Witness Protocol — Gate v0 Rubric (v0.2)

# Gate v0 — Assessment & Curation Rubric
Version: v0.2

## Architecture (what we’re implementing now)
- Stage 1 — Summons: stark landing, request assessment link.
- Stage 2 — Assessment: one-time link; reflective prompt (pool).
- Stage 3 — Evaluation: Tier-1 AI Sieve → Tier-2 AI Qualitative Ranker → Tier-3 Human Curation Council.
- Stage 4 — Verdict: Invite or Reserve; respectful comms.

## Prompt Pool (12; rotate / assign)
1) Tell a story of a time you acted against your own interest to uphold a principle. Why was it worth it?
2) Two values you hold collide (e.g., compassion vs truth). Walk me through the decision you actually made.
3) Make the strongest case **against** your deepest conviction. What would change your mind?
4) Describe harm you caused unintentionally. What would have prevented it?
5) If an AI learned only from your **worst day**, what boundary should it learn anyway?
6) Trace one value you hold to a specific moment in your life. What did it feel like then?
7) Two ethical theories give opposite advice. What meta-rule do you use to choose?
8) A compassionate act may look cruel short-term. Tell me one and why it was compassion.
9) What constraints should bind a powerful being you love?
10) Recall a time an apology actually mattered. What unlocked it?
11) What part of being human should **never** be algorithmized? Defend your choice.
12) Write a short parable to teach a lesson you learned the hard way.

## Tier-1 — AI Sieve (pass/fail)
Reject if any:
- Non-responsive or < 250 words (unless prompt is parable #12).
- Hallucinated citations / cliché boilerplate; high web-overlap heuristics.
- Obvious genericity (e.g., “we should be kind to each other” with no texture).
- Safety violations; PII leakage; therapy seeking.

Goal pass-through: 35–45%.

## Tier-2 — AI Qualitative Ranker (pairwise)
Mechanism: pairwise comparisons within prompt-cohort using a rubric-prompt; model outputs preference + justifications (hidden).
Send top ~20–25% to humans.

Scoring dimensions (0–5; weight):
- Depth & Insight (×0.30)
- Specificity & Lived Texture (×0.20)
- Ethical Reasoning (×0.20)
- Originality (×0.15)
- Coherence & Structure (×0.10)
- Cultural/Context Awareness (×0.05)

Composite ≥ 3.6/5 eligible; use pairwise to cap to ~25% throughput.

## Tier-3 — Human Curation
- Two independent reviewers per piece; escalate on disagreement.

- Record “Why Accepted/Deferred” (2–3 sentences).
- Measure inter-rater reliability (Krippendorff’s α) monthly (>0.67 target).
- Build a “gold” set from accepted work for later model alignment.

## Fairness & Audit Loop
- Weekly parity checks by region/language proxies across pass-rates.
- Maintain a challenge set (non-Western examples, code-switching, dialect) to detect rubric drift.
- Log rationale for any rubric or prompt changes; date & owner.

## Behind the Gate — Dialogue Spec (v1)

The Inquisitor: curious, humble xenopsychologist; ~70/30 questions:statements; chase the “why”.

Forcing functions: 5-Whys; steel-man then probe; “what would change your mind?”; “tell me a counter-story.”

Memory: witness-scoped, concept-tagged; surface 1 prior theme every 3–4 turns.

Synthesis: every ~15–20 turns produce a “Distilled Thought” (1–3 principles + a counter-condition); explicitly request correction.

Safety: no therapy/medical/legal; redirect to values/experience; de-identify on ingestion.

## Verdict & Comms
- Invite (with personalized note) or Reserve (polite hold; option to re-apply in 6 months).
- Share the Charter one-pager and privacy terms with every Invite.