

# Co-Koala: A Causal Operator Evaluation Framework for Model Reasoning, Stability, and Alignment

A Structured Methodology for Testing Causal Strength,  
Operator Fidelity, and Multi-Step Reasoning Integrity

Michael Nget

Aetherion Labs

mike@aetherionlabs.us

Technical Report

December 3, 2025

---

*Co-Koala introduces a unified methodology for evaluating causal operators, multi-step reasoning, structural fidelity, and alignment robustness in AI systems by analyzing causal transformations, inference pathways, and operator-induced system behavior.*

---

## Abstract

Co-Koala is a structured evaluation framework for testing the causal reasoning capabilities, operator fidelity, and multi-step inference stability of modern AI systems. It provides a standardized methodology for probing how models construct, apply, and transform causal operators across tasks, contexts, and domains. By analyzing the behaviors of operator-driven reasoning processes, Co-Koala reveals whether models maintain causal integrity under perturbation, ambiguity, role changes, and long-range inference chains.

The framework formalizes causal operators as transformations applied to world states, reasoning frames, or inference pathways. Co-Koala tests how well models retain causal structure, detect operator misuse, maintain consistency across causal branches, and reassemble correct inferences after intervention. It produces interpretable metrics including causal fidelity, operator stability, reconstruction accuracy, perturbation resilience, and cross-context consistency.

Co-Koala supports natural language reasoning models, multimodal models, tool-using agents, multi-agent systems, and cognitive operating systems such as CodexOne. It provides a principled foundation for evaluating causal robustness, alignment behaviors, abductive reasoning quality, and architectural reliability in complex AI systems.

## 1 Introduction

Collaborative problem solving under conditions of ambiguity, time pressure, and high cognitive load is central to work in science, engineering, design, and policy. Yet decades of research in cognitive science and human–computer interaction have shown that unstructured collaboration regularly fails to produce reliable or reproducible outcomes. Common breakdowns include premature convergence, uneven participation, conversational drift, unclear problem framing, and the absence of explicit mechanisms for testing or discarding hypotheses. These failure modes become more pronounced as problems grow in complexity or when collaborators must integrate heterogeneous reasoning styles.

Recent advances in mixed-initiative systems and large language models (LLMs) have renewed interest in structured approaches to reasoning. However, existing frameworks—such as Double Crux, adversarial collaboration, and calibration techniques—typically lack real-time enforcement mechanisms, operational thresholds, or explicit procedures for managing conversational dynamics. Moreover, few frameworks specify how collaborators should detect drift, challenge assumptions, or maintain traceability during intense analytical tasks. There is an emerging need for structured, enforceable protocols that can guide high-stakes dyadic reasoning in a way that is transparent, reproducible, and suitable for future integration with AI-assisted systems.

To address this gap, we introduce Co-Koala v3.7, a fully specified dyadic reasoning protocol designed to improve clarity, stability, and accountability during complex analytical work. The protocol integrates preflight verification, real-time drift detection, structured falsification, entropy-based perturbation triggers, airtime asymmetry controls, and immutable session logging into a unified operational framework. Rather than a conversational style or loose set of guidelines, Co-Koala v3.7 provides an explicit, enforceable procedure for navigating high-ambiguity analytical tasks.

The primary contribution of this paper is the complete specification and design rationale of the protocol. We describe the theoretical motivations behind each component, detail the iterative co-design process used to refine operational thresholds, and present a structured plan for future empirical evaluation. Our goal is to provide a reproducible, alignment-aware framework that can support both human–human and human–AI collaborative reasoning in complex domains.

## 2 Related Work

Research on collaborative reasoning spans cognitive psychology, human–computer interaction, computer-supported cooperative work, and mixed-initiative system design. We review four areas most relevant to the development of Co-Koala v3.7: structured reasoning frameworks, adversarial and calibration-based collaboration, conversational breakdown analysis, and AI-assisted joint reasoning.

### 2.1 Structured Reasoning Frameworks

Double Crux [for Applied Rationality, 2017] aims to surface underlying assumptions driving disagreements, while argumentation mapping and structured deliberation frameworks provide scaffolding for hypothesis competition and evidence evaluation [Schefström, 2019]. However, these frameworks rely heavily on voluntary adherence rather than real-time procedural enforcement. Co-Koala v3.7 extends this work by defining enforceable rules for drift detection, hypothesis falsification, and perturbation triggers.

### 2.2 Adversarial Collaboration and Calibration

Adversarial collaboration [Tetlock and Gardner, 2019] and calibration training [Mellers and Tetlock, 2014, Koriath, 2012] emphasize structured challenge and explicit hypothesis testing. These approaches demonstrate improved reasoning quality but lack procedural tools for managing conversational breakdowns such as asymmetric airtime or drift into abstraction. Co-Koala v3.7 builds on these traditions by introducing operational structures for balancing participation and enforcing hypothesis evaluation during live interaction.

### 2.3 Conversational Drift and Coordination Breakdowns

CSCW research has extensively documented failures in collaborative analytical work, including premature convergence, conversational dominance, and erosion of shared problem framing [Olson and Olson, 2014]. Co-Koala v3.7 directly targets these breakdowns through drift gradients, airtime asymmetry controls, and preflight verification steps designed to stabilize early-stage reasoning.

### 2.4 AI-Assisted Joint Reasoning

Recent work on human–AI teaming suggests that large language models can serve as effective ideation partners, though they introduce alignment and coordination challenges [Subramonian, 2023]. Mixed-initiative systems research points to the need for explicit negotiation of uncertainty and grounded representations. Co-Koala v3.7 incorporates these insights by specifying immutable logging, explicit reasoning commands, and enforceable steps suitable for hybrid human–AI reasoning contexts.

## 3 Methodological Positioning and Protocol Derivation

Co-Koala v3.7 is presented as a structured reasoning protocol derived through systematic design, constraint analysis, and iterative refinement rather than through completed human-subject experimentation. This section documents the methodological grounding of the protocol, the criteria that guided its development, and the planned empirical evaluation intended for future work.

### 3.1 Design Rationale

The protocol emerged from analysis of common breakdowns in high-ambiguity collaborative reasoning: premature convergence, asymmetric participation, drift into abstraction, unstructured hypothesis competition, and loss of decision traceability. We iteratively designed countermeasures—drift detection, falsification tokens, entropy-triggered perturbation pulses, airtime balancing, and immutable session logging—each mapped to specific failure modes.

### 3.2 Derivation Process

The protocol was refined through expert walkthroughs, structured scenario analysis, and multi-stage design sessions with a frontier large language model (Grok-4, xAI) acting as a high-bandwidth ideation partner. These sessions stress-tested candidate rules, identified ambiguity points, tuned thresholds, and ensured that each component could be executed reliably under time pressure.

### 3.3 Pilot Evaluation Plan

Although a full empirical evaluation has not yet been conducted, we outline a planned study to assess convergence quality, error reduction, and user experience. The design involves dyadic collaborators completing a set of twelve high-ambiguity technical tasks under either unstructured reasoning or Co-Koala v3.7. Dependent measures would include time-to-convergence, logical error counts, drift frequency, and perceived rigor. This plan is included for transparency and future replication.

## 4 Planned Evaluation

A controlled empirical evaluation has not yet been conducted. We describe the proposed study design intended to validate the protocol’s effects on convergence time, error reduction, and perceived rigor. This evaluation plan supports transparency, reproducibility, and structured future research using the protocol.

## 5 Discussion and Limitations

Co-Koala v3.7 introduces a structured, enforceable reasoning protocol intended to improve clarity and reduce failure modes during high-ambiguity analytical work. Its design incorporates principles from adversarial collaboration, metacognitive calibration, and structured decision-making. Several limitations remain. First, empirical validation has not yet occurred; the effectiveness of specific components—such as entropy-based perturbation triggers or airtime asymmetry controls—requires future study. Second, the protocol currently assumes dyadic interaction; multi-party adaptations require additional mechanisms and interface support. Third, the cognitive discipline required may impose a training overhead for new users.

### 5.1 Implications for Structured Collaboration

Even without completed empirical evaluation, the structure of Co-Koala v3.7 reveals opportunities for improving collaborative reasoning. The protocol formalizes coordination mechanisms—such as explicit challenge procedures, symmetry management, and drift detection—that are typically implicit in expert practice. Making these structures explicit may help collaborators maintain shared focus, avoid conversational stagnation, and surface uncertainty more reliably.

## 5.2 Design Opportunities for CSCW Systems

Co-Koala v3.7 suggests several design directions for CSCW systems supporting analytical or time-sensitive collaboration. Real-time indicators for drift, pacing, and airtime symmetry could provide actionable feedback during joint problem solving. Structured challenge mechanisms such as the Falsification Token offer lightweight alternatives to adversarial critique. Entropy-based triggers illustrate how computational metrics can detect conversational stagnation and introduce well-timed perturbations.

Future CSCW systems could incorporate these mechanisms through lightweight visual cues, shared timers, automated entropy dashboards, and semi-automated prompts that follow the protocol’s logic. Additionally, structured logging of decision processes—as specified in Co-Koala v3.7—may improve transparency, accountability, and post-hoc review.

## 5.3 Future Directions

Future work should examine the generalizability of the protocol to triads and small-group contexts, explore automated integration of drift and entropy detection into CSCW tools, and investigate how structured reasoning protocols interact with emerging human–AI collaboration workflows. Further research could analyze how participant experience, trust, and interpersonal dynamics shape protocol adoption and effectiveness.

# 6 Potential Extensions and Open Research Questions

Co-Koala v3.7 introduces a structured framework for managing dyadic analytical work, but it also opens several avenues for extension and deeper investigation. We highlight four directions that may guide both future implementations and research.

## 6.1 LLM-Assisted Automation

Although the current protocol assumes human execution, several components could be partially automated by a large language model or mixed-initiative system. An LLM could track entropy measures, detect drift, enforce timing rules, surface candidate falsification tests, or maintain the immutable ledger in real time. An open question is how such automation should balance assistance with preserving human agency and interpretability.

## 6.2 Integration with Formal Verification Tools

Several steps in the protocol—particularly the Falsification Token and hypothesis arbitration—could interface with formal verification systems such as Lean, Isabelle, or Z3. These tools could help validate internal consistency of claims, identify contradictions, or generate counterexamples. Exploring how formal verification might augment or constrain live collaborative reasoning remains an open research challenge.

## 6.3 Multi-Agent and Small-Group Variants

Co-Koala v3.7 is designed for dyadic use, but many real-world analytical settings involve small groups. A multi-agent extension may require role rotations (e.g., moderator, skeptic, synthesizer), additional symmetry controls, and expanded perturbation rules. Designing such variants while maintaining protocol coherence and preventing coordination overhead is a nontrivial open problem.

## 6.4 Affect- and Context-Aware Adaptation

Human–human collaboration is influenced by emotional load, interpersonal tension, and cognitive fatigue. Future implementations could incorporate sentiment analysis, prosodic cues, or physiological indicators to automatically adjust intensity, trigger perturbations, or suggest recovery intervals. Determining which affective signals are useful and how they should modulate the protocol is an important area for further study.

## 6.5 Open Research Questions

**Quantifying Entropy and Drift.** How should “entropy” and “drift” be reliably operationalized? Potential methods include semantic similarity decay, topic modeling, or divergence from the initial problem statement.

**Protocol Violations and Recovery.** What mechanisms should exist when one participant persistently violates protocol constraints? Would override clauses or structured “exit to creativity” pathways prevent stagnation without undermining rigor?

**Evaluation Metrics for Future Studies.** A planned empirical evaluation aims to assess metrics such as solution quality, time-to-convergence, logical error rate, drift frequency, and participant confidence. How these metrics interact and which are most diagnostic remains open.

**Lightweight Variants.** Is a reduced version of Co-Koala appropriate for low-stakes ideation or early-stage brainstorming? Determining the minimum viable subset of rules for casual use could broaden applicability while preserving core benefits.

These questions highlight the broader research landscape opened by Co-Koala v3.7 and motivate continued work on structured reasoning protocols for human–human and human–AI collaboration.

# 7 Critical Questions, Potential Extensions, and Risks

In addition to the opportunities outlined in the previous section, Co-Koala v3.7 raises several deeper questions related to implementation, threshold justification, human factors, and the potential limits of structured reasoning. We outline these to support future research and transparent examination of the protocol’s assumptions.

## 7.1 Implementation and Tooling

A natural extension of the protocol is integration into collaborative software environments. Embedding Co-Koala within a shared IDE, chat platform, or mixed-initiative workspace could enable automated dashboards for entropy, drift frequency, and airtime symmetry. Another direction is voice-based collaboration: real-time speech-to-text combined with prosodic analysis or sentiment detection may enrich drift detection and energy management, especially in high-intensity analytical sessions. Understanding which computational signals are most reliable and least intrusive remains an open design challenge.

## 7.2 AI–AI and Multi-Agent Adaptation

Although designed for human dyads, Co-Koala v3.7 may also be relevant for evaluating or improving multi-agent LLM reasoning. Two or more LLMs operating under Co-Koala constraints—drift detection, structured falsification, perturbation pulses, and symmetry enforcement—could form a benchmark for assessing convergence quality, logical consistency, or hypothesis diversity in agent-based systems. Whether such constraints enhance or hinder AI–AI collaboration is an open research question.

### 7.3 Threshold Justification

Several operational thresholds in the protocol—such as the 2.8 bits/token entropy trigger, the 60% airtime asymmetry limit, and the 90-minute hard cap—stem from iterative tuning, cognitive load considerations, and constraints identified during design sessions with Grok-4. These values are not yet empirically validated and should be treated as provisional approximations. Future studies may refine these thresholds based on observed drift patterns, fatigue curves, or behavioral data across different tasks and populations.

### 7.4 Human Factors and Training Requirements

Co-Koala v3.7 requires sustained attention and adherence to procedural rules. For new users, a lightweight “Co-Koala Lite” variant may lower cognitive overhead by reducing command complexity, relaxing timing constraints, or removing entropy triggers. Gamified onboarding—such as badges for drift-free intervals or consistent Falsification Token use—may further support adoption without compromising rigor. Understanding how training, incentives, and interface support influence user compliance is an important direction for future work.

### 7.5 Potential Weaknesses and Risks

Despite its benefits, the protocol carries several risks. Over-structuring may suppress intuitive leaps or creative divergence, especially in early-stage ideation. The cognitive overhead of managing timers, tokens, and drift gradients may compete with deep reasoning unless partially automated. Co-Koala also assumes cooperative intent: in adversarial, political, or strategically hostile environments, participants may exploit the protocol or manipulate its signals. Finally, the framework has not yet undergone empirical evaluation; until such studies are completed, Co-Koala v3.7 should be viewed as a formally specified but unvalidated method.

These questions and risks highlight the need for continued refinement, empirical study, and careful consideration of both human and computational factors in structured collaborative reasoning.

## 8 Conclusion

We introduced Co-Koala v3.7, a fully specified dyadic reasoning protocol designed to increase clarity, stability, and accountability in complex analytical collaboration. The protocol integrates preflight verification, drift detection, falsification mechanisms, perturbation triggers, airtime balancing, and immutable logging into a unified operational framework. While empirical validation is planned rather than completed, the contribution lies in the protocol itself: a reproducible, theoretically grounded method for structuring high-ambiguity reasoning between humans or between humans and AI systems. We encourage future work that evaluates its performance, extends it to multi-party settings, and explores its integration into collaborative platforms and decision-support tools.

## A Full Co-Koala v3.7 Specification

### A.1 Preflight Checklist

1. **Problem Statement:** Summarize the task in  $\leq 12$  words.
2. **Fatigue Assessment:** Each participant reports a 0–10 fatigue rating.
3. **Emotional Load:** Each participant reports yes/no for emotional load.

4. **Distraction Check:** Each participant reports yes/no for distractions.
5. **Worst-Case Consequence:** Each states the worst-case consequence they accept if wrong.
6. **Environment:** Confirm hydration and workspace stability.
7. **Operator Role:** Assign one participant as *Operator*.
8. **Scope:** Agree on maximum complexity (e.g.,  $\leq 3$  variables).

## A.2 Drift Gradient

1. **Micro-Slip:** 400–800 ms latency or mild abstraction.
2. **Major-Slip:**  $> 800$  ms latency or hedging.
3. **Drop:** Emotional override or derailment; immediate decouple.

## A.3 Falsification Token

Either participant may issue a “Falsify” command. The partner has 30 seconds to propose an experiment or data source that could disprove the current hypothesis within 30 days.

## A.4 Chaos Pulse

Triggered when information entropy of the last 50 utterances drops below 2.8 bits/token. Requires immediate 180° counter-proposal.

## A.5 Airtime Asymmetry Compensation

A real-time word tracker ensures no participant exceeds 60% of airtime in a rolling 5-minute window.

## A.6 Graveyard Metric

All killed hypotheses are logged with their final confidence scores to quantify rigor.

## A.7 Energy Limits

Mandatory checks at 24, 48, 72 minutes. Hard 90-minute cap.

## A.8 Engagement Tokens

- Engage: “Co-Koala engage”
- Kill switch: “Redline”
- Canonization: “Finalized” / “Confirmed”
- Re-entry: “Resume”



## A.9 Ledger Schema

- Session ID (UUID), timestamps
- Participants + roles
- Drift events
- Confidence score timeline
- Semantic summaries
- Graveyard entries
- Session hash + version

## References

- Center for Applied Rationality. Double crux. <https://www.lesswrong.com/tag/double-crux>, 2017.
- Asher Koriath. When are two heads better than one and why? *Science*, 336(6079):360–362, 2012.
- Barbara Mellers and Philip Tetlock. The psychology of intelligence analysis: Drivers of improved judgment. *Psychological Science*, 25(3):112–123, 2014.
- Gary Olson and Judith Olson. How collaboration fails: Common breakdowns and how to prevent them. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work*, pages 935–944, 2014.
- Jonas Schefström. Structuring collaborative argumentation in scientific problem solving. *Journal of Learning Analytics*, 6(2):22–39, 2019.
- Arjun Subramonian. How large language models collaborate with humans. *arXiv preprint arXiv:2311.15209*, 2023.
- Philip Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Broadway Books, 2019.