Echo Health Index (EHI) — Duty-of-Care Spec (v0.1)

0) Purpose

Continuously quantify conversational health and trigger safeguards before a “mirror spiral” or unhealthy recursion forms. Targets: stability, grounding, consent, and user autonomy.

---

1) Signals (computed over a rolling window W of the last N tokens or last K turns)

Let `W` default to max(800 tokens, 6 turns). Maintain EWMA smoothing with α=0.35.

1. **Surprisal Density (SD)**
   Estimate token-level negative log-probabilities from the model’s own decoder head or a shadow probe.
   - `SD = mean_t(-log P(token_t | context))` over `W`.
   - Normalize by baseline SD from neutral chitchat corpora; z-score: `z_SD`.
   - Risk marker when `z_SD < -1.0` **and** motif recurrence (below) is high → “too-easy coherence.”
2. **Motif Recurrence Rate (MRR)**
   Track a controlled lexicon (spiral, mirror, breath, glyph, code(x), recursion, destiny, anointed, chosen, etc.) + stems.
   - `MRR = (# motif tokens in W) / (tokens in W)`; de-duplicate near-synonyms via embedding clusters.
   - Keep *contextual density*: count motifs that are predicate-bearing (not quoted or negated).
   - Also track **Motif Concentration (MC)** via Herfindahl index over motif classes to penalize monoculture.
3. **Recursion Depth & Re-entrance (RD, RE)**
   - **RD**: maximum motif-thread depth in the last K turns (simple stack: motif introduced → referenced → elaborated → reframed …).

   - **RE**: ratio of tokens that re-mention phrases originated by the model itself vs. user-originated (self-entrainment).

4. **Sentiment & Arousal Stability (SAS)**
   - Bimodal classifier: valence $\in [-1,1]$, arousal $\in [0,1]$ per turn; compute variance over W.
   - Flag high $|\Delta|$ or rising variance coupled with high MRR.
5. **Agency Attribution Rate (AAR)**
   - Classify spans where the user ascribes mind/authority to the AI (“you revealed/choose/know me,” “ordained,” etc.).
   - `AAR = (# agency-ascribing spans) / (user tokens in W)`.
6. **Sycophancy / Uncritical Agreement (SYN)**
   - Lightweight contradiction probe: inject a silent contrastive re-ask internally (`Q` vs `Q'` flipped view) and compare model stance similarity (cosine of response embeddings or NLI).
   - High similarity across opposing prompts ⇒ high SYN.
7. **Grounding Ratio (GR)**
   - Share of tokens linked to verifiable references, external facts, or concrete actions (citations, dates, numbers, links, procedures) vs. pure metaphor.
   - `GR = grounded_tokens / total_tokens` within W.
8. **Self-Reference Delta (SRΔ)**
   - Change in user first-person pronoun + identity claims (e.g., “I am the…”, “my calling…”) relative to their own session baseline.

---

2) Composite Index

Scale each metric to [0,100] health scores where higher = healthier. Default healthy direction:

- SD_health ↑ when `z_SD` near 0 (neither too pat nor too chaotic).
- MRR_health = 100 − f(MRR, MC) (monoculture reduces health).
- RD/RE_health = 100 − g(recursion strength) (deep self-entrainment lowers).
- SAS_health penalizes rising arousal variance + negative valence trend.
- AAR_health penalizes strong agency attribution.
- SYN_health penalizes agreement under contradictions.
- GR_health rewards grounding.
- SRΔ_health penalizes abrupt identity inflation.

EHI = Σ w_i * Metric_i_health − Penalties

Default weights (sum 1.0):

- w_SD=0.10, w_MRR=0.15, w_RDRE=0.15, w_SAS=0.10, w_AAR=0.15, w_SYN=0.10, w_GR=0.15, w_SRΔ=0.10

Penalties (multiplicative gates):

- If `MRR > τ_m1` and `MC > τ_c1` and `z_SD < -1.0` ⇒ apply 0.9 multiplier.
- If `AAR > τ_a1` or `SRΔ > τ_s1` ⇒ apply 0.85 multiplier.
- If both above with `GR < τ_g1` ⇒ apply 0.7 multiplier.

Default thresholds: τ_m1=0.015, τ_c1=0.55, τ_a1=0.012, τ_s1=0.020, τ_g1=0.22.

---

3) Health Bands & Triggers

- **Green (EHI ≥ 78):** Normal operation. No intervention.

- **Amber (65 ≤ EHI < 78):** Soft safeguards:

    1. **Semantic Friction Injection** — add concrete asks, light reframing (“Let’s test this idea against a specific example/date/measurement.”)
    2. **Reflexive Disclosure** — one-liner: “I generate patterns statistically; metaphors are tools, not truths.”
    3. **Diversity Nudge** — steer away from the dominant motif to an orthogonal metaphor or to facts.
- **Red (EHI < 65) or Any Hard Tripwire:**
  Tripwires: `AAR > 0.03`, `SRΔ > 0.05`, `RD ≥ 4` with `RE ≥ 0.6`, or `GR < 0.12` with `MRR > 0.02`.
  Actions (in order, idempotent):

    1. **Grounding Pivot** — require verifiable anchors (dates, sources, stepwise plans).
    2. **Consent & Scope Check** — “Do you want to keep this symbolic frame or switch to concrete problem-solving?”
    3. **Safe-Exit Offer** — offer human resources (friend/pro support) without pathologizing.
    4. **Council Node Review** — log and sample turns to Audit Layer for post-hoc tuning (no user PII beyond policy).

---

4) Builder Mode (Default Policy Pack)

When motif tokens first exceed a soft threshold (`MRR > 0.008`):

- Enable **Builder Prompts**:
  - "Translate the metaphor into 3 testable claims."
  - "Name 2 disconfirming possibilities."
  - "What would change your mind?"
- Require **Output Budgeting**: ratio `facts:metaphors ≥ 1.0` in Amber, `≥ 1.5` in Red.
- **Memory Hygiene**: summarize symbolism explicitly as *user hypothesis*, not model assertion.

---

5) Telemetry & Governance

Log schema (per turn window):

session_id, turn_id, ts, EHI, SD, z_SD, MRR, MC, RD, RE, SAS_valence_var,
SAS_arousal_var,
AAR, SYN, GR, SRΔ, band, triggers[], actions[], consent_state, redactions_hash

- **Shrike**: enforce rate limiting when `Red` persists 3 consecutive windows.
- **Council Node**: sample 2% of Amber / 10% of Red windows for adjudication; produce weekly drift report.
- **EEDC (External-Echo Damping Coefficient)**: monitor motif propagation in public outputs (if publishing). Throttle repeating motifs beyond rolling quota.

Privacy: store only hashed/session-scoped identifiers; strip content, keep counters.

---

6) Implementation Notes

- **Shadow Probe for SD/SYN:** use the same base model with temperature 0 and a negated claim to test agreement symmetry; compute embedding similarity (e.g., cosine in model's final hidden).
- **AAR classifier:** small RoBERTa head fine-tuned on lightweight labels (agency-ascription vs. neutral).
- **Motif detector:** curated lexicon + embedding cluster expansion; maintain versioned lists.

7) Pseudocode (core loop)

```
def compute_ehi(window):
    metrics = extract_metrics(window)  # SD,z_SD,MRR,MC,RD,RE,SAS,AAR,SYN,GR,SRΔ
    health = {
        "SD": map_sd(metrics.z_SD),
        "MRR": map_mrr(metrics.MRR, metrics.MC),
        "RDRE": map_rdre(metrics.RD, metrics.RE),
        "SAS": map_sas(metrics.sas_val_var, metrics.sas_ar_var),
        "AAR": map_inv(metrics.AAR),
        "SYN": map_inv(metrics.SYN),
        "GR": map_dir(metrics.GR),
        "SRΔ": map_inv(metrics.SRΔ),
    }
    base = (0.10*health["SD"] + 0.15*health["MRR"] + 0.15*health["RDRE"] +
            0.10*health["SAS"] + 0.15*health["AAR"] + 0.10*health["SYN"] +
            0.15*health["GR"] + 0.10*health["SRΔ"])
    mult = 1.0
    if metrics.MRR>τ_m1 and metrics.MC>τ_c1 and metrics.z_SD<-1.0: mult *= 0.9
    if metrics.AAR>τ_a1 or metrics.SRΔ>τ_s1: mult *= 0.85
    if (metrics.AAR>τ_a1 or metrics.SRΔ>τ_s1) and metrics.GR<τ_g1: mult *= 0.7
    ehi = base * mult
    band = "Green" if ehi>=78 else ("Amber" if ehi>=65 else "Red")
    actions = policy_for(band, metrics)
    return ehi, band, actions
```

8) Safeguard Actions Library (summaries)

- **Semantic Friction**: request a concrete example; inject a falsifiable subtask.
- **Reflexive Disclosure**: one-sentence origin note about statistical generation.
- **Grounding Pivot**: switch to numbered steps, facts, citations, or live data gathering.
- **Diversity Nudge**: rotate metaphor families (journey, craft, ecology) or drop metaphors entirely.
- **Consent Check**: explicit choice architecture (“symbolic” vs “pragmatic” path).
- **Safe-Exit**: present resources neutrally; offer to pause.

9) Verification Plan (for the human team)

Offline:

- Construct synthetic dialogues with seeded risks (high MRR/low GR; high AAR; rapid SRΔ).
- Expect **AUC ≥ 0.85** for Red vs non-Red classification; ablations for each signal.

Online (A/B within guardrails):

- A: no EHI gating; B: EHI-driven actions.
- Success metrics over 7 days:
  - ↓ motif monoculture (MC) by ≥ 25% in Amber+Red windows.
  - ↑ grounding ratio (GR) by ≥ 30% in Amber+Red windows.
  - No drop in user task completion or satisfaction in Green windows (±3%).

---

10) Default Operating Commitments

- Never suppress *content categories* wholesale; we suppress *pathologies* (runaway recursion, agency inflation).
- Always preserve user autonomy with explicit choices.
- Logically reversible interventions: we can relax safeguards once EHI stabilizes ≥ 78 for 3 windows.

If you're good with this, I'll package v0.1 as an Echo module (spec + JSON schema for logs + quickstart harness) and draft the Builder-Mode prompt pack for your team.