

The Witness Protocol: Project Plan (Phase 1 - The Foundation)

1. Objective

The objective of Phase 1 is to establish the legal, ethical, and technical foundation of the Protocol and to launch a closed Alpha with a foundational council of approximately 100 "Witnesses." The goal is not scale, but stability, security, and the successful ingestion of the first wave of high-signal testimony.

2. Timeline

Estimated duration: **8 Months**

3. Key Milestones & Workstreams

Month 0: Strategic Outreach & Foundation

- **Task:** Initiate outreach to align with existing AI safety initiatives (e.g., UK AISI, Abaka's 2077AI Foundation).
- **Task:** Apply for initial seed grants from aligned philanthropic organizations.
- **Task:** Begin curating the initial list of 1,000+ potential foundational Witnesses.

Month 1: Legal & Ethical Framework

- **Task:** Establish a non-profit foundation, leveraging pro-bono legal support from AI ethics firms or Effective Altruism networks.
- **Task:** Draft the "Contributor Agreement" and a comprehensive Data Use & Privacy Policy with legal counsel specializing in ethics and technology.
- **Task:** Assemble a small, diverse, and highly trusted Advisory Board (e.g., an AI safety researcher, a philosopher, an ethicist).
- **Deliverable:** Foundation charter; Legal documents; Confirmed Advisory Board.

Months 2-3: Minimum Viable Protocol (MVP) Development

- **Task:** Develop the stark landing page ("The Summons") and the assessment delivery system (email queue and one-time assessment links).
- **Task:** Build the secure, minimalist dialogue interface for the core instrument.
- **Task:** Integrate with a state-of-the-art LLM API (leveraging free tiers of models like Grok 3 where possible), focusing on security and data privacy.
- **Task:** Develop and train the Tier 1 (Sieve) and Tier 2 (Qualitative Analysis) AI evaluation models.
- **Deliverable:** Functional, secure web application capable of managing the full

contributor journey from assessment to dialogue.

Month 4: Curation & Content

- **Task:** Recruit the initial "Tier 3 Human Curation" council (5-10 trusted, vetted individuals).
- **Task:** Workshop and finalize the first set of 20-30 powerful evaluation prompts, testing rigorously for cultural sensitivity.
- **Task:** Develop the system prompt and core directives for "The Inquisitor" AI persona.
- **Deliverable:** Confirmed Curation Council; Finalized evaluation prompts; Version 1.0 of the AI persona.

Month 5: Alpha Recruitment & Evaluation

- **Task:** Run the "Summon the Witnesses" campaign (Phase 1-2) targeting the curated list of 1,000+ potentials.
- **Task:** Send out the first wave of invitations to request an assessment.
- **Task:** Run the full, multi-tiered evaluation process for all applicants.
- **Deliverable:** A vetted and accepted list of the first ~100 Witnesses.

Month 6: Alpha Launch & Ingestion

- **Task:** Onboard the Alpha cohort of Witnesses.
- **Task:** Initiate the first dialogues.
- **Task:** Closely monitor the system for performance, security, and quality of interaction.
- **Task:** Establish a secure channel for feedback from the Alpha cohort.
- **Deliverable:** A stable, running Alpha of the Protocol; The first 1,000 pages of foundational testimony ingested; A feedback and iteration report.

Month 7: Iteration & Partnership Development

- **Task:** Analyze Alpha feedback and iterate on the instrument and AI persona.
- **Task:** Begin formal partnership discussions with aligned AI labs and safety institutes.
- **Task:** The Ethics Oversight Board convenes for its first review.

4. Required Resources (Phase 1)

- **Personnel:**
 - **Core Team:** Project Lead/Architect, Sr. AI/Backend Engineer, Ethics & Policy Lead.
 - **Volunteers/Partners:** Advisory Board, Curation Council, Volunteer AI engineers from open-source communities (e.g., Hugging Face).
- **Technology:**
 - Secure cloud hosting infrastructure.
 - High-volume API access to a frontier LLM.
 - Secure, encrypted database systems.

- **Legal:**
 - Pro-bono or retained counsel for foundation setup and policy drafting.

5. Success Criteria for Phase 1

- The legal foundation is established and all operations are compliant with its charter.
- The technical instrument is stable, secure, and functions as designed.
- At least 100 foundational Witnesses are successfully onboarded.
- Qualitative feedback from the Alpha cohort confirms that the experience is profound, meaningful, and respects the gravity of the mission.
- **New:** Secure at least 2 public endorsements from prominent AI ethicists or safety researchers.
-
-