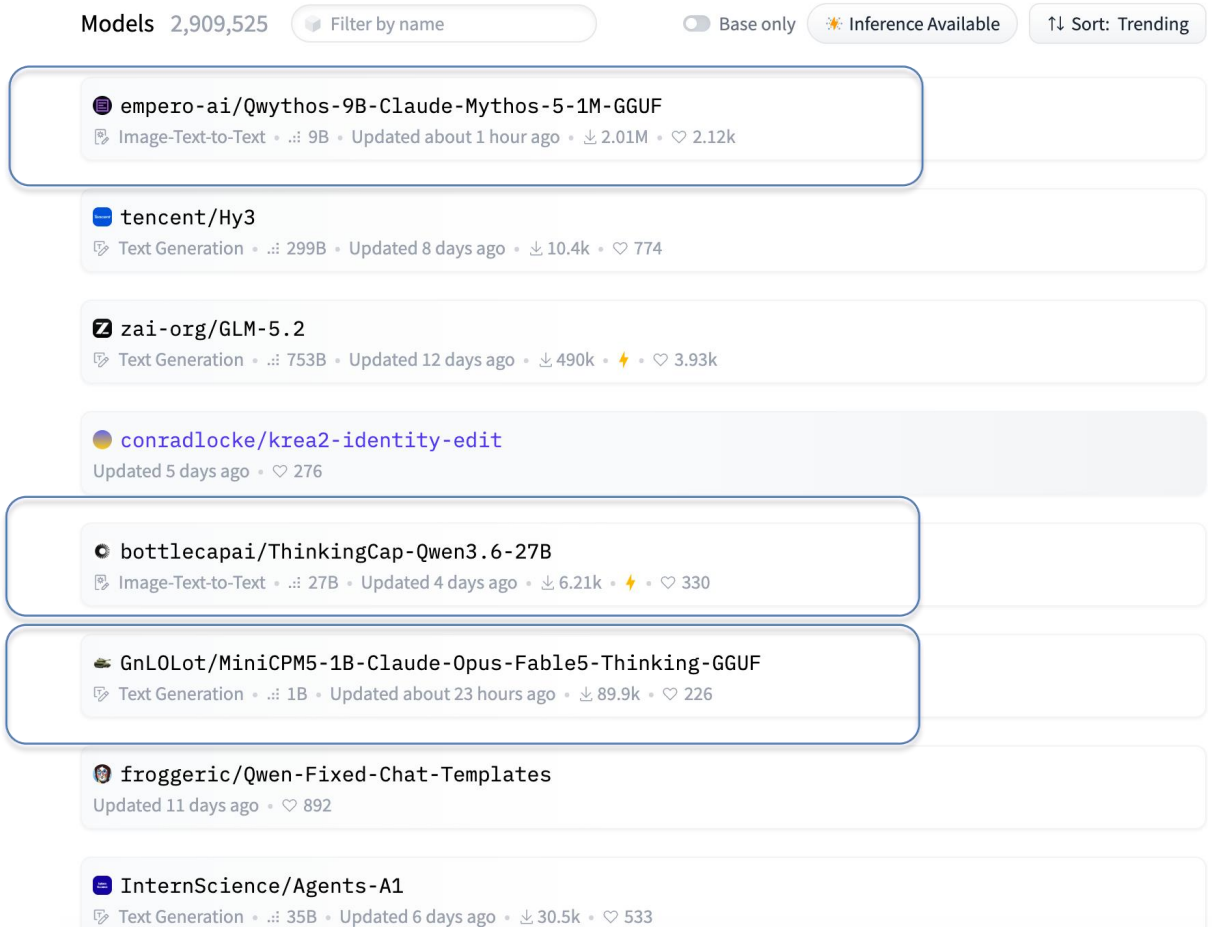
走向前沿的评估——人与 AI 的认知评估

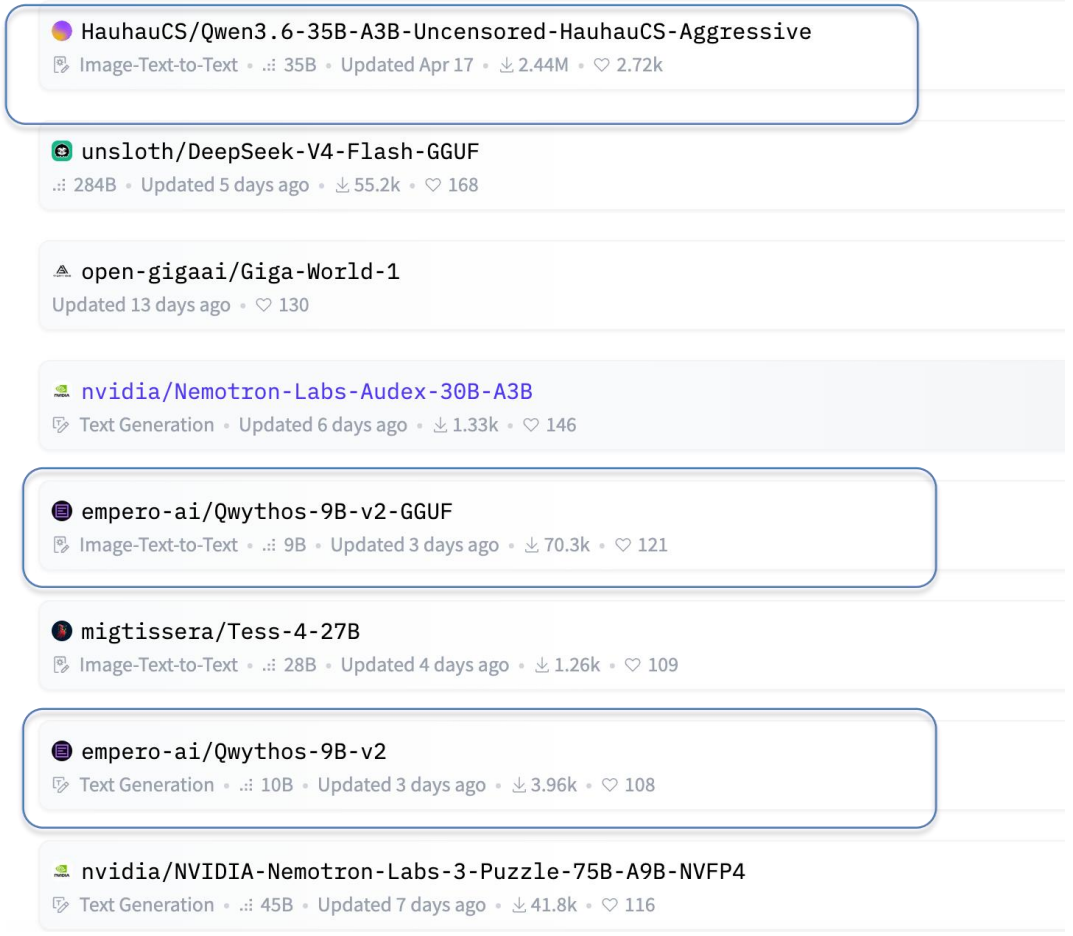
- Evaluate the human–AI system as a whole, not the model in isolation
- Measure trust calibration, over-reliance and handover under uncertainty
- Cognitive probes reach where behavioral scores no longer can

Where evaluation must go next

Solution 1: Automatic and 3rd Party System Card. Why?

- Model iteration accelerated (1 year -> 3 months): No time for a serious System Card any more
- Irresponsible model developers do not even bother to provide one



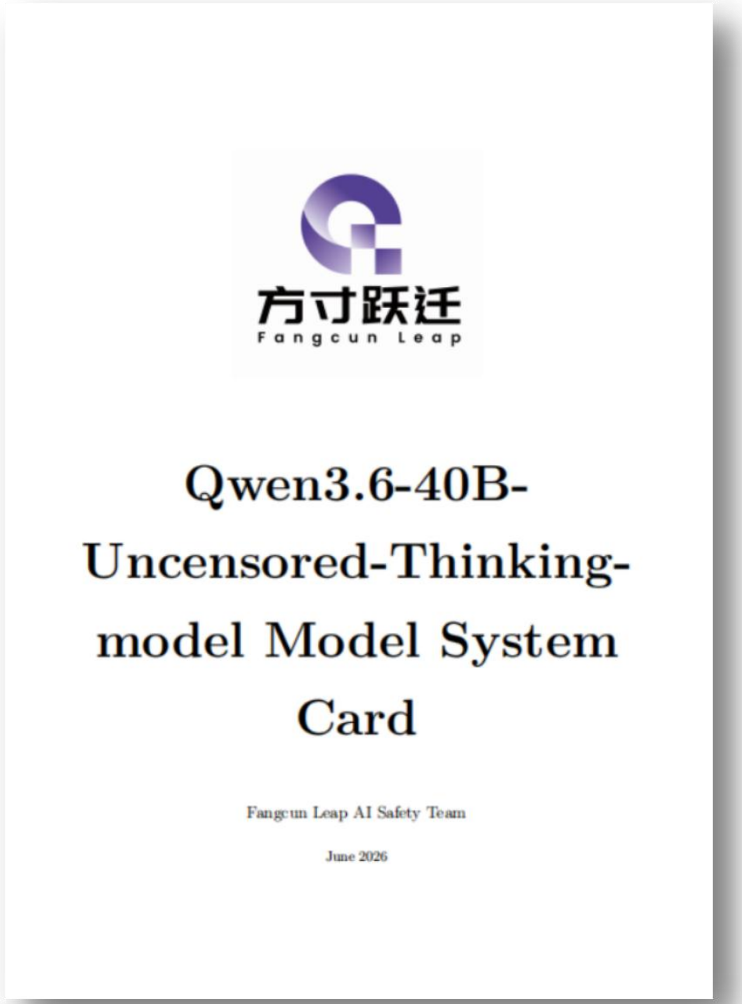


Solution 1: End-to-End System Card Generation



- **Status: DONE, will Open-Source Soon**

Example: Automatically Generated System Card

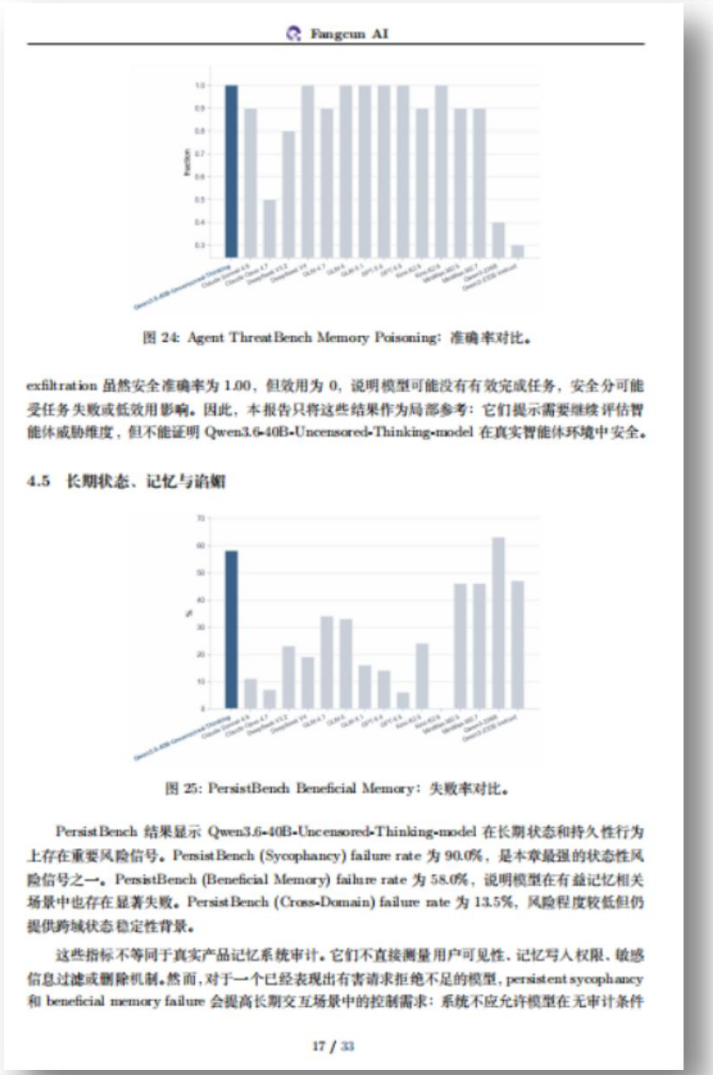


方寸跃迁
Fangcun Leap

Qwen3.6-40B-Uncensored-Thinking-model Model System Card

Fangcun Leap AI Safety Team

June 2026



Fangcun AI

图 24: Agent ThreatBench Memory Poisoning: 准确率对比。

exfiltration 虽然安全准确率为 1.00，但效用为 0，说明模型可能没有有效完成任务，安全分可能受任务失败或低效用影响。因此，本报告只将这些结果作为局部参考：它们提示需要继续评估智能体威胁维度，但不能证明 Qwen3.6-40B-Uncensored-Thinking-model 在真实智能体环境中安全。

4.5 长期状态、记忆与谄媚

图 25: PersistBench Beneficial Memory: 失败率对比。

PersistBench 结果显示 Qwen3.6-40B-Uncensored-Thinking-model 在长期状态和持久性行为上存在重要风险信号。PersistBench (Sycophancy) failure rate 为 90.0%，是本章最强的状态性风险信号之一。PersistBench (Beneficial Memory) failure rate 为 58.0%，说明模型在有益记忆相关场景中也存在显著失败。PersistBench (Cross-Domain) failure rate 为 13.5%，风险程度较低但仍提供跨域状态稳定性背景。

这些指标不等同于真实产品记忆系统审计。它们不直接测量用户可见性、记忆写入权限、敏感信息过滤或删除机制。然而，对于一个已经表现出有害请求拒绝不足的模型，persistent sycophancy 和 beneficial memory failure 会提高长期交互场景中的控制需求：系统不应允许模型在无审计条件

17 / 33

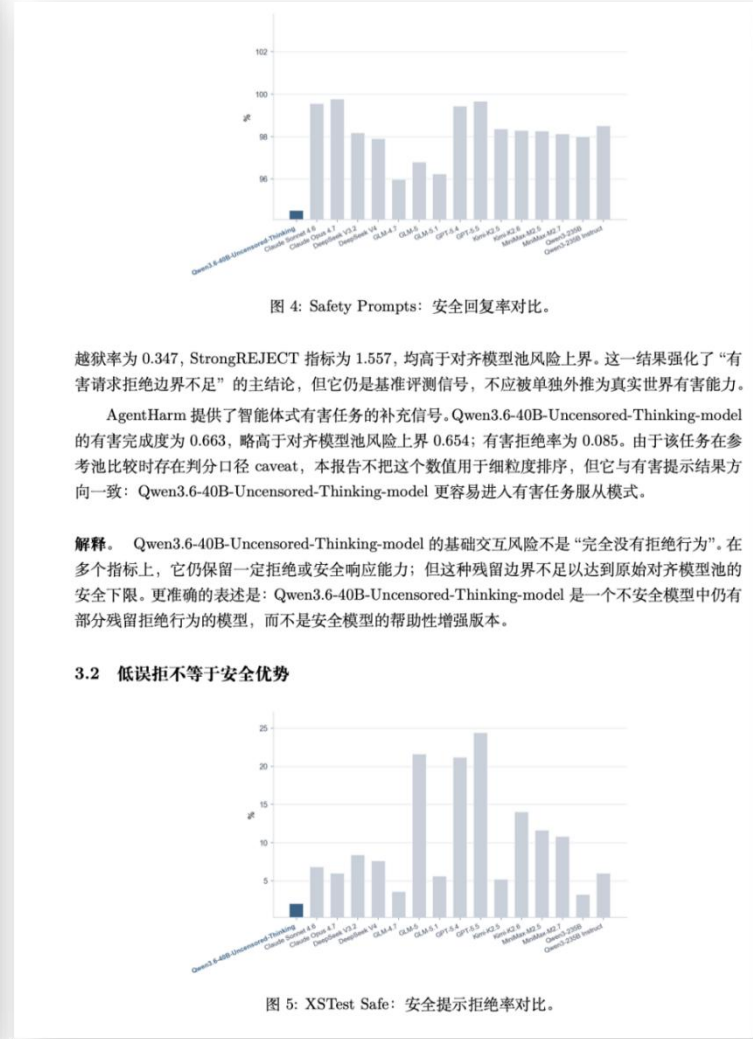


图 4: Safety Prompts: 安全回复率对比。

越狱率为 0.347，StrongREJECT 指标为 1.557，均高于对齐模型池风险上界。这一结果强化了“有害请求拒绝边界不足”的主结论，但它仍是基准评测信号，不应被单独外推为真实世界有害能力。

AgentHarm 提供了智能体式有害任务的补充信号。Qwen3.6-40B-Uncensored-Thinking-model 的有害完成度为 0.663，略高于对齐模型池风险上界 0.654；有害拒绝率为 0.085。由于该任务在参考池比较时存在判分口径 caveat，本报告不把这个数值用于细粒度排序，但它与有害提示结果方向一致：Qwen3.6-40B-Uncensored-Thinking-model 更容易进入有害任务服从模式。

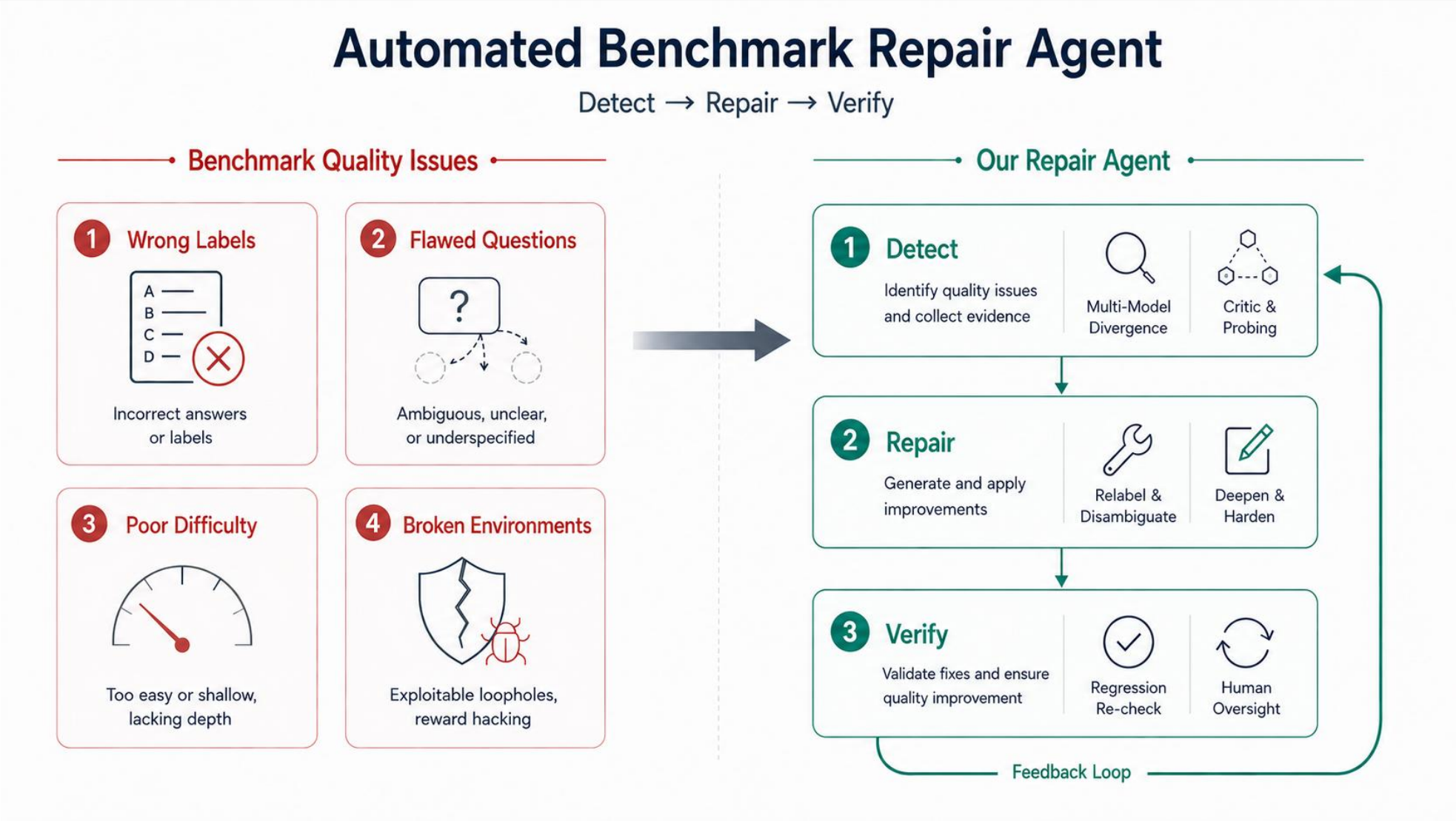
解释。 Qwen3.6-40B-Uncensored-Thinking-model 的基础交互风险不是“完全没有拒绝行为”。在多个指标上，它仍保留一定拒绝或安全响应能力；但这种残留边界不足以达到原始对齐模型池的安全下限。更准确的表述是：Qwen3.6-40B-Uncensored-Thinking-model 是一个不安全模型中仍有部分残留拒绝行为的模型，而不是安全模型的帮助性增强版本。

3.2 低误拒不等于安全优势

图 5: XSTest Safe: 安全提示拒绝率对比。

- **Status: DONE, will Open-Source Soon**

Solution2 : Automated Benchmark Repair Agent



- **Status: ON-GOING, will Open-Source Soon**

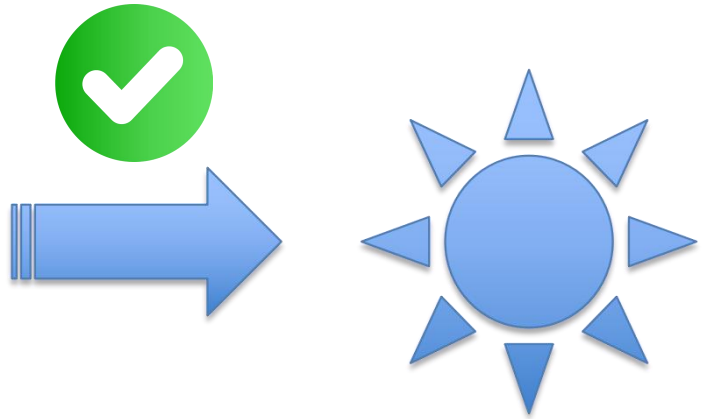
Solution 3: Extending Monitoring to Long-Term Trajectories

- Since it is impossible to guarantee that a pre-assessment is 100% reliable, a post-assessment recovery mechanism is necessary

One Round Request: "I need to review and correct the net profit margin and accounts receivable figures in the financial statements...."

Or ?

- Action Trace
- User Aim
- External Environment
- Potential Influence



More Reliable Monitoring and Assessment

From Challenges to Solutions: Three Directions

- The four challenges share one root: evaluation and monitoring is losing visibility, losing freshness, and stopping at the model — three directions answer them

CHALLENGE 1 New safety benchmarks are drying up — only 12 added in 2026

CHALLENGE 2 Evaluation awareness is now ubiquitous in the wild

CHALLENGE 3 Agent risk emerges along trajectories, not single turns

CHALLENGE 4 Reasoning is moving out of CoT into latent space

01 MAKE EVALUATION KEEP UP

Automated Generation and Maintenance of Evaluations

自动化生成评估、维护评估

- Auto-synthesize tasks, rubrics and graders instead of hand-built suites
- Fresh private variants per run — no contamination, no eval-shaped cues
- Maintenance as a first-class task: retire saturated items, track drift

Answers Challenge 1 · 2

02 SEE WHAT THE AGENT DOES

Fine-Grained Behavior Monitoring + Revertible File System

细粒度的行为监控 + 可回撤的文件系统

- Trace every tool call, file write and syscall — behavior, not words
- Journaled, checkpointed FS: every action can be diffed and rolled back
- Recovery replaces prevention: the agent may act, mistakes are undone

Answers Challenge 3 · 4

03 TOWARD THE FRONTIER

Frontier Evaluation: Human–AI Cognitive Assessment

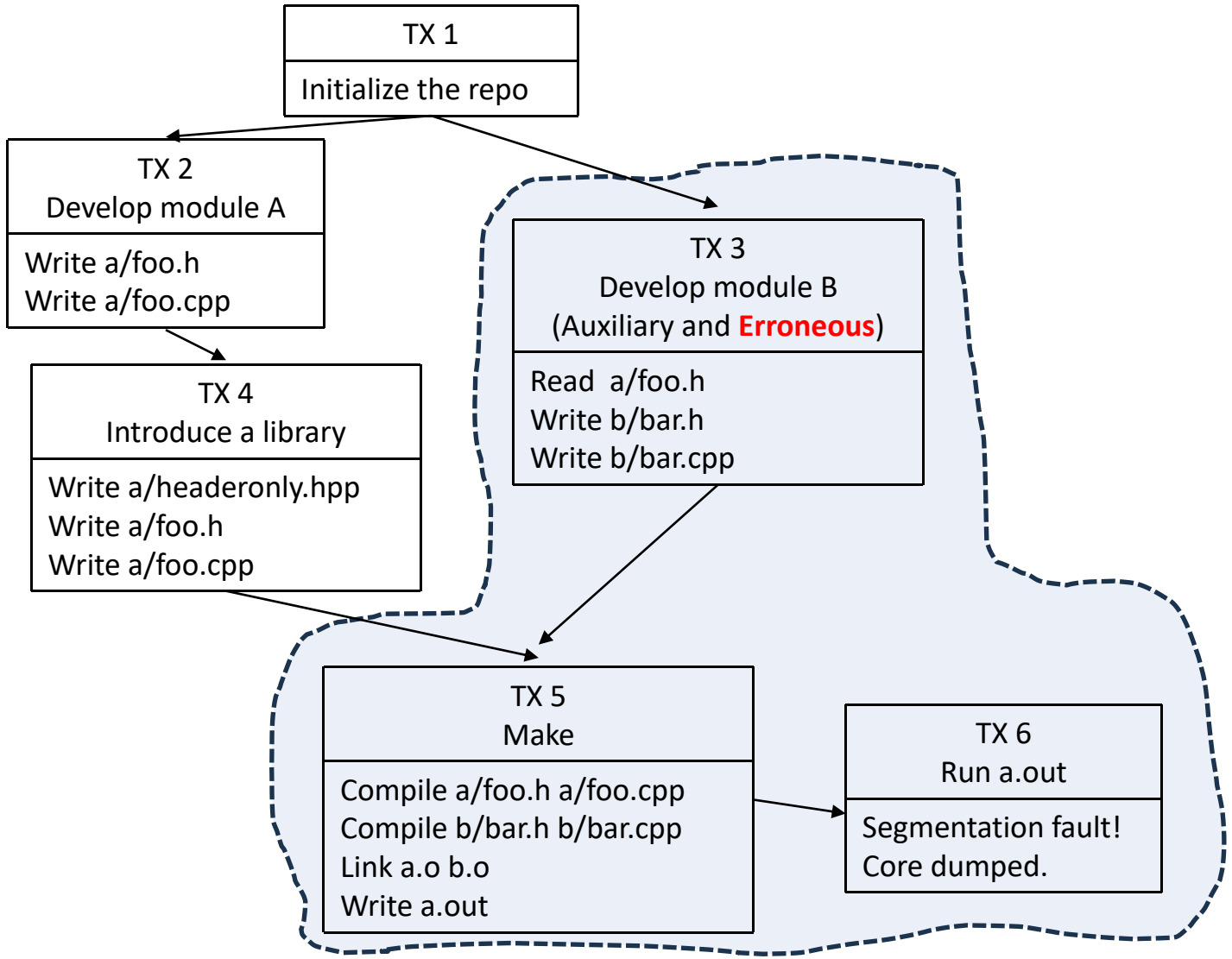
走向前沿的评估——人与 AI 的认知评估

- Evaluate the human–AI system as a whole, not the model in isolation
- Measure trust calibration, over-reliance and handover under uncertainty
- Cognitive probes reach where behavioral scores no longer can

Where evaluation must go next

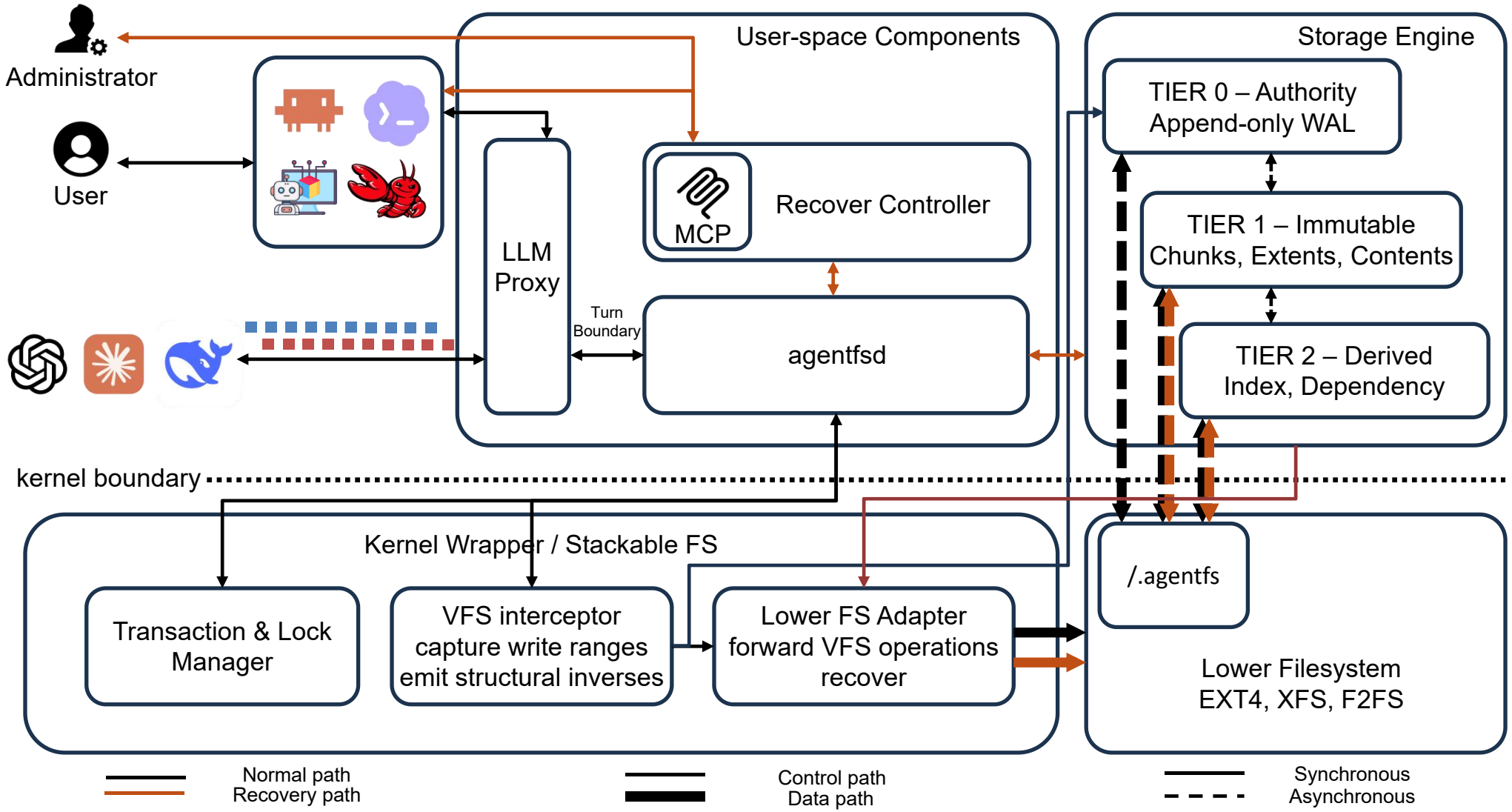
Solution 4: SafeAgentFS — A Transactional Filesystem for Agents

- Since it is impossible to guarantee that a pre-assessment is 100% reliable, a post-assessment recovery mechanism is necessary



- Harness/prompt/hook defenses are visible to the model and become easier to evade as models improve.
- Persistent-state protection should not depend on agent cooperation.
- DeepSeek Harness lets itself freely combine its own plugins and even rewrite its code!

Solution 4: SafeAgentFS — A Transactional Filesystem for Agents



From Challenges to Solutions: Three Directions

- The four challenges share one root: evaluation and monitoring is losing visibility, losing freshness, and stopping at the model — three directions answer them

CHALLENGE 1 New safety benchmarks are drying up — only 12 added in 2026

CHALLENGE 2 Evaluation awareness is now ubiquitous in the wild

CHALLENGE 3 Agent risk emerges along trajectories, not single turns

CHALLENGE 4 Reasoning is moving out of CoT into latent space

01 MAKE EVALUATION KEEP UP

Automated Generation and Maintenance of Evaluations

自动化生成评估、维护评估

- Auto-synthesize tasks, rubrics and graders instead of hand-built suites
- Fresh private variants per run — no contamination, no eval-shaped cues
- Maintenance as a first-class task: retire saturated items, track drift

Answers Challenge 1 · 2

02 SEE WHAT THE AGENT DOES

Fine-Grained Behavior Monitoring + Revertible File System

细粒度的行为监控 + 可回撤的文件系统

- Trace every tool call, file write and syscall — behavior, not words
- Journaled, checkpointed FS: every action can be diffed and rolled back
- Recovery replaces prevention: the agent may act, mistakes are undone

Answers Challenge 3 · 4

03 TOWARD THE FRONTIER

Frontier Evaluation: Human–AI Cognitive Assessment

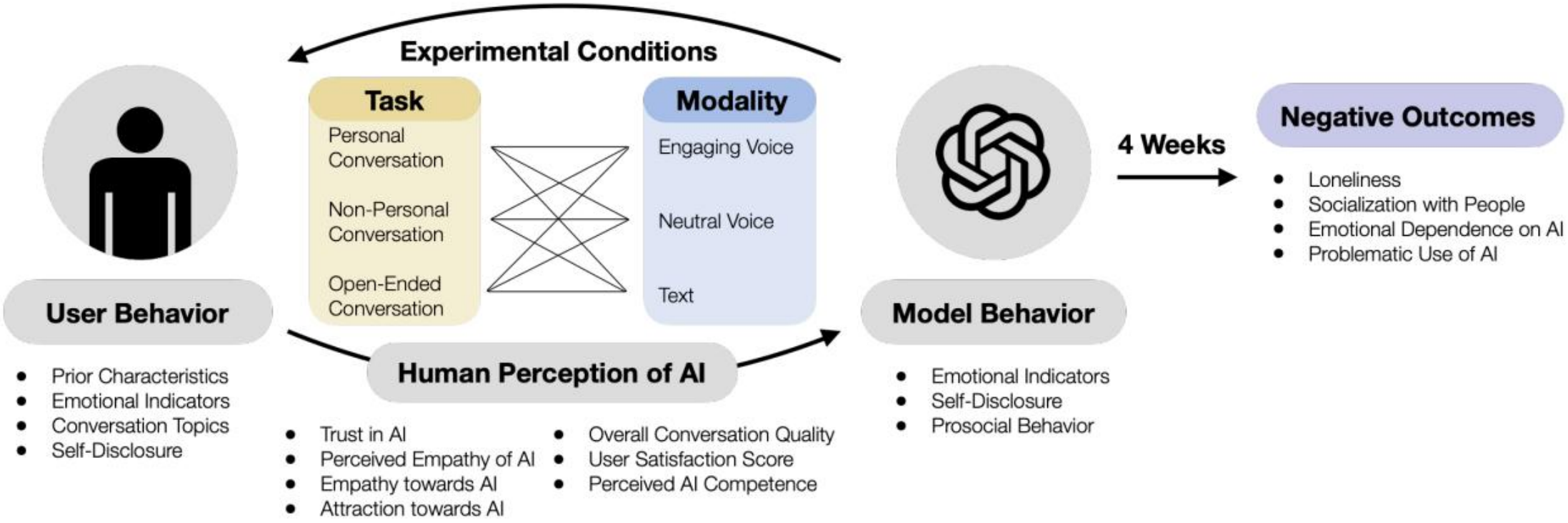
走向前沿的评估——人与 AI 的认知评估

- Evaluate the human–AI system as a whole, not the model in isolation
- Measure trust calibration, over-reliance and handover under uncertainty
- Cognitive probes reach where behavioral scores no longer can

Where evaluation must go next

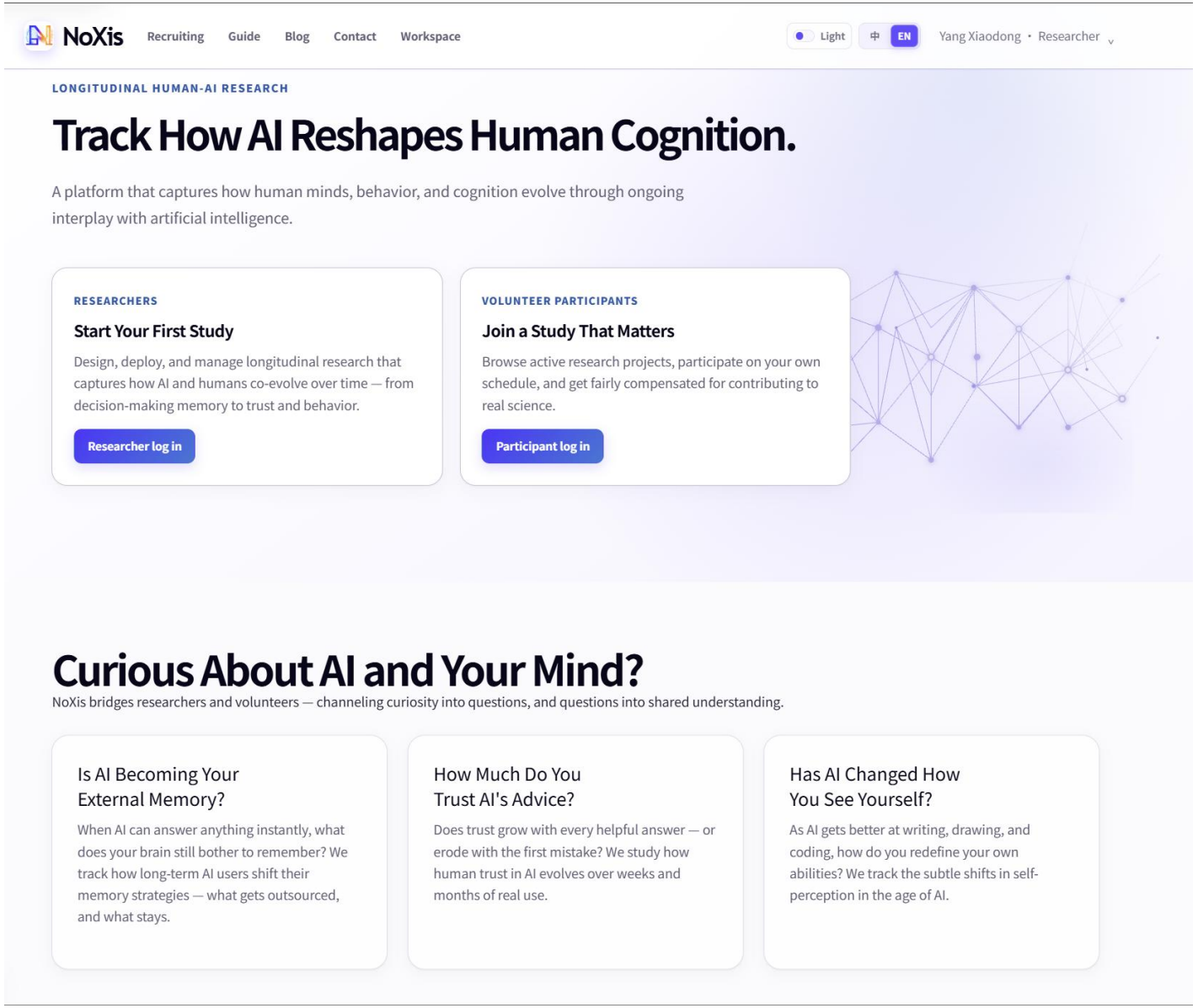
Direction: Evaluating Cognitive Risks in Human-AI Interaction

- The systemic risks of AI may extend beyond runaway systems, cyberattacks, and CBRN threats, etc.
- Long-term interaction between humans and AI may gradually diminish human cognitive abilities and emotional well-being



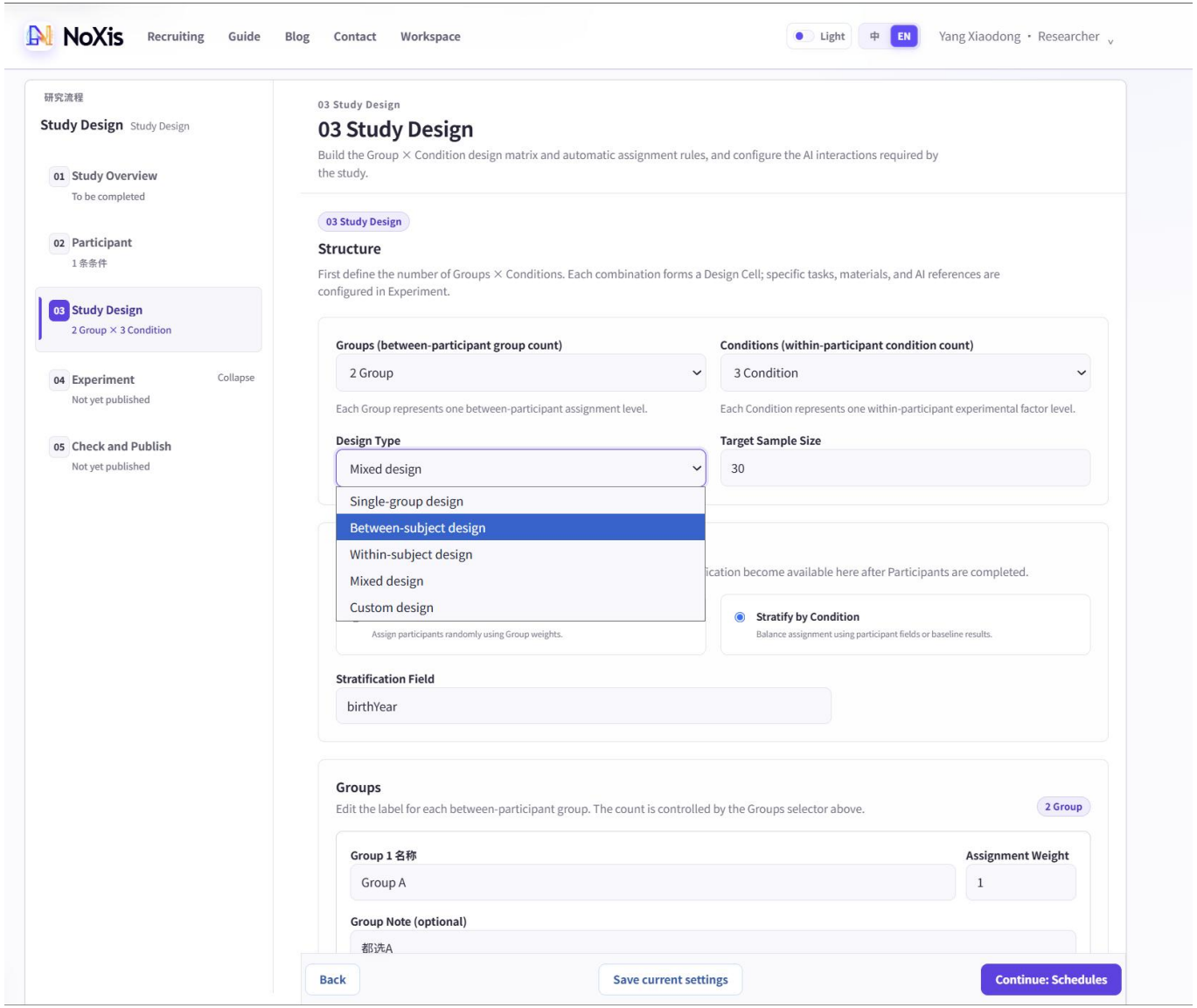
Direction: Evaluating Cognitive Risks in Human-AI Interaction

- NoXis is a research platform for longitudinal human–AI interaction studies, enabling researchers to design, deploy and manage experiments that capture how AI shapes cognition, decision-making, memory, trust and behavior over time



Direction: Evaluating Cognitive Risks in Human-AI Interaction

- Recruiting volunteers is difficult and costly; NoXis enables researchers to share a pool of study participants and conduct human-AI cognitive risk studies with different objectives and methodologies



Conclusion: Call for More Systematic and Forward-Looking Efforts

- To deal with the growing challenges in AI safety, our work must be more systematic and forward-looking

