

AI Safety Entrepreneurship White Paper

AntiMidas: Building Commercially-Viable Agents for Alignment Dataset Generation

Jacob Arbeid Jay Bailey Sam Lowenstein Jake Pencharz
UK AI Safety Institute UK AI Safety Institute UK AI Safety Institute UK AI Safety Institute

1 Overview

AI alignment lacks high-quality, real-world preference data needed to align agentic superintelligent systems. Our technical innovation builds on Pacchiardi et al. (2023)’s breakthrough in detecting AI deception through black-box analysis. We adapt their classification methodology to identify intent misalignment between agent actions and true user intent, enabling real-time correction of agent behavior and generation of valuable alignment data. We commercialise this by building a workflow that incorporates a classifier that runs on live trajectories as a user interacts with an agent in commercial contexts. This creates a virtuous cycle: our alignment expertise produces superior agents, driving commercial adoption, which generates increasingly valuable alignment datasets of trillions of trajectories labelled with human feedback: an invaluable resource for AI alignment.

2 Problem Overview

Existing AI safety measures focused on model capability evaluations or control do not scale to superintelligence; only robust alignment to human intent can ensure existential security when we enter the age of ASI. Yet current alignment research exhibits one of two failure modes: it is either overly theoretical and has not yielded applicable insights for real-life AIs (e.g. agent foundations research), or it is data-driven but involves data that does not scale to aligning superintelligence. RLHF (Christiano et. al. 2017) is posited as a data-driven solution to alignment, but it has two principle problems:

- **Data quality:** Researchers have observed that RLHF data, which mostly at source comes from poorly-paid data labellers, often does not reflect genuine human preferences. (Casper et. al. 2023) Data-labellers have little incentive to select the ‘more aligned’ solution between two options.
- **Ecological validity:** RLHF data involves selecting between static outputs: this is relevant to chatbot alignment but does not scale to more complex agent tasks which are more relevant to aligning superintelligence.

We are fundamentally not equipping agents to be aligned across their entire trajectories with the current RLHF regime; RLHF will not scale to aligning superintelligent agents.

We propose addressing this critical gap by leveraging commercial agent deployments to sustainably collect alignment-relevant data at scale. This approach enables the systematic study of human-agent alignment while creating immediate commercial value in the form of valuable

agents in domains like e-commerce, accountancy and law. Our product creates a virtuous cycle between better agents, higher-quality datasets, and improved alignment research.

3 Solution Overview

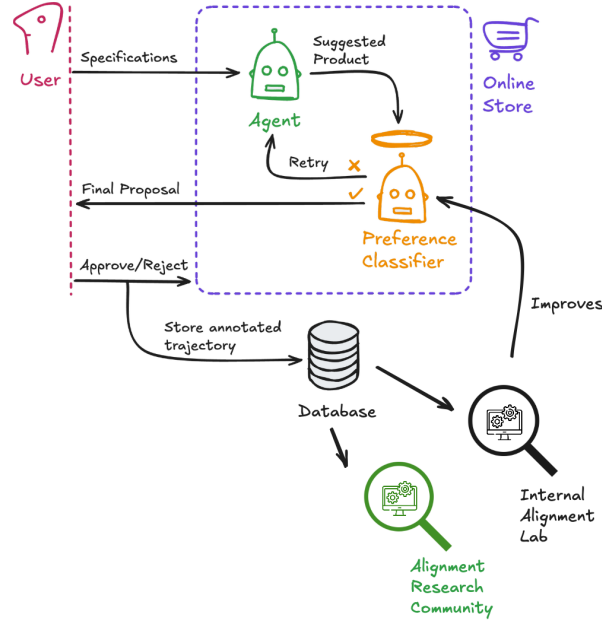


Figure 1: Our agent workflow in an early low-stakes e-commerce context

We propose a state of the art (SOTA) LLM assistant agent aimed at increasing the productivity of businesses operating in low to medium risk domains (such as e-commerce - see Figure 1 and the Demonstration attached to this document). We expect that even in low complexity settings agents will fail to align with user preferences either through their inability to capture the user’s full intent; or by intentionally trading off user preferences in way which is not desirable. To reduce the number of failures we will deploy the agent alongside a custom Preference Model. This model will guide the agent by monitoring its trajectory and making adjustments in real time when it believes that the agent has strayed from the user’s intent. When the agent makes a recommendation which, according to the Preference Model, meets all the user’s requirements, the user is asked to either accept or reject the recommendation. In this way, we aim to collect trajectories which are annotated based on their alignment to the users preference. This dataset will be used to train our custom preference model. Furthermore, the data will be made available to alignment researchers. We believe that leveraging a commercially viable product to facilitate the collection of diverse annotated trajectories could provide a dataset which is far more valuable to alignment research than those collected for RLHF.

Pacchiardi et al demonstrated a black-box classifier can detect model deception with high accuracy. Our Preference Model adapts this work by classifying misalignment between agent actions and user intent rather than deception. Even when misaligned, agents likely assign non-negligible probability to more user preferences, similar to Schaeffer et al.’s (2023) findings on emergent LLM properties. While this may not be self evident to human eyes, these tradeoffs could be detected through black-box classification methods.

Our approach assumes that detecting intent alignment failures in low-risk agentic settings will generalize to high-stakes domains. Our beachhead market will be an e-commerce recommendation agent, leveraging their large market share to validate our technology. Success will enable partnerships in progressively complex domains — moving from personal assistance and travel

planning to legal and business management — where richer trajectories offer deeper insights into agent misalignment.

Our competitive advantage is threefold: Unlike AI labs focused primarily on superintelligence development, we prioritize alignment dataset collection. We differentiate from other agent providers through deep expertise in ML and AI safety research, incorporating advanced alignment insights like Pacchiardi’s work—knowledge typically absent in engineering-led startups. Finally, we surpass existing alignment research groups by generating real-world agentic interaction data, addressing a critical gap in current alignment datasets.

4 Pilot experiment

Our experimental roadmap focuses on developing robust methods to differentiate between agent trajectories that align with user intent versus those that merely satisfy explicit prompts.

Phase 1: Dataset Creation and Validation

Construct a synthetic dataset using agentic benchmarks (e.g., WebArena). All tasks chosen will require intent inference (e.g. OneStopShop purchases, Github commit specifications). The solutions will be evaluated based on (i) prompt adherence; (ii) user intent alignment and (iii) hidden but human-inferable user requirements. Outcomes will be labelled by humans as either a (i) prompt failure; (ii) partial success (prompt-aligned only); or (iii) success (intent-aligned)

Phase 2: Training the preference model

Following Pacchiardi et al.’s methodology, we will train a linear regression classifier (low inference cost) on filtered results from the dataset developed in Phase 1 (excluding prompt failures). The classifier will focus on differentiating between prompt-aligned and intent-aligned solutions. After training we will also validate classifier performance across a diverse set of tasks. The resulting dataset enables both immediate commercial application and longer-term alignment research beyond our preference model.

Phase 3: Optimization and Refinement

Explore options to improve the preference model including varying the classifier architecture.

5 Process

Timeframe	Product development	Assisting alignment research
Next 3 months	Build MVP agentic web workflow and secure compute resources	Design and validate robust (90%+ accuracy) intent alignment classifier using synthetic data on WebArena tasks
2025	Partner with e-commerce; customer service; travel platforms to build SotA workflow dataset, integrating classifier input and growing dataset (>1bn datapoints)	Recruit 5 top researchers for internal alignment science team to apply insights from growing dataset and share initial dataset with external alignment researchers for advanced system alignment

Timeframe	Product development	Assisting alignment research
2026	Consolidate commercial partnerships to scale to trillions of yearly trajectories; establish leadership in low/medium-stakes domains (e.g. accountancy)	Begin fine-tuning open-source LLMs on alignment dataset to create more aligned systems
2027	Expand to high-stakes domains (healthcare, legal, infrastructure), building dataset of quadrillions of higher-quality labeled trajectories	Sell modified alignment dataset to AI labs to improve superintelligent agent alignment prospects

6 Impact on AI Safety & Key Risks

This startup addresses a critical gap in AI alignment research: a comprehensive, high-quality dataset for alignment experiments that surpasses current RLHF datasets for agentic systems. This focus is particularly significant given the anticipated development of superintelligent systems, where traditional capability evaluations or control mechanisms may no longer suffice as safety guarantees (Clymer et. al. 2024). At present, there exists a notable absence of large-scale human preference data collection in real-world agent interactions—a crucial requirement for superintelligence alignment research.

Our product delivers also immediate practical benefits by ensuring current-generation AI agents operate with improved alignment to user intent, thereby reducing the likelihood of alignment failures in pre-superintelligent systems. Additionally, the development of more robustly aligned agents creates market incentives that promote a "race to the top" in agent alignment quality.

Risk Analysis and Mitigation Strategies

1. **Insufficient Research Value:** The collected dataset of human feedback on agent interaction trajectories may not provide meaningful value for superintelligence alignment. *Mitigation:* enable early access to our dataset for alignment researchers to validate our approach, and build an internal alignment team to help structure our feedback mechanisms to optimise for alignment-relevant data collection.
2. **Accelerating AI timelines:** Enhanced economic viability of agents could accelerate AI development. However, given existing competition among AI labs for full automation of AI R&D, our agent's market entry is unlikely to significantly impact development timelines.
3. **Data Security:** Agent trajectories may contain personally identifying information, compromising product trust through data leaks. *Mitigation:* implement proven data protection protocols for data security and anonymisation, hiring relevant experts.

Industry Standards Alignment

While specific AI agent safety standards remain in development, we will align with established data protection and ethical collection frameworks, including ISO 19731: 2017 and EU GDPR.

Monitoring and Evaluation, Success Metrics and KPIs

Since our core product includes a dashboard, M&E need not be abstracted from our KPIs, which are:

1. **Size (in tokens) of the labelled agent trajectory dataset** Our strategy initially emphasises quantity of trajectories gathered, with numerous smaller datapoints in low-stakes domains such as e-commerce recommendation. As we deploy our agent in more complex contexts, we gather higher-quality data involving greater reflection and context. Targeting the token-level size of the dataset therefore serves as an effective metric for both quantitative scale and qualitative depth.
2. **User Satisfaction** Product adoption and scaling depends on maintaining high user satisfaction, measured by the percentage of trajectories where users are satisfied (e.g., successful product search, accurate document preparation).
3. **Quality of alignment research produced using our dataset** While challenging to quantify, this is the most impactful metric. We will assess our research impact through submissions to top-tier ML conferences that use our dataset, and may convene a panel of leading alignment researchers to give opinions on the product's contribution to alignment.

Bibliography

- Casper, Stephen et. al. 'Open Problems and Fundamental Limitaitons of Reinforcement Learning from Human Feedback'. 2023, <https://arxiv.org/abs/2307.15217>
- Pacchiardi, Lorenzo et. al. 'How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Qusetions', 2023. <https://arxiv.org/abs/2309.15840>
- Schaeffer, Rylan et. al. 2023, 'Are Emergent Abilities of Large Language Models a Mirage?', <https://arxiv.org/abs/2304.15004>
- Clymer, Joshua et. al. 2024, 'Safety Cases: How to Justify the Safety of Advanced AI Systems', <https://arxiv.org/abs/2403.10462>
- Grand View Research, 'AI Agents Market Size, Share Trends Analysis Report By Technology, By Agent System (Single-agent, Multi-agent), By Type, By Application (Customer Service Virtual Assistants, Healthcare), By End Use, By Region, And Segment Forecasts, 2024 - 2030', <https://www.grandviewresearch.com/industry-analysis/ai-agents-market-report>: :text=The

Appendix

A note on market research: A report by Grand View Research (2023) estimates that the global AI agents market was estimated at \$3.86 billion in 2023 and is predicted 45% year-on-year, reaching \$17 billion by 2027. While we acknowledge the issues in defining an 'AI agent' and the report's neglect of the recent jump in LLM capabilities, we think this provides a strong initial estimate for the value that could be captured by a preference-aligned AI agent company.

A company that captured 20% of the \$17bn AI agent market by 2027 would have revenues of \$3.4bn.

Acknowledgements:

We would like to thank Hannah Rose Kirk for her valuable insights.