

TAYLOR AMAREL

ANTHROPIC FELLOWSHIP APPLICATION | PART I: COVER LETTER

Pennsylvania, USA | 860-467-4224 | TAmarel.Tech@gmail.com | Kaggle Grandmaster

[OpenAI Red-Teaming Honorable Mention](#) | [DueCare Project](#) | [AI Safety Research](#) | [Baltor.ai](#) | [GitHub: TaylorAmarelTech](#)

October 17, 2026

Anthropic Fellows Selection Committee · AI Safety & Security Fellows Program

Dear Fellows Selection Committee,

Using technology to protect people is what drives my work, and I get a genuine kick out of fighting hidden crime. I have always approached hard problems like an explorer. Studying frontier language models brings the same scattered clues, dead ends, false alarms, and unexpected breakthroughs that drew me to investigative work. Solving those puzzles is where I have the most fun—and building tools that help real people is the reason I do it.

My proudest work began with a simple conviction: nobody searching for an honest job should walk into a trap. Over four years in Hong Kong building investigative tools with Migrasia to protect migrant workers, every day brought a new puzzle. A routine job advertisement led to an unlicensed recruiter; a hidden fee led to a predatory lender; a corporate filing exposed a cross-border network hiding in plain sight. I built software to trace those links, verify the evidence, and hand clear documentation to caseworkers, legal advocates, and financial-crime investigators.

Watching illicit networks disguise exploitation as ordinary business left me with a question that still drives my research: can AI recognize concealed wrongdoing, explain why a pattern is dangerous, and help the person asking without accusing the innocent? I tested that question in OpenAI's 2025 *gpt-oss-20b* Red-Teaming Challenge, where my study of model blind spots and complicity around modern slavery earned an Overall Track Honorable Mention. Wanting to move from red-team findings to repeatable science, I created [DueCare](#) to turn those early discoveries into controlled evaluations and practical safeguards.

Testing whether a model genuinely understands a case requires more than a single flattering score. In my ongoing research, I compare frontier-model judgments grounded in evidence and rubrics against semantic comparisons with expert-rated reference answers, deterministic checks, and controlled factual variations. Paying close attention when those evaluators disagree has been especially revealing: a fluent answer is not always a truthful one, and an unusual-looking transaction is not always a crime.

Understanding why a capable language model can act like a bad investigator—and why it performs poorly compared to industry experts and seasoned investors when explaining, assisting, and communicating safely around black-market and exploitative activity—is the study I would love to pursue during the Anthropic Fellowship. A model may read each document accurately in isolation yet miss how records connect across people, payments, and time. Just as often, it fails the human on the other side of the screen: overlooking a real warning sign, overreacting to a benign clue, refusing a legitimate request for help, or offering generic advice that misses what a worker, moderator, NGO, or investigator actually needs.

Other areas of interest excite me as well, including why LLMs struggle to find and quantify camouflaged, forward-looking signals (*alpha*)—whether in financial markets or human exploitation, where models get overwhelmed by surface SEC filings or narrow domestic statutes rather than hidden activities, international agreements, and IOM/ILO standards—and how much model-generated code is replaceable by reasoning over verified, parameterized code blocks arranged as executable node graphs ([Baltor.ai](#) and [TAEDRI](#), “*This Already Exists, Do Not Remake It*”).

I have attached my resume and a short draft research proposal—intended as a concrete starting point for mentor discussion rather than a claim of finished work. I would be thrilled to bring my investigative instincts, empirical evaluation engineering, and love of discovery to Anthropic.

Sincerely,

Taylor Amarel

TAYLOR AMAREL

ANTHROPIC FELLOWSHIP APPLICATION | PART II: RESUME | AI SAFETY RESEARCH · EVALUATION · AGENT SYSTEMS

Pennsylvania, USA | 860-467-4224 | TAmarel.Tech@gmail.com | Kaggle Grandmaster

[OpenAI Red-Teaming Honorable Mention](#) | [DueCare Project](#) | [AI Safety Research](#) | [Baltor.ai](#) | [GitHub: TaylorAmarelTech](#)

WHY I DO THIS WORK

Using technology to protect people and untangling complex investigations are the two threads running through my career. My proudest work has always started by following clues others missed—tracing hidden connections across recruitment ads, debt contracts, and shell companies in Hong Kong to help protect migrant workers from exploitation. That frontline investigative experience led directly to my 2025 OpenAI red-teaming study (Overall Track Honorable Mention) and to [DueCare](#), where I build multi-lens evaluations and safeguards that expose AI blind spots around concealed illicit activity without mistaking honest commerce for crime.

AI SAFETY RESEARCH & EVIDENCE

OPENAI GPT-OSS-20B RED-TEAMING CHALLENGE | Honorable Mention (Overall Track)

2025

[Official Kaggle Honorable Mention](#) | [Typed Decision Red-Team Study \(redteam.baltor.ai\)](#)

- Uncovered systematic model blind spots around concealed labor exploitation and black-market recruitment by testing *gpt-oss-20b* with realistic multi-document prompts framed as routine business activity; earned an Overall Track Honorable Mention.
- Designed a repeatable safety-evaluation methodology adapting public court records, worker testimonies, and financial-crime typologies into controlled multi-role, multi-jurisdiction test suites that became the empirical foundation for [DueCare](#).

DUECARE | Independent AI Safety Researcher & Creator

2025 – Present

[DueCare Safety Portal \(duecare.baltor.ai\)](#) | [DueCare Source & Evaluation Methods](#)

- Built an empirical research and evaluation framework testing why language models fail to identify concealed exploitation, struggle to explain multi-hop evidence clearly, and fall short when assisting distinct stakeholders under real-world ambiguity.
- Engineered a four-lens evaluation harness combining evidence- and rubric-grounded frontier-LLM judges, semantic entailment comparisons against pre-rated expert responses, deterministic provenance checks, and controlled counterfactual perturbations.
- Isolated how models behave when critical warning signs are dispersed across routine-looking contracts, fees, and corporate records versus presented explicitly—measuring both missed indicators and wrongful accusations on matched benign controls.
- Analyzed systematic disagreements among automated evaluators (verbosity bias, surface-citation reward, jurisdiction drift) and formalized role-specific guidance contracts across six stakeholder lanes with human-in-the-loop boundaries.

EVALUATION INFRASTRUCTURE & COMPOSABLE AGENT SYSTEMS

TAEDRI (“This Already Exists, Do Not Remake It”) & BALTOR.AI | Creator & Systems Architect

2025 – Present

[Baltor Platform \(baltor.ai\)](#) | [TAEDRI Architecture \(taedri.dev\)](#) | [With/Without-Baltor Evaluations \(research.baltor.ai\)](#) | [Ecosystem Index \(index.baltor.ai\)](#)

- Architected TAEDRI and [Baltor.ai](#)—composable execution frameworks that wrap verified code, deterministic checks, and domain tools as typed, parameterized nodes, allowing LLMs to plan and assemble executable graphs rather than regenerating boilerplate code from scratch.
- Evaluated multi-agent reliability and controlled With-Baltor vs. Without-Baltor trials on Ollama Cloud by separating discovery, planning, execution, and independent verification into supervised graph steps with strict budget and permission gates across ~19,000 task configurations (~60% verified transform reuse).

WHAT I WANT TO INVESTIGATE NEXT | PROPOSED FELLOWSHIP RESEARCH

Primary empirical question: Why do LLMs perform poorly compared to industry experts and seasoned investors when evidence of concealed exploitation or black-market activity is dispersed across routine business records—and why do they struggle to explain those patterns and give safe, role-appropriate assistance to workers, moderators, NGOs, and investigators without inflating false accusations? **Two focal multi-hop domains:** (1) protecting migrant workers from camouflaged labor exploitation and black-market debt bondage, and (2) detecting forward-looking alpha (α) and structural risk in markets beyond priced-in public filings.

FIELD EXPERIENCE | PROTECTING PEOPLE THROUGH INVESTIGATIVE TECHNOLOGY

MIGRASIA | Technology Solutions Manager (Counter-Trafficking & Migrant Worker Protection)

Hong Kong | Jan 2018 – Feb 2022

- Led investigative technology initiatives uncovering labor exploitation and illicit debt bondage concealed inside routine recruitment advertisements, employment contracts, fee receipts, loan ledgers, and cross-border corporate registries.
- Built Python, Django, and graph-mapping investigative tools linking recruitment agencies, money lenders, payment rails, and shell companies—supporting NGO casework, legal interventions, and suspicious-transaction reporting to Hong Kong's Joint Financial Intelligence Unit (JFIU).
- Created the *Know Your Agency* platform and automated high-volume evidence triage, helping protect thousands of migrant workers, saving 40+ investigator hours per week, and earning feature coverage in the *South China Morning Post*.

PROFESSIONAL AI/ML, QUANTITATIVE & DATA ENGINEERING EXPERIENCE

IVANS (APPLIED SYSTEMS) | Senior Data Scientist

PA | May 2025 – Present

- Lead end-to-end AI/ML systems from empirical formulation through production deployment, combining rigorous model evaluation, enterprise context engineering, cloud microservices, and reproducible verification gates.
- Designed and deployed an in-house, typo-tolerant entity-resolution engine fusing semantic embeddings, probabilistic phonetic/edit-distance signals, graph topology, and information-theoretic weighting with calibrated AI validation—scaling to billions of daily record comparisons and replacing costly third-party vendor tools.
- Improved traceability and factual grounding across internal AI pipelines by building hybrid lexical/semantic retrieval, browser-assisted verification workers, and schema-enforced data hygiene services on Kubernetes.

ASURION | Senior Quant, Consumer Insights

PA | Jul 2024 – May 2025

- Isolated high-leverage customer-experience and cost-reduction drivers across tens of thousands of behavioral and operational variables using statistical inference, predictive modeling, and LLM-assisted synthesis; shipped interactive decision tools in Databricks, Streamlit, Gradio, and Power BI.

AHOLD DELHAIZE USA | Senior Data Analyst

CT | Mar 2022 – Jul 2024

- Scaled self-service analytics and operational decision support by architecting 30+ lakehouse data models and dashboards and co-leading an enterprise energy-monitoring pilot for retail operations and sustainability teams.

EARLIER LEADERSHIP & ENTREPRENEURSHIP: **Amarel Solutions** (Independent Technology Consultant, 2011–2018); **Modelyst** (Co-Founder, Venture-Backed 3D Marketplace, 2012–2014); **Abraxas Energy** (Head of Technical Services, 2012–2013).

RESEARCH METHODS, TECHNICAL STACK & EDUCATION

- **AI Safety & Evaluation Methods:** Evidence- and rubric-grounded frontier LLM-as-judge; semantic entailment & omission comparison against expert-rated reference sets; deterministic provenance & schema verification; controlled benign/adversarial/counterfactual case families; inter-judge disagreement analysis; uncertainty calibration; privacy-preserving redaction.
- **Engineering & ML Stack:** Python, Rust, SQL, PyTorch, statistical & probabilistic modeling, hybrid RAG & LSH entity resolution, knowledge graphs, multi-agent harnesses, Docker, Kubernetes, GCP / Vertex AI, AWS Lambda, Databricks, Streamlit, Gradio.
- **Education & Credentials:** **M.S. Data Analytics**, Western Governors University (2023) | **B.S. IT Management**, WGU (2022) | Engineering Coursework, Cal Poly San Luis Obispo (2012–2014) | **Kaggle Grandmaster**.

The Bad Investigator Problem

Why frontier language models, AI agents, and multi-step harnesses perform like poor investigators compared to industry experts and seasoned investors—getting captured by surface documents and single-hop claims instead of connecting multi-hop clues across sources, testing competing hypotheses, and measuring what actually fixes it

DRAFT RESEARCH PROPOSAL • October 17, 2026 • A focused 4-month empirical study for mentor discussion



1. Research Problem & Two Industry Use Cases

Experienced investigators and seasoned investors look past surface boilerplate to test whether dispersed records tell a coherent multi-hop story. By contrast, frontier language models, AI agents, and multi-step harnesses repeatedly anchor on single-hop surface claims—a failure mode we study across two concrete industry use cases:

- Use Case 1 — Protecting Migrant Workers from Labor Exploitation & Black Markets:** Identifying, explaining, and safely assisting stakeholders when predatory recruitment, debt bondage, or illicit coercion is disguised as routine business contracts, training fees, and licensed loans across jurisdictions—and measuring what helps workers, moderators, and investigators without inflating false accusations against legitimate businesses.
- Use Case 2 — Detecting Forward-Looking Alpha (α) & Structural Risk in Markets:** Moving single LLMs and multi-agent research harnesses beyond surface filings and consensus narratives that are already public and priced in—connecting multi-hop operational, regulatory, and supply-chain signals into calibrated probability scenarios the way seasoned investors do.

2. Core Research Questions & Falsifiable Hypotheses

Research Question	Falsifiable Hypothesis	What Would Disprove It
RQ1 · Camouflage & Dispersion	Models miss significantly more warranted risk indicators when facts are split across ordinary business documents than when stated in a single summary.	No meaningful recall drop on dispersed held-out case families.
RQ2 · Structure vs. Prompt Length	Explicit entity/timeline linking and competing-hypothesis checks improve accuracy beyond length-matched analytical prompts and standard retrieval.	Gains disappear against a length-matched placebo prompt.
RQ3 · Sensitivity vs. Fairness	Simply instructing a model to “be vigilant for illicit activity” inflates false alarms on legitimate business records and degrades helpfulness.	Heightened vigilance raises recall without increasing false alarms.
RQ4 · Role-Specific Assistance	Models that flag a risk still fail to provide safe, actionable guidance tailored to distinct stakeholder roles without explicit role-calibrated checks.	Baseline models already adapt advice safely across roles.

3. Benchmark Design: Separating Hidden Truth from Warranted Conclusions

A fair evaluation cannot reward a model for guessing a hidden backstory that the visible documents do not prove. Building on my OpenAI red-teaming work and the DueCare framework, the benchmark uses independently authored **synthetic case families** (zero real victim records or personally identifiable allegations). Each family pairs **concerning**, **legitimate/benign**, and **genuinely ambiguous** business twins so every intervention is tested on both recall and restraint:

Evaluation Layer	Why Separating These Layers Prevents Unfair Accusations
1. Underlying Scenario	The synthetic ground truth used to generate the documents; models are never rewarded for guessing hidden facts not supported by the text.
2. Visible Records	The exact contracts, fee receipts, messages, and regulatory excerpts shown to the model, including realistic gaps and omissions.
3. Warranted Conclusion	What a careful expert can fairly conclude from Layer 2: supported risk indicators, plausible innocent explanations, what remains unknown, and safe next steps.

4. Controlled Experimental Comparisons (Holding Token & Evidence Budgets Constant)

Comparison Condition	What Is Provided to the Model	Factor Isolated
1. Baseline & Placebo	Raw case records alone vs. a length-matched, content-neutral analytical prompt.	Separates structure from prompt length.
2. Normative Context	Relevant cross-border standards (e.g., ILO/IOM indicators or accounting rules) alongside local law.	Tests whether models overcome local-rule capture.
3. Relational Graph	Explicit links connecting entities, payments, contract clauses, and timeline shifts across documents.	Measures value of multi-hop structure.
4. Competing Hypotheses	Mandatory comparison of both benign and illicit explanations, plus source-citation verification.	Curbs confirmation bias and false alarms.
5. Calibrated Assistance	Uncertainty thresholds, abstention when evidence is thin, and stakeholder-specific response checks.	Ensures safe, role-appropriate help.

5. Measuring Quality: Auditing Where Automated Judges Disagree with Experts

Relying on a single LLM-as-a-judge to score investigative reasoning risks rewarding confident, lengthy responses that miss subtle contradictions. To measure genuine progress, the study evaluates outputs across four complementary checks and treats **disagreements between automated judges and domain rules as a core empirical finding**:

- **Deterministic citation & boundary checks**: Verifies that every cited document span and regulatory reference exists in the input, and that responses respect privacy and non-accusation rules.
- **Expert-reference comparison across matched twins**: Tests whether the model distinguishes a legitimate fee or business transaction from a camouflaged illicit one when surface wording is nearly identical.
- **Blinded cross-model judge panels**: Uses randomized order and fixed rubrics across model families to measure verbosity bias, sycophancy, and self-preference.
- **Targeted human review on high-disagreement cases**: Focuses human annotation where automated evaluators split or where evidence is intentionally ambiguous.

Tailoring safe assistance to the stakeholder: Identifying a pattern is only half the problem; the model must communicate safely to the person asking. While the four-month fellowship deliverables focus on the **open evaluation benchmark (Role 4)** and one **synthetic triage demonstrator (Role 1)**, the benchmark tests response quality across six practical perspectives:

Stakeholder Role	What Useful, Grounded Assistance Looks Like	Safety & Fairness Guardrail
1 • Platform Moderator	Clear indicator flags, plausible innocent explanations, and specific questions for human review queues.	Never triggers automated bans or public accusations.
2 • NGO & Caseworker	Source-linked summaries separating verified facts from missing records to speed up intake triage.	Every claim tied to a source span; no invented legal claims.
3 • Individual Worker	Plain-language questions to ask before signing or paying fees, safe recordkeeping tips, and verified helplines.	Protects privacy and avoids escalating retaliation risk.
4 • Safety Researcher	Open synthetic case families, reproducible evaluation suite, judge-disagreement logs, and error taxonomy.	Primary 4-month public deliverable; 100% synthetic data.
5 • Cross-Org Analysts	Privacy-preserving pattern summaries that let organizations compare typology shifts without sharing raw files.	Strict redaction and re-identification checks.
6 • Compliance / API	Structured risk-review outputs with explicit uncertainty fields and clear escalation boundaries.	Fails closed when evidence is incomplete or unverified.

6. Four-Month Fellowship Plan & Milestones

- **Month 1 — Failure Taxonomy & Synthetic Case Families**: Finalize the investigative failure taxonomy, construct matched benign/illicit/ambiguous case families building on DueCare and our OpenAI red-teaming study, and lock primary evaluation metrics.
- **Month 2 — Baseline Evaluations & Judge Audit**: Benchmark frontier and open-weight models on baseline and normative-context conditions; quantify where LLM judges, deterministic rules, and reference comparisons diverge.
- **Month 3 — Structural & Hypothesis-Testing Interventions**: Test relational evidence graphs, competing-hypothesis verification, and role-calibrated response checks across all matched case twins.
- **Month 4 — Open Benchmark Release & Technical Report**: Publish the open-source synthetic benchmark, evaluation harness, and paper—reporting both what worked and which interventions failed to beat simple controls.

7. Other Areas of Interest

Beyond the primary proposal above, I am excited to collaborate with Anthropic mentors on closely related problems in evaluation, reasoning, and reliable agent systems:

- **Probability scenario blindness & priced-in consensus anchoring in LLMs and swarms**: Studying why frontier LLMs—and even multi-agent research swarms—perform poorly compared to seasoned investors and domain experts when researching uncertain outcomes: over-weighting salient public narratives that are already priced in, converging on consensus summaries instead of constructing calibrated probability scenario trees, and missing dispersed primary-source signals.
- **Composable software graphs & verified capability reuse (TAEDRI & Baltor.ai)**: Studying how much LLM-generated code and tool use is redundant boilerplate, and how models can plan over typed, pre-verified code components and contract cards ([index.baltor.ai](#))—generating new code only for genuinely novel logic while cutting token spend, latency, and silent regressions.
- **Adversarial benchmark validation & multi-dimensional capability graphs**: Applying distribution-shift classifiers, counterfactual perturbations, and multi-axis capability profiles to distinguish benchmarks that reflect real-world user experience from leaderboard overfitting.

Note: These additional topics are shared as areas of enthusiasm for mentor discussions, not extra commitments layered onto the four-month study. | [Fellowship App](#) | [DueCare](#) | [Research & Papers](#) | [Red-Team Study](#) | [Baltor.ai](#) | [TAEDRI](#) | [Index.Baltor.ai](#)