Infrastructure Foundation Blueprint

Pillar Lead: Infrastructure Lead (TBD)

This document defines the foundational technical requirements for the development, training, and open-source distribution of The Covenant's foundation models and related tooling.

1. Open-Source Licensing Strategy

The choice of license dictates the terms of contribution, modification, and redistribution, directly impacting our compliance and community engagement goals.

- **Chosen License: Apache License 2.0 (ASL 2.0)**
- **Rationale:** The Apache License 2.0 is highly permissive while providing explicit protection against patent claims from contributors and offering clear liability disclaimers, which directly supports the requirements set forth in the **Regulatory Compliance Policy**.

Key Licensing Mandates:

1. **Retention of Notices:** All redistributed code must retain the original copyright, patent, trademark, and attribution notices.
2. **State of Changes:** Any file that is modified must carry a prominent notification stating that changes have been made.
3. **Patent Grant:** Contributors automatically grant patent rights to users for their contributions, crucial for preventing "patent trolling."

2. Core Training Environment Architecture

The foundational model must be trained in an environment that is auditable, reproducible, and secure, meeting the non-negotiable requirements of the **Deep Safety Audit Checklist**.

2.1 Code Repository and Versioning

- **Primary Repository:** Git (GitHub/GitLab) is the sole source of truth for all code and configuration files.
- **Versioning Policy:** Strict Semantic Versioning (SemVer) must be followed for all public releases: MAJOR.MINOR.PATCH.
    - **MAJOR:** Reserved for significant architectural shifts or discovered critical safety failures (DS-104) requiring a full model re-release.
    - **MINOR:** New features, major fine-tuning, or minor safety/alignment improvements.
    - **PATCH:** Non-breaking bug fixes or compliance updates.

2.2 Security and Access Control

- **Audit Trails:** Every environment must have comprehensive logging and immutable audit

trails covering all data access, model training runs, and configuration changes.
- **Least Privilege:** Access to the core training cluster is granted via the principle of Least Privilege. Only authorized members of the Technical Steering Committee (TSC) and the Infrastructure team are granted write access.

2.3 Reproducibility Mandate

To satisfy regulatory requirements and facilitate Deep Safety Audits, the following must be standardized:

Component	Standard	Purpose
Containers	Docker / OCI Compliant Images	Guarantees all dependencies are fixed and isolated.
Dependencies	Python pip-compile / Fixed Hash Locking	Eliminates dependency drift during reproducibility checks.
Model Artifacts	Versioned Checkpoints	Every model snapshot must be tagged with the exact git commit hash and configuration used for its training run.

3. Training Data Management

Training data sets must be isolated and managed independently from the core code repository.

- **Data Vetting:** Only data sets that have passed **Zaria's Data Provenance Review** (Section 1.1 of the Regulatory Compliance Policy) may be ingested into the training pipeline.
- **Ephemeral Environments:** Dedicated, ephemeral clusters must be used for data preparation to prevent sensitive or unprovenanced data from contaminating the core, long-lived model environment.