CoTEP: A Multi-Modal Chain of Thought Evaluation Platform for the Next Generation of SOTA AI Models

Alyssia J
Independent, Oxford Alignment Research Bootcamp

Martin CL
Independent Research Engineer

Overview

As advanced state-of-the-art models like OpenAI's o-1 series, the upcoming o-3 family, Gemini 2.0 Flash Thinking and DeepSeek display increasingly sophisticated chain-of-thought (CoT) capabilities, our safety evaluations have not yet caught up. We propose building a platform that allows us to gather systematic evaluations of AI reasoning processes to create comprehensive safety benchmarks. Our Chain of Thought Evaluation Platform (CoTEP) will help establish standards for assessing AI reasoning and ensure development of more robust, trustworthy AI systems through industry and government collaboration.

1 Problem Overview

We propose that robust safety evaluations are essential for AI alignment: *what gets measured gets managed*. Without clear insights into an AI's reasoning processes, ensuring alignment with safety-critical principles is impossible.

The urgency of this work is highlighted by the rapid evolution of AI, with a projected $276 billion safety assurance market by 2030. Major AI labs, such as Anthropic, are dedicating billions toward AI safety. At the same time, governments worldwide are committing unprecedented resources to safety oversight, yet tools for evaluating reasoning remain sparse. Evaluating how AI arrives at decisions is no longer optional—it is fundamental to building trust, governance, and scalability in AI safety.

Beyond Input/Output Safety Evaluations and Benchmarks

Current capability benchmarks, such as MMLU (Massive Multitask Language Understanding), APPS (Automated Programming Progress Standard), ARC-AGI (Abstract Reasoning Corpus), SWEbench (Software Engineering Benchmark), and Frontier Math, primarily assess model performance without addressing the reasoning processes behind outputs. Similarly, public safety benchmarks, such as the UK AI Safety Institute (AISI), have made significant strides in evaluating safety-critical AI systems. AISI's recent calls for proposals on dangerous capability evaluations and autonomous system assessments demonstrate their commitment to advancing safety standards and addressing emerging challenges.

While these efforts are foundational, there remain opportunities to supplement their work by providing additional tools and frameworks to evaluate reasoning pathways in domains like cybersecurity, programming, and multi-agent systems. Our platform, CoTEP, is designed to support and enhance these initiatives by addressing gaps in reasoning transparency and providing robust evaluations to complement AISI and other safety organization's ongoing contributions.

While these efforts are foundational, there remain opportunities to supplement their work by providing additional tools and frameworks to evaluate reasoning pathways in domains like

cybersecurity, programming, multi-agent systems and beyond. We propose that our platform, CoTEP, be aimed to support and enhance these initiatives by addressing gaps in reasoning transparency and providing robust evaluations to complement AISI's ongoing contributions.

Transformative Advances in Chain of Thought Reasoning

We propose to address safety challenges stemming from the rapid evolution of CoT models. AI systems like OpenAI's o-1, the o-3 family, Gemini 2.0 Flash Thinking and DeepSeek are advancing CoT reasoning but expose significant gaps in current evaluation methodologies. These models often produce outputs through complex reasoning paths that remain opaque and potentially unsafe.

Unverified Reasoning Steps

We propose to build upon existing benchmarks like GSM8K and StrategyQA, which evaluate reasoning outputs but not the decision paths leading to those outputs. GSM8K, for instance, focuses on step-by-step math reasoning, but its scope is limited, and it lacks rigorous safety evaluations for other critical domains. The REVEAL framework, designed to verify reasoning chains, offers a foundation but requires substantial expansion to address real-world use cases.

Escalating Risks in Critical Domains

We propose solutions for mitigating unsafe reasoning in high-stakes fields like programming, finance, and healthcare, where systemic vulnerabilities or cascading failures could have catastrophic consequences. For example, a single logic error in AI-generated code could jeopardize entire software infrastructures.

Multi-Agent Complexity

We propose developing tools to monitor and evaluate multi-agent AI systems, whose interactions may produce emergent behaviors that are difficult to predict or control. Such dynamics significantly increase risks of unintended consequences, including competitive failures and resource misallocations.

Governance and Liability Gaps

We propose filling gaps in traceability for AI reasoning processes to enable responsibility assignment and effective regulatory oversight. These gaps currently undermine public trust and compliance with safety standards.

2 Proposed Solution

We propose the **Chain of Thought Evaluation Platform (CoTEP)** to evaluate and ensure the safety of CoT reasoning through:

a) Expert Session Capture System

- Capturing expert problem-solving sessions with multi-modal inputs: verbal, textual, and visual inputs.
- Annotating decisions with safety principles and context-specific risks.

b) Knowledge Graph Construction

- Translating expert and AI reasoning into structured, comparable formats.
- Linking decision points to safety-critical concepts and principles.

c) Evaluation Framework

- Benchmarking AI reasoning against expert baselines.
- Identifying reasoning vulnerabilities, including logical gaps and potential misuse scenarios.
- Expanding metrics from frameworks like REVEAL to include safety-focused criteria.

Competitive Edge

- First CoT platform explicitly targeting safety-critical domains.
- Scalable design, integrating with industry tools and regulatory standards.
- Collaborative partnerships with government and private sector entities for adoption.

3 Pilot Experiment or Demo

For this hackathon, we built:

1. **A Proof of Concept**:
    - A multi-modal platform integrating with IDEs to capture real-time expert reasoning (demo prototype video attached to submission!)
    - An initial schema for knowledge graph construction, linking decisions to cybersecurity principles.
2. **A Prototype Evaluation Framework**:
    - Sample benchmarking AI outputs against annotated expert traces.
    - Sample visualized gaps in reasoning chains with generated reports.

4 Process

Timeframe	What will you do?
Next 3 months	• Build initial MVP evaluation suite focusing on chain of thought evaluations. • Secure 1-3 pilot partnerships with AI labs for early testing and feedback. • Submit comprehensive evaluation frameworks to government AI safety agencies. • Publish research paper on chain of thought evaluation benchmarks. • Present findings at major AI safety conferences.
2025	• Launch beta version of our core evaluation platform. • Develop licensing model for our evaluation software. • Expand evaluation suite to cover additional safety benchmarking categories. • Secure first annual contracts with AI labs. • Build partnerships with 2-3 research institutions.
2026	• Integrate evaluation tools with major AI development frameworks. • Establish certification programs for safe AI systems. • Publish longitudinal studies on safety improvements from CoT evaluations. • Expand partnerships with government regulatory agencies.
2027	• Achieve global adoption of CoTEP. • Standardize CoT safety evaluations as part of international AI safety regulations. • Build advanced multi-agent evaluation frameworks. • Host an annual conference on AI safety evaluation innovations.

Table 1: Proposed timeline for CoTEP development and commercialization.

5 Impact on AI Safety & Key Risks

1. **Direct Safety Benefits**: Mitigates risks of reasoning failures in safety-critical applications; enhances traceability and accountability in AI decision-making; aligns AI development with evolving regulatory standards.

2. **Indirect Benefits**: Advances AI interpretability research; establishes benchmarks for safe AI governance; builds public trust in AI systems through transparent evaluations.

3. **Risks & Mitigation**: Address adoption risks through early partnerships with industry and government; mitigate scalability challenges via modular design and automated annotation tools.

6 References

1. Introducing OpenAI o1-preview - Introducing OpenAI o1-preview and CoT capabilities.
2. OpenAI o3 and o3-mini—12 Days of OpenAI: Day 12 - OpenAI o3 and o3-mini—12 video announcement during 12 days of OpenAI.
3. Introducing Gemini 2.0: our new AI model for the agentic era - Gemini 2.0 model features rapid CoT reasoning through "Flash Thinking".
4. Introducing DeepSeek-V3 - DeepSeek's new model with CoT.
5. AISI Dangerous Capabilities and Autonomous Agent Scaffolding Bounty Announcement / Call for Proposals - Official announcement for new evaluations.
6. Anthropic to raise at $60B valuation - Anthropic raising billions for AI safety.
7. Introducing the AI Safety Institute - UK Government Formation of AISI to address AI safety.
8. MMLU Benchmark - Comprehensive benchmark for multitask language understanding.
9. APPS Benchmark - Automated Programming Progress Standard for AI evaluation.
10. StrategyQA benchmark - A question answering benchmark where the required reasoning steps are implicit in the question.
11. Frontier Math Benchmark - A math benchmark testing the limits of AI.
12. Abstraction and Reasoning Corpus for Artificial General Intelligence (ARC-AGI) - In-depth evaluation framework for abstract reasoning.
13. ThoughtSource Dataset - Central resource for CoT reasoning datasets.
14. REVEAL Framework - Benchmark for verifying reasoning chains, requiring further expansion to address broader use cases and generate extensive safety evaluations.
15. 2024 AI Assurance Technology Report - Insights into AI risk management and governance.
16. Scaling Language Models for Chain-of-Thought Reasoning - Advances in CoT reasoning.
17. Emergent Behaviors in Multi-Agent Reinforcement Learning - Safety risks in multi-agent systems.
18. GSM8K Dataset - Step-by-step grade school math reasoning dataset, highlighting the need for broader safety evaluations across domains.