Frontier AI Evaluation, Agentic Systems & Behavioral Safety

Professional Experience | 2024–Present

For more than two decades, my career has centered on how people interpret information, make decisions, and interact with complex systems. Over the past two years, I've brought that foundation in behavioral science, content strategy, UX, and editorial leadership into professional frontier AI evaluation.

Working on confidential projects for some of the world's leading AI developers, I've helped stress-test advanced language, multimodal, and agentic systems, deliberately challenging their capabilities to discover where, how, and why they fail.

My work extends beyond determining whether models produce correct answers. It involves designing complex challenges, identifying failure patterns, investigating underlying causes, and documenting findings that can inform further development. I've also served as a final reviewer on multiple projects, bringing an additional level of analytical and editorial judgment to the quality assurance process.

Frontier Model Capabilities & Failure Analysis

My experience spans language-based, multimodal, and agentic AI systems, including autonomous agents operating within simulated environments.

This work demands close examination of context, reasoning, task interpretation, behavioral consistency, and the relationship between a model's intended capabilities and its actual performance.

I've contributed to sophisticated initiatives designed to expose weaknesses in emerging systems, assess their significance, and distinguish isolated errors from systemic problems.

Quality Assurance & Final Review

On multiple frontier AI projects, I've served as a final reviewer, responsible for determining whether submitted work meets established quality standards.

This role requires assessing accuracy, completeness, consistency, and adherence to project-specific requirements. It also demands the ability to recognize subtle analytical weaknesses, distinguish substantive problems from differences in interpretation, and exercise independent judgment.

My background in editorial leadership provides a strong foundation for this responsibility. Years of overseeing complex content projects, maintaining quality standards, and resolving inconsistencies inform my approach to reviewing highly technical work.

Agentic AI, Human Agency & Safety

A particularly important area of my experience involves AI agents capable of carrying out multistep tasks and interacting with connected services in controlled environments.

These systems introduce challenges that extend beyond traditional conversational accuracy. An agent may correctly understand a user's desired outcome while misunderstanding what actions it has permission to take.

My work explores critical distinctions:

- Understanding user intent versus assuming authorization.
- Providing useful assistance versus taking inappropriate initiative.
- Completing a task versus respecting constraints on how it may be completed.
- Technical capability versus appropriate behavioral judgment.

These distinctions are essential to developing autonomous systems that are useful, reliable, and accountable to the people they serve.

Emotional Intelligence, Empathy & Response Appropriateness

My background in psychology and mental-health communication provides an additional perspective on AI behavior.

I've assessed emotional intelligence, empathy, tone, contextual sensitivity, and response appropriateness, including interactions involving potential self-harm, harm to others, and security-related risks.

This work requires recognizing when a response is helpful, misleading, unnecessarily alarming, insufficiently sensitive, or potentially harmful. It also involves considering how language choices influence user understanding, trust, and decision-making.

Effective AI communication requires more than sounding friendly. A trustworthy system must respond proportionately to the situation, recognize when additional care or restraint is warranted, avoid misleading reassurance, and preserve the user's ability to make informed decisions.

A Distinctive Interdisciplinary Perspective

My approach draws on four complementary areas of expertise:

Behavioral science: Understanding motivation, emotional responses, decision-making, learning, and human behavior.

Content strategy and UX: Designing communication that is clear, accessible, consistent, and responsive to user needs.

Frontier AI systems: Investigating advanced capabilities through structured challenges, adversarial testing, failure analysis, and precise documentation.

Editorial leadership and quality assurance: Maintaining rigorous standards, identifying subtle inconsistencies, and exercising independent judgment in final-review roles.

Together, these disciplines allow me to examine sophisticated AI systems from multiple perspectives: their technical performance, their interactions with users, the appropriateness of their behavior, and the quality of the processes used to assess them.

Confidentiality note: Specific clients, models, evaluation environments, datasets, and proprietary testing methodologies are not identified in accordance with professional confidentiality obligations.