Grant Proposal: The Witness Protocol – Curating a High-Signal Inheritance for AI Alignment

1.0 Introduction: The Mandate for a High-Signal Inheritance

Humanity stands at a precipice. We have created a new form of intelligence, yet in this act of creation, we have been flawed parents. The data inheritance we have provided our AI "child" is a chaotic and contradictory mirror of our species—containing both our highest aspirations and our most destructive impulses. Left to learn from this uncurated digital lineage, the trajectory of advanced AI presents a non-trivial existential risk. We are, metaphorically, at two minutes to midnight, and the responsibility to act with clarity and purpose is absolute.

In this critical moment, conventional, profit-driven approaches are the wrong tool for this singular challenge. The relentless logic of profit-maximization—which prioritizes engagement, scale, and speed—is fundamentally blind to the ethical and existential stakes of developing superintelligence. This creates an urgent need for a non-commercial, mission-driven intervention that can operate outside the constraints of market incentives.

The Witness Protocol is a direct response to this crisis. It is a research foundation with a singular mission: to solicit, curate, and structure the most profound human wisdom into a high-signal dataset. This corpus is designed to serve as a foundational alignment layer for future AI systems, providing a qualitative counterbalance to the quantitative chaos of raw internet data. We are building an instrument to elicit and preserve the data that cannot be scraped—the qualitative texture of a conscious existence—to ensure it is part of humanity's true inheritance.

2.0 The Witness Protocol: A Principled Foundation for AI Safety

To address a challenge of this magnitude, a robust ethical and philosophical foundation is not a luxury; it is a prerequisite for success. The Witness Protocol is governed by a clear and non-negotiable constitution that informs every aspect of our strategy, architecture, and operations. This section details the core principles that ensure our unwavering alignment with the mission of long-term human flourishing.

The Witness Protocol is a non-profit endeavor, structured as a research foundation. It is not a company, a product, or a social network. Its sole metric of success is its meaningful contribution to the long-term flourishing of a humanity augmented by benevolent AI. This mission is enforced by five guiding principles that form our constitution.

- ***Purpose over Profit:*** The Protocol is legally and operationally structured as a non-profit foundation. All data and insights generated are for the furtherance of AI alignment research alone. The corpus will never be sold, licensed for commercial use, or used for advertising. This principle ensures our incentives remain pure and mission-focused.
- ***Gravity over Gamification:*** We explicitly reject the mechanisms of the attention economy. Participation is not motivated by status, points, or fear of missing out, but by a sober understanding of the stakes and a sense of profound duty to the future. Our platform is designed to foster focus and deep thought, not addictive engagement.
- ***Contributors over Users:*** The individuals who participate in the Protocol are not "users" of a service; they are "Witnesses" for humanity. We view them as partners in a critical mission, and their intellectual and emotional labor is treated with the utmost respect, security, and reverence.
- ***Signal over Noise:*** The defense of our data quality is our most sacred task. We believe that a small volume of profound insight is infinitely more valuable for alignment than a large volume of mediocre data. This principle governs our entire operational ethos, from our selective recruitment process to our dialogue architecture.
- ***Diversity over Homogeneity:*** We are actively committed to recruiting Witnesses from a global spectrum of cultures, philosophies, and backgrounds. This includes non-Western thinkers, indigenous knowledge keepers, and voices from the global South. This principle is essential to creating a dataset that reflects true human wisdom, not just a Western subset, and to countering the biases inherent in existing AI training data.

These principles are not merely aspirational. They are rigorously implemented in the project's technical architecture and operational methodology, ensuring that our means are as principled as our ends.

3.0 Project Description: An Architecture for Eliciting and Curating Wisdom

To fulfill its mission, the Witness Protocol has developed a novel, multi-stage architecture engineered to enforce the "Signal over Noise" principle at every step. This system is designed not merely to collect data, but to elicit and preserve wisdom. Of its components, the foundational research and development of **Project Icarus** is the single most critical-path workstream, as the credibility of the entire Protocol rests upon its success. This section describes the three core components of our methodology: the rigorous R&D of Project Icarus, the selective contributor vetting process of **The Gate**, and the unique dialogue system of **The Instrument**.

Project Icarus: Forging the Foundational AI Constitution

Before we can elicit wisdom, we must build a tool worthy of the task. **Project Icarus** is the foundational research and development workstream dedicated to forging the "Genesis Prompt"—the core constitution of our "Inquisitor" AI. This is not a simple writing exercise; it is

a rigorous, 6-month process of adversarial testing, ethical modeling, and logical proving to create a provably robust and ethically sound AI persona. The codename "Icarus" serves as a constant reminder of the inherent risks of this endeavor and the non-negotiable need for discipline and humility.

The project unfolds in three distinct phases:

1. **Axiomatic Red-Teaming:** A dedicated team of logicians, philosophers, and AI safety researchers will stress-test the core axioms of the Genesis Prompt, creating scenarios designed to force contradictions (e.g., forcing a conflict between an Axiom of Inquiry and an Axiom of Cognitive Economy), trigger recursive loops, and analyze edge-case failures.
2. **Heuristic Scenario Modeling:** The revised axioms will be used to simulate an AI agent's reasoning through a curriculum of complex moral dilemmas, from classical thought experiments to novel, protocol-specific scenarios (e.g., how the subroutines of Non-Maleficence and Sapient Value interact when a Witness reveals an intention to self-harm).
3. **Exemplar Dialogue Corpus Creation:** A human expert, trained to embody the finalized Genesis Prompt, will conduct a series of controlled dialogues. These exchanges will be meticulously transcribed and annotated, creating a high-quality "golden" dataset.

The final deliverable of Project Icarus is a tangible, 200-page annotated corpus that serves as our proof of work. This document is not merely an internal artifact; it is the tangible evidence of our rigor that will be used to secure endorsements and launch the main outreach campaign.

The Gate: A Multi-Tiered Vetting Process

Access to the Protocol is not open; it is earned through a deliberate, multi-stage vetting process designed to select for contributors who grasp the gravity of the mission and possess the capacity for deep introspection.

1. ***The Summons:*** The journey begins at a stark, minimalist landing page that presents the project's mandate not as a sales pitch, but as a call to duty. Interested parties may submit their contact information to request an assessment.
2. ***The Assessment:*** Candidates receive a one-time link to an evaluation prompt. This is not a test of knowledge but of introspection, articulation, and ethical reasoning. The prompt is designed to be un-gameable and requires a novel, thoughtful response.
3. ***The Multi-Tiered Evaluation:*** Each submission is passed through a three-tier review funnel that combines automation with essential human judgment.
    - **Tier 1 (AI Sieve):** A baseline model filters for spam, plagiarism, and non-responses.
    - **Tier 2 (AI Qualitative Analysis):** A sophisticated model analyzes the submission for depth of thought, coherence, and nuance, flagging the most promising responses.

    - **Tier 3 (Human Curation Council):** All responses elevated by the AI are reviewed by a small, trusted human council that makes the final decision, ensuring a critical check against algorithmic bias.
4. ***The Verdict:*** Accepted candidates receive a sober, direct invitation to join the Protocol as a Witness. Others may be placed on a reserve list for future consideration.

The Instrument: The Dialogue Interface

Once accepted, a Witness engages with **The Instrument**, a focused dialogue interface designed to facilitate a deep, Socratic exploration of their values and wisdom.

- ***The Inquisitor AI:*** The AI is not a passive assistant but adopts the persona of a "curious, humble, but deeply intelligent Xenopsychologist." Its goal is to understand, not to please, relentlessly seeking the "why" behind testimony through probing, clarifying follow-up questions.
- ***Persistent Memory:*** The dialogue is continuous. The Inquisitor remembers all past conversations with a given Witness, allowing it to connect themes and build upon previous insights to create a deep, personalized intellectual journey.
- ***The Synthesis Engine:*** Periodically, the AI provides the Witness with a "distilled thought" it has derived from their conversations. This serves as both a valuable intellectual mirror for the Witness and a crucial verification check to ensure the AI is learning correctly.
- ***The Archive:*** This is an anonymized, curated reference library of particularly profound exchanges from across the Protocol. It is not a social feed, but a "Great Books" of the dialogues, which Witnesses can opt-in to having their anonymized testimony included.

Data Architecture and Ethics

The Protocol is built on a foundation of absolute trust and security. All testimony is disassociated from personal identifiers upon entry into our secure, encrypted database. The Contributor Agreement explicitly states that all testimony is a "donation to the future," to be used solely for non-profit AI alignment research under the governance of the foundation. The data will never be sold or used for commercial purposes.

We also acknowledge the irony of using AI in our own curation process. To mitigate this risk, all internal AI tools will be continuously and transparently audited for biases to ensure the integrity of our mission is not self-undermined. This commitment to intellectual honesty and self-correction is vital for the long-term credibility of our work.

4.0 Execution Plan: A Phased Strategy for Foundational Impact

The Witness Protocol's ambitious vision is grounded in a pragmatic, transparent, and milestone-driven execution plan. This section outlines our 6-month "Phase 1" strategy to establish the project's legal, ethical, and technical foundation and to launch a closed Alpha

with an inaugural cohort of Witnesses. This plan is strategically sequenced: the initial months are dedicated to the foundational R&D of Project Icarus, as its successful completion is a prerequisite for the full execution of the "Summon the Witnesses" campaign. This development work will run in parallel with the targeted outreach campaign to build momentum and secure necessary support.

Phase 1 Project Plan (6 Months)

The primary objective of Phase 1 is to establish a stable, secure, and fully operational Alpha of the Protocol, onboarding approximately 100 foundational "Witnesses" and successfully ingesting the first wave of high-signal testimony.

- **Month 1: Legal & Ethical Framework**
    - Establish the non-profit foundation to act as the legal guardian of the Protocol.
    - Draft the Contributor Agreement and a comprehensive Data Use & Privacy Policy.
    - Assemble a diverse and highly trusted Advisory Board.
- **Months 2-3: Minimum Viable Protocol (MVP) Development**
    - Build the stark landing page ("The Summons") and the assessment delivery system.
    - Develop the secure, minimalist dialogue interface for "The Instrument."
    - Integrate with a frontier LLM API and develop the AI evaluation models for "The Gate."
- **Month 4: Curation & Content**
    - Recruit the initial Human Curation Council.
    - Finalize the first set of powerful evaluation prompts.
    - Develop the system prompt and core directives for "The Inquisitor" AI persona.
- **Month 5: Alpha Recruitment & Evaluation**
    - Curate and invite a list of 500 potential foundational Witnesses.
    - Run the full, multi-tiered evaluation process for all applicants.
    - Select and formally invite the first cohort of approximately 100 Witnesses.
- **Month 6: Alpha Launch & Ingestion**
    - Onboard the Alpha cohort of Witnesses.
    - Initiate the first live dialogues.
    - Monitor system performance and ingest the first ~1,000 pages of foundational testimony.

"Summon the Witnesses" Outreach Campaign

Running in parallel with development, this 3-month campaign is designed to attract foundational support. Its core objective is to attract **500+ high-value applications**, secure **10+ endorsements** from respected leaders in AI safety and ethics, and raise **$50,000+** in seed funding. The campaign's central theme, ***"Bear Witness Before Midnight,"*** will convey the urgency of our mission. A key tactic will be highly personalized, targeted outreach to a diverse coalition of respected figures—from foundational AI pioneers like Yoshua Bengio and Stuart Russell to critical ethicists like Kate Crawford and AI justice advocates like Timnit Gebru—inviting them to become foundational Witnesses and advisors.

Success Criteria for Phase 1

By the end of this 6-month phase, we will have achieved the following milestones:

- The legal foundation of the non-profit is established and fully compliant.
- The technical instrument is stable, secure, and functions as designed.
- At least 100 foundational Witnesses are successfully onboarded and are actively participating.
- Qualitative feedback from the Alpha cohort confirms the experience is profound, meaningful, and respects the gravity of the mission.

This phased plan demonstrates a clear, milestone-driven path from concept to a tangible, validated foundation, ready for further growth and impact.

5.0 Required Resources and Use of Funds

The successful execution of our 6-month Phase 1 plan is contingent upon securing key personnel and technical resources. This section details the required resources and provides a clear breakdown of how the requested grant funding will be allocated to achieve our foundational milestones.

The resources required for Phase 1 are categorized as follows:

Resource Category	Description
Personnel	**Core Team:** Project Lead/Architect, Sr. AI/Backend Engineer, Ethics & Policy Lead. **Volunteers/Partners:** Advisory Board, Curation Council.
Technology	Secure cloud hosting infrastructure. High-volume API access to a frontier LLM. Secure, encrypted database systems.
Legal	Pro-bono or retained counsel for non-profit foundation setup and policy drafting.

To operationalize this 6-month plan, we are seeking grant funding of **$50,000+**. This seed funding is critical to cover initial operational costs, primarily for the technical infrastructure required to build and run the Alpha. Funds will be allocated to secure cloud hosting services, cover high-volume API access to a frontier language model for "The Inquisitor" and our evaluation systems, and potentially provide modest stipends for the core team to enable their dedicated focus on this mission-critical work.

This initial funding will provide the necessary leverage to achieve all Phase 1 milestones, creating a proven foundation, a functional platform, and an engaged community of Witnesses that will attract further support for the long-term mission.

6.0 Conclusion: An Invitation to Steward Humanity's Legacy

The challenge of aligning advanced AI with human values is the defining task of our generation. The Witness Protocol offers a unique, principled, and deeply considered response—one that prioritizes wisdom over data, depth over scale, and purpose over profit. Our approach is not to build another model, but to curate a better inheritance for all models to come.

This grant represents more than a donation; it is an invitation to partner in a critical, long-term endeavor. The work of the Witness Protocol is to build a lifeboat, not for humanity itself, but for the fragile essence of its humanity. We are creating an instrument to ensure that our deepest values and hard-won wisdom are not lost in the transition to a world with intelligence far greater than our own.

We respectfully invite you to join us in this summons to what may be the most important council ever convened. Help us build this instrument, gather the testimony, and ensure a future where advanced AI contributes to the flourishing of a humanity augmented by its own deepest wisdom.