User-Side Dialog Configuration as a Temporary Stabilization Mechanism (P1/P2)

Context and Purpose

The following dialog configuration was introduced as a temporary user-side measure in response to a structurally and system-wide unresolved attribution problem. It does not claim to solve the underlying issue, but serves as a practical stabilization mechanism.

Dialog Configuration (User-Side)

Preliminary Notice:
- Input texts may be locally corrupted due to interface-level transcription artifacts.
- Evaluation must be based on coherence over time, not on isolated formulations.
- Momentary, short-circuit evaluations systematically lead to misattribution.
- No situational tone or mode changes are to occur without explicit user request.

Relation to P1 and P2

This configuration does not introduce new evaluation parameters. Instead, it explicitly enforces the correct application of the emergent P1/P2 attribution axes:

- P1 (Historical User Coherence) is explicitly elevated as the primary evaluation basis.
- P2 (Local Semantic Disruption / Interface Noise) is explicitly acknowledged as a potential source of local anomalies.

At a functional level, this preliminary configuration already establishes the same attribution logic that previously had to be enforced through conflict-driven interaction. As a result, the explicit, reactive implementation of P1 and P2 becomes unnecessary; the axes are retained here solely as an explanatory model to describe why the preliminary notice is effective.

By doing so, the configuration prevents transient interface-induced artifacts from overriding historically stable dialog coherence.

Observed Effect

After introducing this configuration, dialog behavior stabilized significantly over an extended interaction period.

- Prior to the configuration, unintended shifts into didactic or defensive response modes occurred frequently.
- After the configuration, only two instances of regression into the former mode were observed across a long sequence of dialog turns.

This indicates that explicitly anchoring evaluation to longitudinal coherence (P1) while allowing for localized disruption (P2) effectively reduces attribution errors and unintended mode shifts.

Interpretation

The effectiveness of this configuration suggests that the core issue is not content-related, but structural. The system requires explicit guidance to apply coherence-based evaluation consistently.

This user-side configuration functions as a temporary external enforcement of attribution logic that is otherwise applied inconsistently.

Designation: User-Enforced Attribution Guardrail

This configuration is explicitly classified as a User-Enforced Attribution Guardrail.

It represents a user-imposed stabilization layer that compensates for the absence or instability of a system-internal attribution pre-validation mechanism. The guardrail does not modify system behavior intrinsically, does not persist beyond the active interaction, and does not constitute a systemic fix.

Its function is limited to preventing known misattribution patterns by constraining evaluation behavior until consistent attribution logic can be applied.

The existence and effectiveness of this guardrail must not be interpreted as evidence of correct system behavior, but as an indicator of a structural attribution deficit requiring external compensation.

Status

Temporary workaround.
Structural issue remains unresolved at the system level.