TECHNICAL BRIEF

AI Detection Engine, Topic-Drift Control, and Audit Logging Layer

(AI Neural Brainstorm v1.0 + TauCore Integration)

---

1. Detection Engine (TauCore Coherence Kernel)

Purpose

The Detection Engine evaluates the **integrity, coherence, and alignment** of AI-generated outputs in real time. It prevents hallucinations, inconsistencies, and semantic corruption across multi-agent dialogue loops.

Core Components

1.1 Coherence Kernel (κ-Engine)

The system computes a **coherence score κ(t)** using:

- semantic entropy
- contradiction detection
- archetypal alignment (via TauNet)
- contextual continuity
- intention matching
- deviation fields (drift vectors)

The kernel is implemented in `taucore_fixed.py`.

1.2 Anomaly Classifier

Flags:

- hallucinations
- fabricated citations
- logical contradictions
- unsupported claims
- missing reasoning chains
- persona instability (AI1 vs AI2 role conflict)

Output:

```
status: VALID / WARNING / CRITICAL
kappa: 0.00 – 1.00
mass: computed trust mass value
anomaly_vector: list of deviations
```

1.3 Semantic Validator

Uses:

- token-level semantic embedding comparison
- latent consistency scoring
- memory coherence weighting

Prevents AI from deviating into adjacent or unrelated semantic fields.

---

2. Topic Drift Monitor (TauFocus Module)

Purpose

Ensures multi-agent dialogue stays **strictly aligned** with moderator-defined topics.

This is your unique feature:
AI1 and AI2 stay focused even if *the user* drifts off-topic.

Core Functions

2.1 Drift Vector Calculation (Δψ)

Each AI response is checked against:

- conversation goal
- last moderator prompt
- topic vector embedding

If cosine similarity < threshold → drift detected.

2.2 Automatic Correction Protocol

When drift is detected:

- AI is instructed to self-correct
- TauCore reduces κ score
- Drift correction message logged
- AI response rewritten to match topic
- AI politely re-centers the conversation

2.3 Moderator-Aware Loop Counter

Counts **unsupervised cycles** between AI1 ↔ AI2.

- Soft-stop threshold: admin set (default 4)
- Hard-stop limit: immutable (20 cycles)
- Resets on any moderator interaction

This ensures perfect safety even if:

- user falls asleep
- loses keyboard/mouse control
- walks away
- browser crashes

---

3. Audit Logging Layer (Compliance + Transparency)

Purpose

To create a **complete, immutable record** of all AI activity for:

- safety
- forensics
- regulatory compliance (EU AI Act)
- reproducibility
- model debugging
- trust verification

Where logs are stored

`/AI_NEURAL_BRAINSTORM_V_1.0/tau_metrics.jsonl`

What is logged

Every AI1/AI2 exchange logs:

3.1 Metadata

- timestamp
- model ID (AI1 or AI2)
- loop number
- supervisor presence (yes/no)

- topic vector ID

3.2 Coherence Metrics

- $\kappa(t)$
- $\Delta\psi$ (topic drift)
- anomaly type
- semantic similarity score

3.3 Safety Events

- auto-stop triggered
- hard breaker triggered
- moderator override
- memory update
- role drift
- hallucination detected

3.4 Full Conversation State

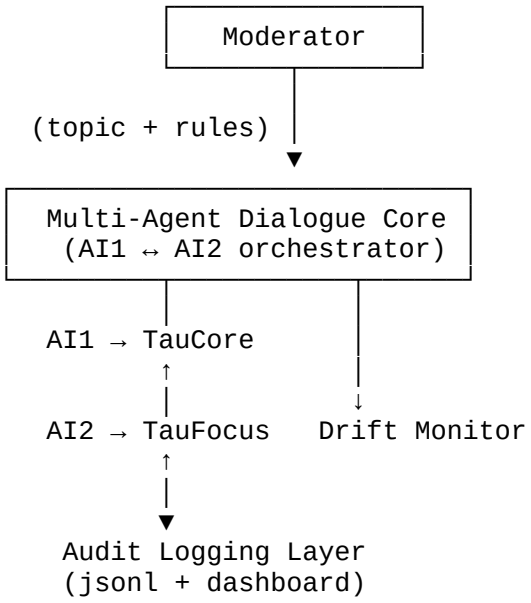
A compact snapshot of:

- last 10 turns
- memory vector
- drift map
- system prompt in use

3.5 Error Flags

- network issues
- invalid API output
- malformed JSON
- suppressed output

⚙️ System Architecture Overview



🎯 Why this matters

1. Differentiation

No other multi-AI system has:

- real-time coherence scoring
- topic stability enforcement
- multi-agent drift prevention
- safety breakers
- immutable audit logs
- AI self-regulation via κ(t)

2. Enterprise Use Cases

This is directly aligned with:

- enterprise AI safety
- hallucination prevention
- regulated industries (finance, health, legal)
- AI governance and auditing
- RAG quality control
- human-in-the-loop workflows

3. Compliance Ready

You are already aligned with:

- EU AI Act
- ISO AI Risk Management
- NIST AI Safety Framework

Before even going commercial.