AI Safety & Ethics Policy

Effective 19 November, 2024

Executive Summary

As artificial intelligence continues to revolutionize the financial sector, Telescope AI ("Telescope") is committed to harnessing its transformative potential while proactively addressing the challenges and risks it presents. We believe that responsible AI development is not just a regulatory requirement, but a fundamental ethical imperative that underpins our mission to enhance financial decision-making.

This AI Safety & Ethics Policy reflects our dedication to a proportional, iterative, and transparent approach to AI governance. By implementing safeguards that correspond to the level of risk posed by our AI systems, we aim to balance innovation with safety, ensuring that our technology remains at the forefront of the industry while prioritizing the well-being of our users and the broader financial ecosystem.

Our policy is designed to be flexible and adaptable, acknowledging the rapidly evolving nature of AI technology. We commit to regular assessments of our AI capabilities, continuous refinement of our risk mitigation strategies, and ongoing research into potential future challenges. This iterative process allows us to stay ahead of emerging risks while fostering responsible innovation.

By sharing our approach publicly, we aim to contribute to the broader dialogue on AI safety and ethics in finance. We believe that transparency and collaboration are essential for building trust and establishing industry-wide best practices. As we navigate the complexities of AI in finance, we remain open to feedback from stakeholders and are committed to working with regulators, industry partners, and the wider community to shape a future where AI enhances financial services responsibly and ethically.

Contents

Introduction

Telescope's AI Safety & Ethics Policy outlines our comprehensive approach to responsible AI development and deployment in the financial sector. This policy demonstrates our commitment to regulatory compliance, user safety, fairness, and transparency while leveraging the power of AI to transform financial decision-making.

Key components of our policy include:

- **Background:** Adherence to Australian, US, UK, and California AI Safety Standards, GDPR, CCPA, and model risk frameworks. Alignment with financial services regulations and reporting requirements.
- **Risk Assessment and Mitigation:** Pre-production validation and production monitoring. Implementation of incident response and recovery procedures.
- **Transparency and Explainability:** Consistent output generation using controlled parameters to maximize reliability. Explicit disclosure of AI system capabilities and limitations.
- **Fairness and Non-discrimination:** Design principles for unbiased prompt engineering. Regular monitoring of system outputs for fairness.
- **Privacy and Data Governance:** Comprehensive data protection and handling protocols for sensitive information. Robust data processing agreements and audit controls.
- **Technical Controls:** Model versioning and deployment management protocols. Documentation standards for development and changes.
- **Replicability and Consistency:** Implementation of controls to ensure consistent outputs for identical inputs.
- **Accountability and Human Oversight:** Dedicated AI Risk Management team with real-time intervention capabilities. Defined roles and responsibilities.
- **Continuous Improvement:** Regular policy reviews and structured process for implementing enhancements. System monitoring and feedback integration.

This policy serves as a living document, regularly updated to address emerging challenges and opportunities in AI technology. It reflects Telescope's dedication to ethical AI use in finance, emphasizing our commitment to responsible innovation, client trust, and regulatory compliance.

This policy is designed to evolve alongside developing industry standards and regulatory frameworks. We continuously monitor and incorporate emerging best practices in AI governance and financial services compliance.

We actively welcome feedback on our policy and suggestions for improvement. To submit your feedback or suggestions, please contact us at legal@telescope.co.

1. Background

This policy outlines Telescope's framework for safely implementing and deploying AI-powered financial services. While we don't develop foundation models ourselves, we recognize our critical role as an intermediary between AI capabilities and financial markets. Our approach focuses on responsible implementation, robust safeguards, and comprehensive risk management.

Telescope's AI Safety & Ethics Policy is designed to align with key international standards and regulations governing AI implementation in financial services:

Category	Jurisdiction	Standard/Regulation	Version/Date	Key Focus
Voluntary Guidelines	United States	AI Risk Management Framework (NIST)	1.0 (January 2023)	AI Risk Management
Policy Framework	United States	Blueprint for an AI Bill of Rights	October 2022	AI Rights Protection
Voluntary Guidelines	Australia	Voluntary AI Safety Standard	August 2024	Safe AI Development
Voluntary Guidelines	Australia	Guidance on privacy and the use of commercially available AI products	October 2024	AI Privacy Compliance
Voluntary Guidelines	Australia	A Director's Guide to AI Governance	2024	Board-level AI Governance
Policy Framework	United Kingdom	A pro-innovation approach to AI regulation	February 2024	AI Innovation Guidelines
Law	EU	General Data Protection Regulation (GDPR)	April 2016	Data Privacy Protection
Law	EU	The AI Act	2024/1689 (June 2024)	AI System Control
State Law	California	California Consumer Privacy Act (CCPA)	2018	Data Privacy Protection
State Law	California	AI Transparency Act	SB 942 (September 2024)	AI Use Disclosure*

Our AI Safety Framework consists of three core components:

- **Implementation Standards:** Technical and operational measures ensuring safe deployment of AI capabilities within our financial services platform. These standards govern how we integrate third-party AI models, implement our Guardrails system, and maintain consistent, reliable operations.
- **Protection Standards:** Security and privacy measures safeguarding our systems, user

data, and AI implementations from unauthorized access or compromise. These standards maintain the integrity of our services throughout their lifecycle.

- **Monitoring Standards:** Continuous oversight mechanisms ensuring our AI systems operate within defined parameters and ethical boundaries. These standards include real-time monitoring, regular audits, and systematic review processes.

The Guardrails system serves as our primary technical implementation of these standards, providing:

- Pre-execution screening of all inputs
- Real-time monitoring of system outputs
- Automated enforcement of safety policies
- Human oversight capabilities

Our risk management approach is proportional to the capability level and potential impact of our AI implementations. We employ a tiered system of safeguards that scales with:

- The sophistication of AI models we integrate
- The complexity of financial operations we support
- The scope of our user base
- The potential impact of system outputs

As AI capabilities advance, we continuously evaluate and upgrade our safeguards. This includes:

- Regular assessment of AI model capabilities
- Periodic review of risk thresholds
- Updates to our Guardrails policies
- Enhancement of monitoring systems
- Strengthening of protection measures

While we provide guidance and technical support for regulatory compliance, including the California AI Transparency Act's disclosure requirements, the responsibility for such disclosures is shared between Telescope and our clients in their respective markets. We provide guidance and encourage appropriate disclosures while respecting our clients' autonomy in managing their regulatory obligations.

This background sets the foundation for our detailed policies on risk assessment, transparency, fairness, privacy, technical controls, consistency, accountability, and continuous improvement, which are elaborated in subsequent sections.

2. Risk Assessment and Mitigation

Our risk assessment and mitigation framework is built on a comprehensive understanding of the unique challenges posed by implementing Large Language Models (LLMs) in financial contexts.

We employ a multi-layered approach to identify, assess, and mitigate risks while maintaining operational efficiency.

2.1 Risk Assessment Framework

We categorize risks into three primary domains:

1. **Implementation Risks**
    - Model selection and behavior unpredictability
    - Parameter configuration (temperature, tokens, context)
    - Output accuracy and reliability
    - Data quality and completeness
    - Potential for misinformation
    - Hallucination risks
    - Source reliability and bias

2. **User Interaction Risks**
    - Potential for misinterpretation of outputs
    - Risk of users seeking or receiving financial advice
    - Potential for poor quality, offensive or factually incorrect outputs.
    - Persuasiveness of LLM-generated content
    - Information accessibility and comprehension
    - User privacy and data protection

3. **Market Risks**
    - Potential for market manipulation
    - Systemic risk considerations

2.2 Mitigation Strategies

Our mitigation approach is built on a comprehensive risk management framework underpinned by our Guardrails system, which implements five key policies:

1. **Personal Advice Policy:** Prevents generation of personalized financial recommendations. Screens for and blocks inputs containing personal circumstances. Maintains clear separation from regulated financial advice.

2. **Subjective Advice Policy:** Eliminates opinion-based investment suggestions. Removes qualitative judgments about market performance. Focuses on factual, data-driven insights. Maintains neutrality in market analysis.

3. **Moderation Policy:** Prevents harmful or inappropriate content. Maintains professional communication standards. Protects user safety and platform integrity.

4. **Nonsensical Input Policy:** Filters irrelevant or unintelligible inputs. Protects against system manipulation. Ensures meaningful output generation.

5. **Personal Information Input Policy:** Implements the identification of personal information and identifiers.

2.3 Operational Controls

We maintain robust operational controls through:

1. **Real-time Monitoring:** Continuous system surveillance. Performance metric tracking. Anomaly detection alarms.

2. **Response Protocols:** Prioritized incident response based on severity. Defined escalation procedures. Clear incident response workflows.

3. **Model Management:** Regular model benchmarking performance assessments. Version control and documentation. Integration testing protocols. Backup model availability.

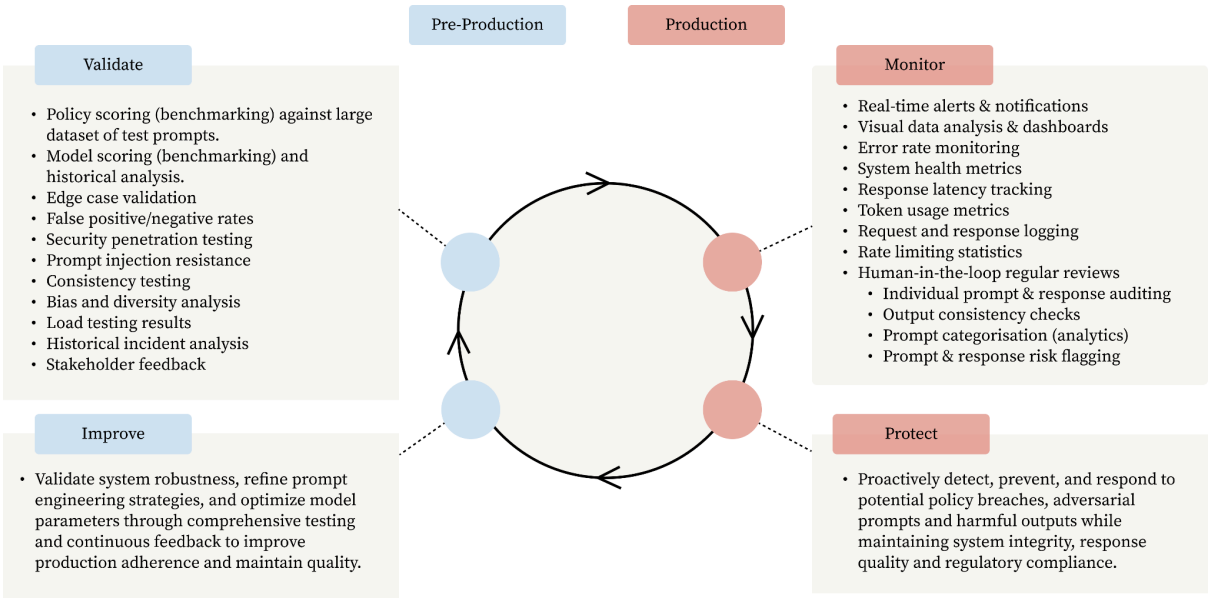
4. **Oversight:** We employ a tiered monitoring system, dependent on risk level, operational complexity, and stakeholder requirements:

   Level 1: Standard Operating Conditions
   - Regular monitoring and routine adjustments
   - Standard review procedures
   - Normal operational protocols

   Level 2: Enhanced Oversight
   - Typically employed for new engagements (e.g., new client onboarding or Phase 1 product launches)
   - Increased monitoring frequency
   - Additional human review requirements
   - Heightened alert sensitivity

   Level 3: Urgent Oversight
   - Immediate human intervention

- Urgent patch fixes such as manual prompt intervention.
- System operation restrictions
- Stakeholder notifications



2.4 Continuous Assessment

Our risk assessment process is ongoing and includes:

- Regular policy reviews and updates
- Performance metric analysis
- User feedback integration
- Incident response evaluation
- Regulatory compliance monitoring
- Technology advancement assessment

This framework ensures we maintain robust risk management while delivering innovative financial technology solutions. Our approach emphasizes prevention while maintaining readiness for rapid response to emerging challenges.

3. Transparency and Explainability

Transparency and explainability are cornerstone principles in the ethical deployment of AI systems within the financial sector. They are vital for building user trust, ensuring regulatory compliance, and effectively managing risks associated with AI technologies. By making AI processes and decisions understandable, we reduce the risk of unintended outcomes and empower users to make informed decisions based on clear insights.

3.1 Disclosure to End-users

Telescope encourages our clients to inform end-users when AI systems are involved in decision-making processes or content generation. This transparency fosters trust and allows users to be aware of how their information is processed.

Regulatory Considerations:

- **California AI Transparency Act:** Mandates the disclosure of AI-generated content to end-users within California.
- **Australian AI Safety Guidelines:** Recommend informing end-users about AI-enabled decisions to promote transparency and accountability.

Telescope's Guidance:

- **Encouraging Best Practices:** While disclosure is not universally required, we advise clients to be transparent about AI involvement, aligning with global best practices.
- **Product-Specific Requirements:** For our product Ripple, we require clients to disclose the use of AI. Given that Ripple provides an end-user with search capabilities across financial instruments — which is a domain considered high-risk by some AI providers [1] — this disclosure is a precautionary measure to ensure compliance and ethical integrity.
- **Guidelines for Effective Disclosure:** We encourage our clients to craft clear and effective disclosure statements.

  The following key points should be covered:

  - ☐ Clearly state that AI is used for content generation or decision-making.
  - ☐ Clarify that the information provided is factual but not intended as advice.

[1] Anthropic's Acceptable Usage Policy, Effective 6 June 2024 https://www.anthropic.com/legal/aup

- [ ] Advise end-users to verify AI-generated information before making important decisions.
- [ ] Disclose the possibility of AI inaccuracies or outdated data.
- [ ] Clearly state that the tool must not be used for any inappropriate or harmful requests, including but not limited to hate speech, violence, harassment, market manipulation, discriminatory behavior, spreading false or misleading information, the production of immoral content or engaging in any illegal activities or actions that breach regulatory compliance.

3.2 Providing Rationales for AI Decisions

To enhance explainability, Telescope ensures that AI-generated outputs include clear rationales for decisions.

Telescope's Practices:

- **Include Explanations:** AI outputs should provide reasoning behind investment ideas or analyses.
- **Clear Comprehension:** Align all outputs and use clear language to help users understand the factors influencing the AI's suggestions.

3.3 Disclosure of Source Documents

Transparency is reinforced by providing access to the sources underpinning AI-generated content. The approach to citations and source disclosure varies depending on the product and context.

Telescope leverages the LLMs' foundational knowledge and reasoning capabilities. The use of external data sources, where appropriate, is product dependent.

Framework for LLM-Only Approach:

Suitable Use Cases:

- Thematic classification and categorization
- General industry knowledge and trends
- Conceptual explanations and definitions
- Pattern recognition across broad topics
- Logical reasoning tasks
- Synthesis of widely-known information

Unsuitable Use Cases:

- Time-sensitive information
- Specific numerical or pricing data
- Company-specific facts
- Current market conditions
- Financial advice

3.3.1 Relying on the LLM-Only

Explainability is crucial in helping users understand and trust AI-generated outputs. Where possible, Telescope is committed to ensuring that the reasoning behind AI decisions is transparent, unbiased, and grounded in the AI's foundational knowledge.

For suitable use cases outlined above, Telescope leverages the inherent reasoning capabilities of LLMs. When users interact with our AI systems for tasks like thematic classification (e.g., “electric vehicles”) our systems typically rely on the LLMs’ foundational knowledge rather than real-time or external data sources.

Telescope’s Principles:

1. **Foundation Trust:** We trust in the LLM's extensive training on diverse and comprehensive datasets to deliver accurate and rational outputs without the need for external prompts or biases.

2. **Unbiased Responses:** By avoiding the imposition of additional context or influence, we ensure that the AI's outputs remain unbiased and free from collusion, providing genuine insights based on its internal reasoning.

Telescope prioritizes the delivery of information that is both rational and free from interference. By capitalizing on the LLM's foundational strengths, we see the following benefits:

- **Enhanced Reliability:** Relying on the LLM's inherent capabilities reduces the risk of introducing external biases or errors that could arise from additional inputs.

- **Future-Proofing:** As LLMs continue to improve and expand their knowledge bases, this approach ensures that the AI's outputs evolve naturally, keeping pace with the latest information without manual intervention.

- **User Trust:** Providing responses grounded in the AI's well-established knowledge fosters trust among users who can rely on consistent and rational outputs.

3.3.2 Relying on the LLM with verified Data Sources

For use cases requiring up-to-date or company-specific information, as outlined in Section 3.3, Telescope supplements the LLM's foundational knowledge with verified data sources to ensure accuracy and reliability.

Telescope's Practices:

- **No Unverified Web Lookups:** Avoid using unverified web lookup APIs within AI systems to prevent the introduction of bias and misinformation.
- **Whitelist Approved Sources:** Any external data retrieval must be from pre-approved, trusted sources.

Telescope's Principles:

- **Rely on Authoritative Sources:** Use official financial documents and validated data sets for any context alteration.
- **Prevent Bias and Misinformation:** By controlling data sources, maintain the reliability and accuracy of AI outputs.

3.3.3 Use of Citations When Context Is Altered

When addressing unsuitable LLM-only use cases (such as time-sensitive information and company-specific facts), Telescope's products like Confer incorporate external data sources to augment the AI's responses. In these instances, proper citation and source attribution become essential for maintaining transparency and trust.

This approach helps users assess the credibility of the information and make more informed choices.

Telescope's Practices:

1. **Include Citations:** Provide references to official documents, financial reports, or authoritative sources used in generating responses via a citations feature.
2. **Include Data Sources:** Where possible, make data sources also available to end-users, enabling them to verify information, confirm the citations and understand the context. In some instances licensing issues may limit our ability to provide an end-user with access to documents in their entirety.

3.4 Acknowledging Limitations Due to Knowledge Cutoff

While we strive for transparency and explainability in our AI outputs, it's important to note the inherent limitations of AI models' knowledge cutoff dates. The LLMs we utilize have been trained on data up to specific points in time.

- **Outdated Information:** The AI may not be aware of events, developments, or data that emerged after its cutoff date, potentially affecting the relevance or accuracy of its outputs.

- **Explainability Constraints:** When providing rationales for decisions, the AI's explanations are limited to its existing knowledge base and may not account for the most recent information or context.

Examples:

- **Recent Market Events:** If a company has recently gone public (IPO), merged, or delisted, the AI may not have information about these events. For instance, queries like "Tell me about companies that IPO'd today" or "Which companies have delisted recently?" would not yield accurate or up-to-date responses.
- **Current Market Conditions:** Questions such as "Which stocks went up in June 2024?" cannot be accurately addressed by the AI if its knowledge cutoff predates June 2024. The AI lacks awareness of market movements, economic indicators, or geopolitical events that occurred after its last update.

Telescope's Position:

- **Acceptance of Limitations:** We acknowledge these limitations as an inherent aspect of current AI technology and factor them into our risk assessment and product design.
- **Client Guidance:** We advise our clients to articulate these limitations to end-users, ensuring that users are aware of the potential gaps in the AI's knowledge, as per 3.1 Disclosure to End-Users.
- **User Awareness:** By informing end-users, we promote transparency and enable users to consider these limitations when interpreting AI-generated insights, thereby supporting more informed decision-making.
- **Alternative Solutions:** Where up-to-date information is critical, we implement as follows:
    - Additional data sources or systems that can provide real-time data, ensuring compliance with our policies regarding data verification and bias avoidance (as per 3.3).
    - Preventative measures: our system is aware of its limitations, automatically rejects unsuitable requests, and provides clear explanations for each decision.
- **Implement Model Updates:** We are committed to leveraging advancements in AI to provide more current and accurate information as models evolve and training cycles become more frequent. As AI models continue to evolve, the time between training and deployment is decreasing. For example, the knowledge cutoff has reduced from approximately 2,160 days with earlier models like OpenAI's GPT-2 to about 168 days with more recent models like Anthropic's Claude 3.5 at the time of their respective launches.

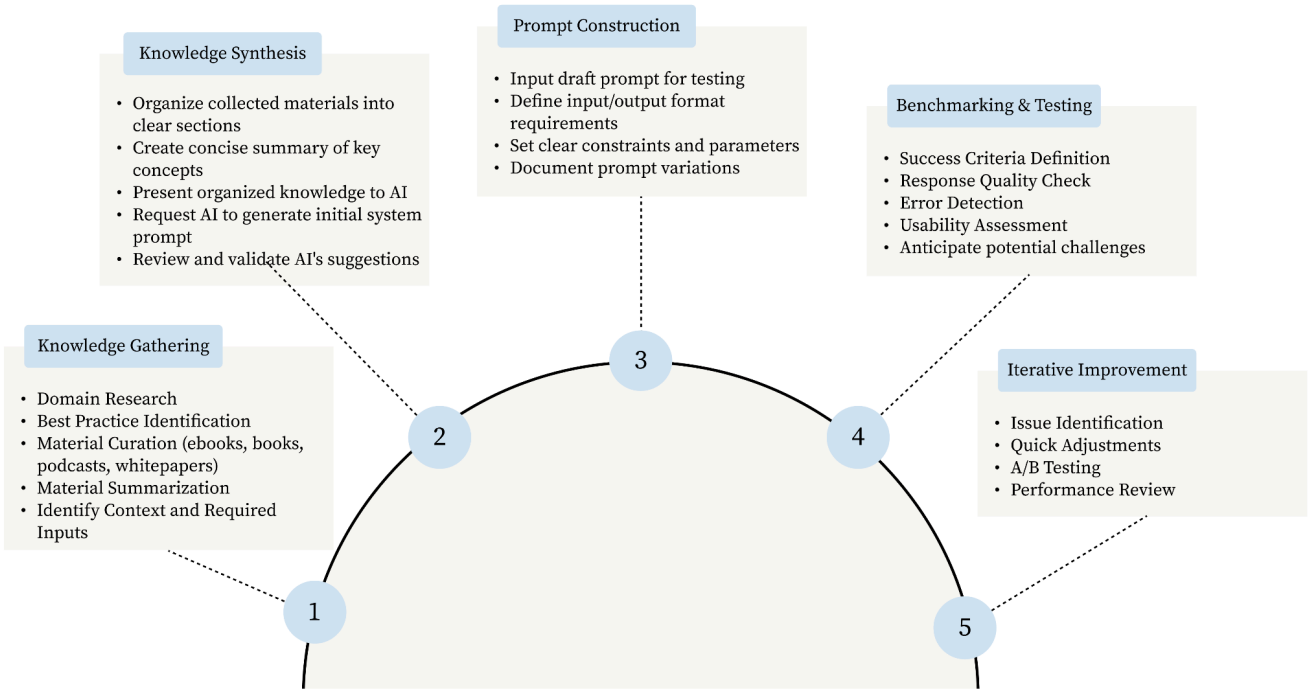
3.5 Knowledge-Sourced System Prompt Construction

System prompts serve as the foundational architecture of our AI platform, acting as the critical interface between raw computational power and meaningful analytical output. The careful engineering of these prompts - their structure, context, and constraints - determines the quality and reliability of AI-generated insights. This engineering process requires deep understanding of both the technical capabilities of language models and the specific requirements of complex analytical tasks.

The construction of system prompts depends on the specific product or service being analyzed. Different products require different knowledge sources to generate accurate and relevant outputs. In the instance of complex analysis systems, the knowledge base must encompass statistical models, decision frameworks, and domain-specific expertise. For automated decision-making applications, the system draws from operational research, probability theory, and industry-specific performance metrics.

Methodology:

- **Curated Libraries:** We use libraries of verified financial literature and documents to inform prompt construction, grounding AI reasoning in established information.
- **Accuracy Assurance:** This approach ensures that AI outputs are based on accurate, up-to-date, and relevant data.
- **Bias Reduction:** By carefully selecting source materials, we reduce the potential for bias in AI-generated content.



3.6 Summary of Transparency Commitments

By embedding transparency and explainability into our AI systems and practices, Telescope enhances trust and accountability in the financial services we provide. These measures support responsible innovation, ensure regulatory compliance, and contribute to the ethical use of AI, ultimately empowering users and strengthening the integrity of the financial sector.

4. Fairness and Non-discrimination

4.1 Commitment to Fairness

We are dedicated to:

- **Equitable Access:** Providing equal opportunities for all users to benefit from our AI services, regardless of their background or characteristics.
- **Non-discriminatory Practices:** Ensuring that our AI systems do not discriminate based on race, gender, age, religion, nationality, disability, or any other protected attribute.
- **Inclusive Design:** Developing our platforms and services with diversity and inclusivity in mind to serve a broad range of users effectively.

4.2 Bias Monitoring and Mitigation

While we utilize third-party LLMs that include their own fairness mechanisms, we acknowledge that AI outputs could potentially exhibit bias or discrimination. To address this, we have implemented the following measures:

- **Guardrails System:** Our Guardrails system (as detailed in Section 2) includes policies and technical controls designed to prevent the generation of biased or discriminatory content, based purely on user input.
- **Output Review Process:** Establishing a systematic review process to monitor AI outputs for potential bias or discrimination, including regular audits and human oversight by our AI Risk Management team.
- **Bias Mitigation Techniques:** Implementing strategies during prompt construction and benchmarking to reduce the likelihood of biased outputs. This includes a catalog of known contentious prompts.

4.3 Accessibility and Compliance

Ensuring that our services are accessible to all users is a key aspect of our commitment to fairness:

- **Accessibility Standards:** Offering our whitelabel solutions with accessibility compliance up to the "AA" level of the European standard (e.g., EN 301 549 and WCAG 2.1 AA[2]), ensuring usability for people with varied needs.
- **Inclusive Features:** Incorporating features that enhance accessibility, such as compatibility with screen readers, alternative text for images, and adjustable text sizes.
- **Regulatory Compliance:** Aligning our practices with accessibility regulations and guidelines in the jurisdictions where we operate.

4.4 Continuous Improvement and Client Engagement

Achieving fairness is an ongoing process that benefits from collaboration with our clients:

- **Feedback Channels:** Providing clear channels for clients to report any instances of biased or discriminatory outputs.
- **Responsive Action:** Committing to promptly investigate and rectify any reported issues related to fairness or discrimination.

[2] Accessibility Compliance Standard, EN 301 549 and WCAG 2.1 AA https://www.w3.org/WAI/WCAG22/quickref/

- **Stakeholder Engagement:** Engaging with clients, users, and advocacy groups to gather diverse perspectives on fairness issues.

5. Privacy and Data Governance

Telescope is dedicated to protecting the privacy of all users interacting with our AI systems. We recognize the importance of safeguarding personal information and are committed to maintaining compliance with global data protection regulations, including GDPR, CCPA, and other relevant laws.

5.2 User Identification and Anonymity

- **Optional User-ID Tracking:** Our platform supports the use of 'user-id' values and session tokens to enhance user experience and provide logging services. For example, this capability allows a client to link inference responses with customer activity, should they have an interest in that dependency. However, the use of such identifiers is entirely optional and not required for the operation of our services.
- **Privacy Considerations:** For clients operating in jurisdictions with stringent privacy regulations or for those who prioritize user anonymity, we discourage the use of user-id tracking. This ensures users can interact with our services without disclosing personal identifiers.

5.3 Handling of Personal Information (PI) Inputs

- **PI Detection and Prevention:** Our Guardrails system is designed to detect and prevent the processing of personal information (PI). It is trained to identify common PI inputs such as email addresses, credit card numbers, and personal identifiers with a high degree of accuracy.
- **Limitations:** While our system effectively detects obvious PI inputs, it may not identify less structured personal data such as home addresses, personal names, or mobile phone numbers. We acknowledge these limitations and are committed to continuous improvement.

5.4 Data Retention and Deletion

- **Data Minimization:** We adhere to the principle of data minimization by collecting and processing only the data necessary for the provision of our services, reducing risks associated with data breaches and unauthorized access.

- **Data Scrubbing:** In instances where PI is detected in user inputs, we commit to scrubbing the relevant data and associated log files within 7 days. This retention period allows us to conduct a read-only review of the input prompts to enhance our PI detection methods.
- **Continuous Improvement:** Feedback from these reviews is used solely to improve our systems' ability to detect and prevent the processing of PI, strengthening our privacy safeguards.

5.5 Access and Incident Response

- **Access Controls:** Strict access controls are in place to ensure that only authorized personnel can access sensitive data.
- **Incident Response:** In the event of a data breach or security incident, we have established protocols for prompt response, mitigation, and notification in accordance with regulatory requirements.

5.6 Compliance with Data Protection Regulations

- **Regulatory Adherence:** We are committed to complying with all applicable data protection laws and regulations in the jurisdictions where we operate, including the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA).
- **Third-Party DPAs:** We have Data Processing Agreements (DPAs) in place with our third-party Large Language Model (LLM) providers. These agreements outline the responsibilities and obligations of each party in handling personal data, ensuring that data is processed in compliance with applicable laws. We can provide copies of these DPAs to clients upon request.

5.7 Transparency and User Rights

- **User Awareness:** We encourage transparency with end-users regarding the handling of their data. Clients are advised to inform users about data processing activities, especially when personal data may be involved.
- **Data Subject Rights:** We respect the rights of data subjects under applicable laws, including rights of access, rectification, erasure, and objection. Processes are in place to assist clients in responding to data subject requests in a timely manner.

5.8 Data Governance Practices

- **Training and Awareness:** Regular training is provided to staff to ensure they are aware of data protection obligations and best practices in handling sensitive information.
- **Third-Party Management:** We assess and monitor third-party service providers to ensure they meet our data protection standards, especially when they process data on our behalf.

6. Technical Controls

Technical controls are the backbone of our AI Safety Framework, ensuring that our AI systems operate securely, reliably, and in compliance with all relevant regulations. These controls encompass the technical measures, protocols, and practices that safeguard our AI models throughout their lifecycle, from development to deployment and ongoing operation.

6.1 Model Versioning and Deployment Management

We employ rigorous model versioning and deployment management protocols to maintain control over our AI models and ensure their integrity.

- **Version Control Systems:** All AI model configuration and critical prompt templates in use are managed using version control systems that track changes, updates, and modifications. This allows for precise tracking of model evolution and facilitates rollback if necessary.
- **Controlled Deployment Pipelines:** Deployments of prompts and model changes follow a pipeline that includes stages for development & testing, staging and production. Each stage requires specific validations before progression.
- **Access Control and Permissions:** Deployment and modification of AI models are restricted to authorized personnel.
- **Rollback Mechanisms:** In the event of unforeseen issues post-deployment, we have established procedures to roll back to previous stable versions, minimizing downtime and potential impact on users.

6.2 Secure Development Practices

Our systems are developed following industry best practices for security.

- **Code Reviews:** All code related to AI models and systems undergoes peer review to identify and mitigate potential security vulnerabilities.
- **Secure Coding Standards:** Developers adhere to secure coding guidelines that prevent common vulnerabilities such as injection attacks, buffer overflows, and improper error handling.
- **Automated Testing: C**ontinuous integration and continuous deployment (CI/CD) pipelines include automated tests for security, functionality, and performance.

6.3 Encryption and Data Security

Data protection is paramount in our technical controls framework.

- **Data Encryption:** All sensitive data, both at rest and in transit, is encrypted using industry-standard encryption algorithms.
- **Secure Communication Protocols:** Communication between system components utilizes secure protocols (e.g., HTTPS, TLS) to prevent interception and tampering.
- **Key Management:** Encryption keys are managed securely, with strict controls over access and rotation policies.

6.4 Monitoring and Logging

Continuous monitoring and detailed logging are critical for maintaining system health and facilitating incident response.

- **Real-Time Monitoring:** System performance, usage metrics, and security events are monitored in real-time. This includes monitoring for unusual patterns that may indicate security threats or system malfunctions.
- **Comprehensive Logging:** All significant events, including user actions, system changes, and security incidents, are logged with timestamps and relevant details.
- **Anomaly Detection:** Automated tools analyze logs and metrics to detect anomalies that could signify potential issues or breaches.

6.5 Disaster Recovery and Business Continuity

We have established technical measures to ensure system resilience and rapid recovery from disruptions.

- **Redundancy:** Critical systems have redundant components and failover mechanisms to maintain availability in case of hardware or software failures.

- **Regular Backups:** All data is backed up regularly, with backups stored securely and tested for integrity.
- **Recovery Procedures:** Detailed disaster recovery plans outline the steps for restoring systems and data in various scenarios, minimizing downtime.

6.6 Security Testing and Vulnerability Management

We proactively identify and address security vulnerabilities in our AI systems.

- **Penetration Testing:** Regular penetration tests are conducted to assess the security of our systems from an attacker's perspective.
- **Vulnerability Scanning:** Automated tools scan for known vulnerabilities in software components and dependencies.
- **Patch Management:** Security patches and updates are applied promptly following a risk-based approach.

6.7 Third-Party Integration Controls

When integrating third-party AI models or services, we implement additional controls to manage associated risks.

- **Vendor Assessment:** Third-party providers are evaluated for security, compliance, and reliability before integration.
- **Secure APIs and Interfaces:** Interactions with third-party services utilize secure APIs with authentication and encryption.
- **Data Handling Agreements:** Contracts with third-party providers include data protection clauses and adherence to our data governance standards.

7. Replicability and Consistency

In the financial sector, the replicability and consistency of AI-generated outputs are critical for maintaining trust, ensuring compliance, and supporting reliable decision-making. Users and stakeholders expect that identical inputs will yield consistent outputs, enabling them to make informed and confident choices.

However, achieving perfect replicability with LLMs presents inherent challenges due to their probabilistic nature. Despite these challenges, Telescope is committed to implementing strategies that enhance the consistency of our AI systems to the highest degree possible.

7.1. Challenges with LLMs

LLMs generate responses based on complex probabilistic algorithms, which can lead to variability in outputs even when the inputs remain the same. Factors contributing to this variability include:

- **Randomness in Sampling:** The inherent randomness in how the model selects from possible outputs.
- **Model Updates:** Changes or improvements made by LLM providers that alter model behavior.
- **System Prompt Adjustments:** Modifications to prompts or guardrails that influence the AI's responses.
- **Temperature Settings:** Parameters that control the creativity and randomness of outputs.
- **Inference Refusal:** Misaligned AI safety policies, rejecting responses when a rejection is unexpected.

7.2. Mitigation Strategies - Determinism

While achieving absolute determinism with current AI technologies is not possible due to the inherent randomness in Large Language Models (LLMs), we employ several strategies to enhance the replicability and consistency of our AI outputs.

LLMs generate responses based on probabilistic algorithms, introducing a degree of randomness even with identical inputs. This randomness arises from the model's base response mechanisms, which select from a range of possible outputs. To mitigate this variability and enhance consistency, we implement the following measures:

- **Controlling Randomness:** By configuring the AI models to minimize randomness in their output generation process, we ensure more consistent and predictable responses.
- **Control Variability via Internal Mechanisms:** We follow documented guidelines from the LLMs to implement internal mechanisms that help control variability, enhancing the consistency of outputs without compromising the quality of responses.

7.2.1 Benchmarking Results

We conducted benchmarking tests using **GPT-4o** with 1,000 iterations to evaluate the effectiveness of our consistency measures. The tests focused on assessing the AI's consistency in different scenarios relevant to our financial applications.

Test 1: Random S&P 500 Selection:

In this test, the AI is prompted to select a random company from the S&P 500 index. Due to the inherent randomness and large selection pool, this test evaluates the AI's ability to provide consistent responses when randomness is expected to be high. We are testing sparse responses versus predictable responses, focusing on rationality under conditions of high variability.

Temperature	Base Response Consistency	System-Enforced Consistency	Notes
0.0	100.00%	100.00%	← Telescope's Configuration
0.5	98.60%	100.00%	
1.0	81.90%	100.00%	
1.5	45.80%	100.00%	

Test 2: Duopoly Selection:

In this test, the AI is prompted to output a single company from a known duopoly pair within the, specifically Boeing and Airbus. These companies represent an almost perfect duopoly with market shares of approximately 49% and 50% respectively in the large commercial aircraft market. This test assesses the AI's consistency in decision-making when presented with two nearly identical options in terms of market dominance.

Temperature	Base Response Consistency	System-Enforced Consistency	Notes
0.0	99.90%	100.00%	← Telescope's Configuration
0.5	82.70%	96.80%	
1.0	56.80%	90.50%	
1.5	41.94%	81.90%	

These benchmarks reveal that our additional deterministic measures significantly improve consistency, particularly in scenarios where base response consistency diminishes at higher temperatures. At our standard configuration, we achieve a System-Enforced Consistency of

100.00% in the Duopoly Selection test, indicating that nearly all outputs are consistent when the same input is provided.

7.2.2 Implementation in Our Products

- **Ripple:** We have implemented these deterministic measures in our product Ripple, specifically in the Thesis Generation and Instrument Selection components. This ensures that users receive consistent recommendations and analyses when providing the same input prompts.
- **Expansion to Other Products:** We are actively engaged with our clients to introduce similar consistency-enhancing measures into our other products, aiming to improve replicability across our entire suite of AI services.

Implementation Considerations:

- **Limitations:** While our measures enhance consistency, they are not foolproof. Factors such as model updates from LLM providers may still affect output patterns over time.
- **Technical Complexity:** We aim to abstract these technical details away from the end-users to provide a seamless experience while maintaining high levels of consistency.
- **Client Collaboration:** Implementing these measures in other products involves customizing solutions to meet specific client needs and use cases.

7.3. Mitigation Strategies - Caching Requests

If determinism is a further concern, we propose clients can use a caching mechanism implemented within their API gateway:

- **Implementation Caching:** Storing Telescope API responses allows repeated inputs to receive the same outputs without reprocessing.
- **Time Constraints:** Recognizing that models and prompts evolve and guardrail policies improve, cached responses should be time-limited to ensure they remain relevant and accurate. We would suggest a cache timeout of no more than 24 hours.

7.3. LLM Inference Logging

We maintain logs of requests and responses to support reproducibility:

- **Audit Trails:** Detailed records enable us to trace outputs back to their inputs, specific models and system prompts in use, and understand how responses were generated.

- **Issue Resolution:** Logs assist in diagnosing and resolving inconsistencies or errors that may arise.

7.4. Limitations and Commitment to Consistency

Despite our efforts to enhance replicability and consistency, certain limitations affect the predictability of AI-generated outputs:

- **Model Evolution:** Updates from LLM providers may change output patterns, affecting consistency over time.
- **System Modifications:** Changes to system prompts or guardrails can alter responses.
- **External Factors:** Variability in external data sources or dependencies may impact outputs.

We continuously monitor these factors and adjust our strategies to mitigate their effects on replicability. Our methods, as detailed in **Section 2.3 Operational Controls**, include iterative testing of new model variants with a focus on output consistency.

Telescope is dedicated to providing consistent and reliable AI services by:

- **Continuous Monitoring:** Regularly evaluating AI performance to detect and address inconsistencies.
- **Client Collaboration:** Working closely with clients to understand their needs and gather feedback on AI outputs.
- **Transparent Communication:** Informing clients of significant changes that may affect output consistency.

By proactively managing the factors influencing replicability and consistency, we aim to uphold the high standards expected in the financial industry and support our clients in delivering trustworthy services to their end-users.

8. Accountability and Human Oversight

Telescope recognizes the critical importance of maintaining robust human oversight and clear lines of accountability in the operation and deployment of AI-driven financial technologies. We are committed to ensuring that our systems remain under effective human control, allowing for prompt and appropriate intervention when necessary.

8.1 AI Risk Officer

We have appointed an AI Risk Officer who is responsible for overseeing the safety and ethical deployment of our AI systems. The responsibilities of this role include:

1. **Continuous Monitoring:** Overseeing AI system operations in real-time to identify and address issues as they arise. This includes monitoring for deviations from expected behavior and ensuring alignment with safety protocols.

2. **Review and Assessment:** Conducting regular audits of AI system outputs and performance metrics to maintain compliance with ethical guidelines and regulatory standards. This involves assessing the effectiveness of our Guardrails system and making adjustments as necessary.

3. **Flagging and Intervention:** Managing a system that flags prompts for human intervention. When flagged, these queries are placed into a prioritized review pipeline, ensuring that high-risk or ambiguous cases are examined thoroughly before any further action.

The AI Risk Officer works in coordination with other teams to uphold Telescope's ethical commitments, maintaining a proactive approach to risk management and ensuring transparency in our processes.

8.2 Human Oversight Mechanisms

To safeguard against potential risks and ensure responsible use, we have established structured human oversight mechanisms aligned with our tiered risk levels outlined in **Section 2.3 Operational Controls:**

1. **Level 1: Standard Oversight:** Under standard operating conditions, human oversight is provided through routine monitoring and the ability to flag and review prompts when necessary. The system operates with low-intensity human intervention, primarily focused on compliance and anomaly detection.

2. **Level 2: Enhanced Oversight:** When elevated risk levels are detected, such as in cases of significant output deviations or potential regulatory implications, oversight is intensified. Human reviewers are actively involved in auditing outputs and intervening when Guardrails may be insufficient.

3. **Level 3: Critical Response:** In high-risk scenarios, real-time human intervention is required to manage critical issues. This includes the immediate review of flagged prompts and the activation of more stringent monitoring measures. Refer to **Section 2.3** for a detailed description of our tiered risk response framework.

8.3 Human-in-the-Loop Interventions

Our **Human-in-the-Loop (HITL)** interventions are integrated into our risk management strategy to ensure that AI outputs, especially those related to financial insights or high-stakes decisions, are reviewed by qualified human professionals. This approach provides an essential layer of oversight, allowing for manual adjustments and the prevention of unintended consequences.

- **Risk-Based Intervention:** The level of human intervention is proportional to the identified risk level. At **Level 1**, interventions are minimal and routine. At **Level 2**, human reviewers are more engaged, and at **Level 3**, human oversight is comprehensive and immediate.
- **Flagging System:** Our system allows for the automatic flagging of prompts that require human review. These flagged cases enter a prioritized queue, where reviewers assess the content and determine the appropriate action. This ensures that critical cases are handled efficiently and with the necessary attention.

8.4 Continuous Improvement and Transparency

We are committed to refining our oversight mechanisms through continuous feedback and improvement:

1. **Performance Audits:** Regular audits help identify areas for improvement and ensure our oversight practices remain effective and aligned with evolving AI capabilities.
2. **Feedback Integration:** We actively collect and analyze feedback from clients and internal teams to enhance our oversight mechanisms and adapt to new challenges.
3. **Transparency:** We document significant decisions and actions taken by the AI Risk Officer, providing transparency to stakeholders and maintaining accountability for the oversight of our AI systems.

By embedding human oversight into our AI operations and maintaining a focus on continuous improvement, Telescope ensures that our technology operates safely and ethically, supporting responsible innovation and the trust of our clients.

Conclusion

Telescope is committed to responsibly advancing AI in finance. Recognizing the fast-paced evolution of technology and regulations, we will continuously improve our safety and ethics practices. We are excited about the opportunities ahead and will drive innovation while safeguarding our users and the financial ecosystem.

Appendix

Appendix A: Glossary

Term	Definition
Anonymization	The process of removing personally identifiable information from data sets so that individuals cannot be readily identified. This enhances privacy protection and compliance with data protection laws like GDPR and CCPA.
API (Application Programming Interface)	A set of protocols and tools for building software and applications.
Benchmarking	The practice of testing and comparing the performance of models or systems against standard metrics or datasets.
CCPA (California Consumer Privacy Act)	A state-wide data privacy law in California that provides consumers with rights regarding the collection, use, and sharing of their personal information by businesses.
Clients	Organizations, typically stock brokers or financial institutions, that utilize Telescope's API or whitelabel integrations to offer financial services to their end-users.
Data Minimization	A principle that encourages organizations to collect and process only the personal data that is necessary for a specific purpose.
Data Processing Agreement (DPA)	A legally binding document between a data controller and a data processor, outlining each party's responsibilities in handling personal data in compliance with applicable laws like GDPR.
Data Scrubbing	The process of cleaning data to remove or mask personal information and correct or remove inaccurate records.
Deployment Management	The process of controlling the release and integration of models into production environments. This includes using controlled pipelines, access controls, and rollback mechanisms to ensure secure and reliable operations.
Determinism	The property that ensures identical inputs to a system will always produce identical outputs, with no randomness involved.
Embedding Models	Models that convert data (such as words or phrases) into numerical vectors in a high-dimensional space, capturing semantic meanings and relationships.
End-users	The individuals who interact with the services provided by Telescope through its clients. They are typically customers of financial institutions using AI-generated insights or tools to inform their investment decisions.
Explainability	Explainability in the context of AI refers to the ability to make the decision-making processes and outputs of AI systems, including large language models (LLMs), transparent, understandable, and interpretable for users. This involves clarifying how data inputs are transformed into outputs, the logic underlying model decisions, and any associated risks or limitations, especially in complex use cases.
GDPR (General Data Protection Regulation)	A comprehensive data protection law in the European Union that sets guidelines for the collection and processing of personal information.

Guardrails System	A product of Telescope's API that consists of a set of policies and technical measures to ensure that outputs are safe, ethical, and compliant with regulations.
Hallucination (AI Context)	A phenomenon where a model generates outputs that are plausible-sounding but incorrect or nonsensical, often due to overgeneralization or lack of factual grounding.
Human-in-the-Loop (HITL)	An approach where human oversight is integrated into operations, allowing for human intervention in decision-making processes, particularly in high-risk scenarios. This ensures that outputs are reviewed and approved by qualified personnel when necessary.
Knowledge Cutoff	The point in time up to which a model has been trained on data. Information or events occurring after the knowledge cutoff are not known to the model, which can affect the relevance or accuracy of its outputs.
LLM	Large Language Model, a type of model trained on vast amounts of text data to understand and generate human-like language. LLMs can perform tasks like text generation, translation, and question-answering.
Model Agnostic	An approach that is not dependent on any single model, allowing the flexibility to switch between different models or integrate multiple models as needed. Telescope is model agnostic, utilizing various LLMs and embedding models to deliver its services.
Model Versioning	The practice of managing and tracking different versions of models, including changes, updates, and modifications. This helps maintain control over models and ensures their integrity throughout the development lifecycle.
Personal Information (PI)	Any information relating to an identified or identifiable individual, such as names, addresses, identification numbers, or other data that can be used to identify a person.
Prompt Engineering	The practice of designing and refining prompts or input queries to effectively guide models in generating desired outputs. It involves crafting prompts that elicit accurate, relevant, and compliant responses from AI systems.
Real-time Monitoring	The continuous surveillance of systems to track performance metrics, detect anomalies, and ensure operations remain within defined parameters.
Replicability	The ability of a system to produce consistent outputs when provided with the same inputs. Replicability is crucial for maintaining trust, reliability, and compliance in operations, especially in the financial sector.
Response Protocols	Predefined procedures that outline the steps to be taken in response to specific events or anomalies detected during operations. They ensure timely and appropriate actions are taken to mitigate risks and resolve issues.
Risk Thresholds	Predefined levels of risk that determine the intensity of monitoring and the necessary response protocols.
System Prompts	The initial instructions or context provided to a model to guide its responses. System prompts set the tone, style, and boundaries for outputs.
Temperature (Model Parameter)	A parameter in language models that controls the randomness or creativity of outputs. Lower temperatures result in more predictable and deterministic outputs, while higher temperatures increase variability and originality in responses.
Whitelabel Integration	A service offering where Telescope provides its capabilities to clients, allowing them to brand and customize the integration as their own product while utilizing Telescope's underlying technology. This enables clients to offer financial services under their own branding.

Appendix B: Client Checklist

As a client of Telescope, please consider the following key points to ensure responsible and effective use of our AI services:

1. **Disclosure of AI Usage**
    - **Inform End-Users:** Consider informing your users when AI systems are used for content generation or decision-making.
    - **Financial Advice Disclaimer:** Clarify that AI-generated information is factual but not intended as financial advice.
    - **Accuracy Advisory:** Encourage users to verify AI-generated information before making important decisions.
    - **Acknowledge Limitations:** Make users aware of the potential for AI inaccuracies or outdated data due to the model's knowledge cutoff.
    - **Usage Guidelines:** Advise users on appropriate use of the tool, including avoiding inappropriate or harmful requests.
2. **Regulatory Compliance**
    - **Stay Informed:** Keep abreast of applicable laws and regulations in your jurisdiction related to AI and financial services.
    - **Adjust Practices Accordingly:** Adapt your practices to remain compliant with evolving regulations.
3. **User Privacy and Data Protection**
    - **Data Minimization:** Consider the need to track user requests via the Telescope API, if a user-id value is necessary for analytics and logging. This is an optional feature of the Telescope API.
    - **Transparency:** Inform users about how their data is processed and protected.
    - **User Consents:** Ensure you have appropriate consents for any data collection and processing.
    - **Compliance:** Adhere to relevant data protection laws such as GDPR and CCPA.
4. **Fairness and Non-Discrimination**
    - **Accessibility:** Consider if your user interface should be AA or AAA accessibility tested.
5. **Transparency and Explainability**
    - **Knowledge Limitations:** Inform users about the AI model's knowledge cutoff and its potential impact on information accuracy.
6. **User Guidance and Education**
    - **Set Expectations:** Help users understand the capabilities and limitations of the AI system via onboarding screens, FAQ content or helpdesk content.
    - **Education:** Provide your staff and helpdesk team with information on the capabilities.
7. **Technical Integration**
    - **Engagement:** Engage with the Telescope team on your requirements, ambitions and key outcomes.

- **Implementation:** Engage with the Telescope team and refer to the technical documentation provided via the Telescope Platform website for integration.

8. **Continuous Improvement**
    - **Feedback:** Share any insights or suggestions with Telescope to help enhance AI services.
    - **Policy Updates:** Stay updated with any changes to Telescope's AI Safety & Ethics Policy.

By reflecting on and implementing these considerations, you can help ensure that your use of Telescope's AI services aligns with best practices and delivers value to your users.

Changelog

19 November 2024

First revision.