AI–Mediated Discipline in Scientific Ideation: A Transcript–Based Analysis of Hypothesis Generation in Mitochondrial Transport Studies

Abstract

We analyze four AI–mediated discussions conducted in a virtual laboratory where role–structured agents—a Principal Investigator, Method Architect, and Critical Reviewer—generated and refined hypotheses linking RNA–seq signatures to mitochondrial transport phenotypes. Extending the Zou–group's open–source framework, this system introduced a novel approach to hypothesis generation that disciplined idea formation through explicit observables, comparators, and acceptance criteria [1]. Hypotheses were required to specify observable endpoints, baseline comparators, and predefined conditions for acceptance or rejection, with critiques systematically converted into validation tests. Success criteria were framed explicitly in quantitative terms: improvement in area under the precision–recall curve (AUPRC) over gene–set baselines, with calibration bounded by an expected calibration error (ECE) ≤ 0.05 to guard against shortcut learning under class imbalance. From these machine–learning–generated transcripts, we show that hypotheses concerning MFN2 advanced because they combined mechanistic plausibility with stable observables, while motility–proxy claims, such as those tied to rotenone, were held back until viability assays and refined definitions could stabilize the phenotype. TRAK1 advanced only conditionally, gated by replication. We conclude that the primary contribution of AI mediation lies in its ability to structure ideation through precommitment, adversarialization, and transparent ledgering—thereby yielding robust, information–dense advances.

1 Introduction

Large language models (LLMs) are often introduced to biology as accelerators of scale: faster literature reviews, larger pools of candidate hypotheses, or increased throughput in ideation. Our analysis suggests that their most consequential effect may instead be discipline. The system we study is not a generator of unbounded novelty but a structured, AI–mediated "virtual laboratory" in which multiple agents—a Principal Investigator (PI), Method Architect, and Critical Reviewer—engage in role–specific dialogue governed by explicit agenda rules. These rules require that ideas be framed against baselines, that confidence levels be assigned to every decision, that disagreements be surfaced as identifiable cruxes rather than left implicit, and that convergence be recorded in a decision ledger.

The system, adapted from the Zou group's open–source repository and extended here for mitochondrial transport studies, is intentionally architectural (MIT License, commit from February 2025). Each agent embodies a complementary function in scientific reasoning. The PI introduces falsifiable hypotheses tied to observable endpoints and decisive acceptance criteria. The Method Architect translates these into specifications: dataset assumptions, baseline comparators, split policies, and ablation plans. The Critical Reviewer acts as adversary, probing fragile assumptions, registering risks, and demanding tests that could overturn the hypothesis if failed. Together, these roles prevent the conversational drift of brainstorming and instead scaffold inquiry into a preregistration–like format.

From the opening prompts, hypotheses are treated not as exploratory suggestions but as testable commitments. Each is paired with measurable endpoints, decisive thresholds, and designated owners of next steps. For instance, one early framing proposed that MFN2 knockdown alters distal mitochondrial density in neurites, immediately binding the claim to explicit performance criteria and calibration checks. Parallel formulations were made for the motile fraction and soma–to–neurite ratios. These early entries illustrate the broader pattern: AI mediation does not replace scientific judgment but embeds it in a structure that converts every proposal into a claim, every critique into a stress test, and every convergence into an auditable decision.

2 Analytic Framework and Coding Scheme

Our analysis is grounded in four structured sessions, each of which enforced role–specific logging by the Principal Investigator (PI), Method Architect, and Critical Reviewer. Agenda rules required every participant to adopt a baseline–first stance, assign confidence levels to each decision, and surface disagreements explicitly as cruxes rather than as passing objections. The PI framed hypotheses with observable endpoints and decisive acceptance criteria. The Method Architect converted critique into precise specifications and ablation plans. The Critical Reviewer recorded weaknesses, risks, and adversarial tests. This design meant that hypotheses could not advance without baselines and thresholds, critiques had to be paired with stress tests, and convergence was never tacit but had to be entered into a decision ledger with owners and residual risks.

To analyze the resulting corpus, we developed a coding scheme that sorted transcript segments into four categories of reasoning signals: hypothesis statements, critiques, refinements, and decisions. Coding was guided by explicit textual markers. For example, when the PI stated, "Hypothesis 1: MFN2 knockdown alters distal mitochondrial density in neurites, detectable via RNA–seq signatures... Primary metric: AUPRC +0.10 over baseline B2; calibration: ECE ≤ 0.05" (Trial 1), the segment was classified as a hypothesis statement because it introduced a falsifiable claim tied to metrics and thresholds. In contrast, when the Critical Reviewer wrote, "Viability assays and refined phenotype definitions must precede any trust assignment" in response to the rotenone line (Trial 2), this was coded as a critique because it identified a weakness and specified conditions under which the hypothesis would fail. Similarly, the Method Architect's proposal for "paired samples under MFN2 knockdown, leave–one–intervention–out (LOIO) splits, and a minimal ablation set including shuffle–labels, negative wells, batch–drop, and gene–set drop" was classified as a refinement, since it translated critique into operational requirements. Finally, statements such as "Proceed with batch correction using ComBat or SVA, validated by PCA plots, silhouette scores, and differential expression analysis... Confidence: 4" (Trial 2) were coded as decisions, as they recorded outcomes, confidence levels, and assigned owners.

Coding proceeded in two stages. First, one analyst coded all sessions independently, guided by the role headers in the transcripts (PI, Method Architect, Critical Reviewer). Second, a validation step re–coded 20% of the material after a two–week interval. Agreement between the two passes was 94%. Discrepancies were mostly borderline cases between critiques and refinements; these were resolved by adopting a stricter rule: only segments that specified a concrete change to baselines, metrics, or validation procedures were classified as refinements, while all other challenge–oriented interventions were coded as critiques. This procedure ensured reproducibility and conservative application of categories.

The coded signals revealed a temporal gradient: early discussions emphasized hypothesis articulation and critique, while later sessions focused on refining experimental specifications and ledgering decisions. This dynamic was evident in the progression from speculative framing toward convergent decision–making, as hypotheses were clarified, critiques were operationalized into tests,

and acceptance criteria were fixed in place. The comparative coding further highlighted differences among the central hypotheses—MFN2, rotenone, TRAK1, and RHOT1—across dimensions of novelty, mechanistic plausibility, RNA–seq observability, and resilience to confounders. MFN2 advanced because it met all of these conditions, while the others lagged until issues of observability, replication, or line heterogeneity were resolved. These descriptive profiles are not statistical claims but structured summaries that make the reasoning process visible and show how critiques functioned as epistemic tests.

3 Hypothesis Formation within a Falsification–First Framework

Idea generation in these sessions did not proceed as open–ended brainstorming but as the construction of falsifiable propositions tied to explicit observables, baselines, and acceptance thresholds. In the opening frame, the PI declared that MFN2 knockdown would alter distal mitochondrial density in neurites, and immediately bound the claim to predefined criteria for success. Comparable structures were set for the motile fraction and the soma–to–neurite ratio (Trial 1).

The baseline ladder was explicitly operationalized: B0, a constant predictor; B1, a minimal dose–time or line–only model; and B2, a gene–set logistic regression seeded by prior biological knowledge (Trial 1). Progress was defined not as an absolute score but as improvement over B2, with calibration constrained to prevent spurious gains under class imbalance [2]. These dual gates—relative performance and calibration—ensured that success could not be claimed by clearing trivial benchmarks or by overfitting unstable signals [3]. Together, the baselines encoded distinct epistemic stances—agnostic (B0), minimal (B1), and biologically scaffolded (B2)—which acted as guardrails against shortcut learning.

Risk registration occurred at the same moment as hypothesis declaration rather than being deferred to a later "threats" section. For MFN2, the PI and Critical Reviewer jointly flagged batch imbalance and weak feature pairing as immediate vulnerabilities, which in turn committed the team to batch–aware splits and calibration checks downstream (Trial 1). For motility, the record was sharper: "labels are noisy; viability can masquerade as transport" (Trial 2), a warning that prefigured later demands for viability assays and refined phenotype definitions before trust could be assigned. TRAK1 was explicitly tagged as fragile on replication grounds—"insufficient n, line heterogeneity"—which elevated leave–one–line–out (LOLO) validation and replication as gating conditions rather than optional checks (Trial 3). These annotations mattered analytically because they transformed generic caveats into design constraints: each identified risk was coupled to a required test or split policy that, if failed, would veto advancement.

The observables themselves were also specified concretely, closing the common loophole of post–hoc reinterpretation. Measures such as "distal mitochondrial density per neurite" and "soma–to–neurite mitochondrial ratio" were operationalizable endpoints directly tied to measurement channels and evaluation artifacts. The system further enforced continuity with prior work through Intervention–Driven Study (IDS) cards: MFN2 was mapped to IDS–EXP–01, rotenone to IDS–EXP–02, and TRAK1 to IDS–EXP–03 (Trial 1). This mapping was not decorative; it required that new claims be evaluated against their strongest prior instantiation, making novelty insufficient unless paired with superior, well–calibrated observables under the agreed baselines.

In effect, the combination of relative performance gates, calibration ceilings, and IDS lineage created an epistemic contract: proposals were born with the criteria by which they could be rejected, and prior art was fixed as the floor they had to clear. That contract bent ideation away from impressionistic plausibility and toward disciplined falsification at inception.

4 Adversarial Testing of Proxy Definitions and Batch Correction

Once hypotheses were scaffolded, the discourse shifted toward adversarial critique, where weaknesses were exposed and translated into operational tests. Two recurring issues defined the crux of these debates. The first was whether transcriptome–only surrogates could faithfully capture motility–linked phenotypes. The second was whether batch correction could remove spurious artifacts without erasing true biological variance.

On batch correction, consensus emerged early that adjustments were necessary, but not in a perfunctory way. The team specifically considered ComBat [4], an empirical Bayes framework for harmonizing expression data across batches, and Surrogate Variable Analysis (SVA) [5], which estimates hidden sources of variation and incorporates them as covariates in downstream models. Both methods are widely used in transcriptomic analysis, but the transcripts show that the group insisted they be validated rigorously. Checks such as principal component analysis (PCA) and silhouette clustering were judged insufficient on their own; additional differential expression analyses before and after correction were required to verify that genuine biology was preserved. This specification was elevated into the decision ledger: "Proceed with batch correction using ComBat or SVA... Confidence: 4," with implementation assigned to named owners and residual risks acknowledged as over–correction or subtle missed effects.

The Method Architect systematically operationalized these critiques into specifications that treated risks as potential failure modes. For the MFN2 line, paired samples were required, evaluated under LOIO splits. Models were restricted to ridge and lasso regressions on fixed hyperparameter grids to avoid opportunistic tuning. A minimal ablation set—including shuffled labels, negative wells, batch drops, and gene–set drops—was preregistered, with the explicit rule that failure on any ablation would flip the decision (Trial 1). In this way, critique was translated into ablation, and ablation into decisive criteria for advancement.

The fragility of motility proxies became an even sharper point of debate. The decision card for rotenone specified a decisive test of "reduction greater than 20% in motile fraction," but this threshold was immediately qualified by concerns that apparent reductions might simply reflect loss of cell viability rather than true changes in transport [8]. The Critical Reviewer emphasized that viability assays and refined phenotype definitions were non–negotiable prerequisites before trust could be placed in such hypotheses (Trial 1). Here, adversarial reasoning made explicit that the very definition of the phenotype was unstable, and that stability would have to be secured through additional assays before claims could move forward.

5 Structured Ledgering of Decisions

In the transcripts, convergence was not inferred from tone or tacit alignment but codified in decision cards that functioned like living preregistrations. Each entry carried explicit headers for evidence, risks, confidence, and ownership, making decisions both auditable and actionable. For example, the Trial 3 batch–correction entry mandated use of ComBat or SVA, validated by PCA plots, silhouette scores, and differential expression analysis (Trial 3). Responsibility for implementation was assigned to the Method Architect, while validation and risk monitoring were delegated to other roles. This division of labor prevented authority from concentrating in one individual and embedded accountability directly into the record.

The same ledger logic extended across other hypotheses. In Trial 1, the group resolved: "Keep the current approach with modifications for validation and batch correction". This was not a retreat but a dialectical stance—the idea survived provisionally, contingent on additional validation. For motility, the rotenone decision card required that a >20% reduction in motile fraction would only count

if paired viability assays and refined phenotype definitions were satisfied, embedding definitional stability as a prerequisite (Trial 2). TRAK1's ledger entries similarly conditioned advancement on expanded replication and LOLO designs, elevating replication from a recommendation to a formal requirement.

The cumulative effect of this ledger mentality was to make convergence visible, conditional, and reviewable. Agreement was no longer an impressionistic "sense of the room" but a recorded artifact backed by confidence scores, risk registers, and assigned owners. These decision logs created an auditable trail that could later be revisited to test whether hypotheses advanced under conditions of rigor or slipped forward without sufficient validation. In doing so, the system converted the fragile process of agreement into a transparent decision architecture—one designed to support replication, external audit, and downstream accountability.

6 Comparative Evaluation of Hypotheses

A comparative reading of the central hypotheses shows clear divergence in their readiness for advancement, as reflected in the transcripts. MFN2 consistently emerged as the most decision–ready. Its claim—changes in distal mitochondrial density under MFN2 knockdown—was both mechanistically plausible and tied to well–specified observable endpoints (Trial 1) [9]. The Method Architect further linked MFN2 to a rigorous ablation panel (shuffle–labels, negative wells, batch/geneset drops), with failures explicitly defined as decision–flipping. These layered safeguards meant that MFN2 was anchored not only to observability but also to a preregistered validation plan, positioning it as the line with the greatest epistemic resilience.

By contrast, rotenone and RHOT1 remained fragile because their acceptance hinged on unstable proxies. The rotenone line depended on a >20% reduction in motile fraction, but the Critical Reviewer repeatedly warned that such reductions could reflect viability loss rather than genuine transport effects (Trial 2 & 4) [8, 10]. Advancement was therefore conditioned on paired viability assays and refined phenotype definitions—gates that had not yet been cleared. RHOT1 suffered from a similar fragility: while mechanistically plausible as a transport regulator, its connection to transcriptomic signatures was judged weak, leaving it in a speculative state without decisive observables (Trial 2 & 4).

TRAK1 occupied an intermediate position. It was mechanistically strong—anchored to prior knowledge of kinesin/dynein adaptor functions—and its observable endpoint, the soma–to–neurite mitochondrial ratio, was clear (Trial 3) [11–13]. Yet the transcripts repeatedly stressed line heterogeneity and low replication as gating risks, requiring LOLO validation and expanded cohorts before advancement (Trial 3). TRAK1 was thus limited not by definitional instability, as with motility, but by external validity constraints.

These contrasts can be summarized as follows: MFN2 leads on observability and confounder resilience, underpinned by ablation–anchored validation. Rotenone and RHOT1 stall on proxy fragility, awaiting definitional and assay stability. TRAK1 scores strongly on observability and mechanism but remains gated by replication. The comparative pattern underscores how the system differentiated rigor across lines: MFN2 advanced under explicit safeguards, motility proxies were paused until stabilized, and TRAK1 was held behind replication thresholds.

7 Structural Mechanisms of AI–Mediated Inquiry

Across discussions, a recurrent pattern emerged: the value of AI mediation lay less in generating novelty and more in enforcing epistemic norms. Three mechanisms structured this discipline. First, hypotheses were precommitted to observables, baselines, and thresholds, preventing drift into

unfalsifiable speculation. Second, critiques were systematically converted into operational ablations or validation gates, ensuring that risks were tested rather than noted and ignored. Third, convergence was ledgered, recorded not as tacit consensus but as auditable decision entries. Together, these mechanisms increased the likelihood that advancing ideas were both mechanistically meaningful and empirically tractable.

An illustrative case is the Trial 3 synthesis of the batch–correction debate. Here, a general worry that batch adjustment might erase true biology was reframed as a formal requirement: PCA and silhouette validation [6] had to be complemented by differential expression checks. What began as a diffuse concern was converted into a concrete acceptance gate, exemplifying how AI mediation enforced rigor by transforming uncertainty into explicit, testable conditions.

8 Limitations and Validity Considerations

It is important to emphasize that our claims concern the process of ideation, not the empirical truth of MFN2, rotenone, or TRAK1 effects. Motility definitions remain intrinsically unstable: calibration may narrow error bounds, but shifting operational criteria and residual label noise introduce variability that no transcript–based refinement can fully resolve without independent functional assays. Batch correction, even when carefully parameterized, carries the unavoidable risk of attenuating or erasing genuine biological signals; acknowledging this limitation does not prevent it [7].

9 Conclusion

Our transcript–grounded analysis supports a precise claim: LLM mediation improved the quality of ideation not by inflating the number of ideas but by enforcing epistemic discipline at the point of invention. Three mechanisms did this work. First, hypotheses were precommitted to observable endpoints, baselines, and acceptance gates, constraining them to testable form. Second, critiques were adversarialized into ablations and validation checks whose failure was preregistered as decision–flipping. Third, convergence was ledgered, recorded with evidence, risks, confidence, and ownership. These mechanisms did not operate as post–hoc annotations but actively shaped the space of admissible claims in real time.

Cross–line comparisons made this structure visible. MFN2 advanced because it satisfied all three conditions: mechanistic plausibility, crisp observables (distal mitochondrial density), and a validation plan robust to shortcut risks. Rotenone stalled not for lack of plausibility but because its defining proxy, motile fraction, was unstable and confounded by viability; until assays and definitions stabilized, the ledger withheld trust. TRAK1 occupied an intermediate position: mechanistically sound and observable but constrained by replication limits, and therefore gated behind LOLO policies and expanded cohorts. Readiness, in this system, was earned through observability, resilience to confounders, and institutionalized accountability—not through narrative plausibility alone.

This discipline carries risks of its own. Constraining ideas to metric–gated frames can induce Goodhart pressure (optimizing ΔAUPRC/ECE at the expense of biological validity), prematurely stabilize fragile proxies, or lock norms once criteria are written. The transcripts themselves anticipated these pitfalls: batch correction required pre–and post–differential expression checks to guard against signal loss, and motility claims were made conditional on orthogonal assays. From these precedents, we generalize three safeguards: (1) dynamic gates that can be revised as assumptions shift, (2) orthogonal validation through independent assays or replication across lines, and (3) rotation of roles and red–teaming to prevent ossification of a single stance.

The framework also yields testable predictions for future deployments. If discipline is doing

real work, we should observe fewer hypotheses admitted to active testing but higher replication rates among those that pass; shorter time–to–decision for well–instrumented hypotheses and longer, conditional timelines for proxy–fragile ones; and ablation–triggered reversals recorded explicitly in the ledger rather than silent pivots. These constitute measurable auditables for evaluating AI–mediated science.

We therefore propose reading the LLM not as an idea generator but as a protocol amplifier and institutional memory that implements a decision architecture of precommitment → adversarialization → ledgering. When made public through machine–readable decision logs, preregistered gates, and replication policies, this architecture aligns naturally with open–science norms and makes reasoning inspectable. The result is fewer but higher–integrity advances: hypotheses that are testable, calibrated, and accountable. That, in our corpus, is the real contribution of AI mediation—it exchanges quantity for rigor and transforms convergence into an auditable artifact capable of withstanding scientific scrutiny.

Reproducibility, Responsibility, and Autonomy Disclosures

Research Autonomy and Human Oversight

The AI–mediated framework required role–specific logging and confidence ratings at every step, but human agency remained central throughout. Transcripts specify that the Principal Investigator convened sessions and framed prompts, setting the scientific agenda and defining what counted as a legitimate hypothesis. The Method Architect translated these prompts and critiques into explicit specifications—datasets, baselines, ablations, and validation gates—while also editing or refining prompts to ensure clarity and feasibility. The Critical Reviewer authored adversarial tests and flagged risks, but humans collectively determined which critiques were escalated into ledger entries.

In addition, this manuscript itself was drafted in collaboration with GPT–5, which assisted with writing, editing, and refinement while human authors retained control over content, framing, and final interpretation. The transcripts analyzed here were generated through our custom virtual lab setup, in which role–structured GPT–5 agents (Principal Investigator, Method Architect, Critical Reviewer) conducted discussions under human–defined prompts, goals, and evaluation gates. This combination of automated transcript generation and human oversight ensured both consistency in logging and fidelity to scientific norms.

Beyond role assignments, human participants also set goals for the discussions and assessed the viability of outputs. This included judging whether transcript–generated observables were biologically meaningful, whether acceptance criteria were realistic, and whether proposed validation strategies would yield interpretable results. Humans further evaluated the quality of transcript material itself, distinguishing between reasoning that could be carried forward and reasoning that was speculative, incomplete, or unconvincing. Human oversight was active and continuous, guiding prompts, steering goals, and adjudicating the credibility of proposed designs.

Responsible AI and Risk Disclosure

The transcripts identify three technical risks: batch effects that distort signals, phenotype proxies for motility that conflate viability with transport, and information shortcuts that undermine interpretability. Mitigations were explicitly preregistered: tripartite validation of batch correction with PCA, silhouette, and differential expression; viability assays and refined phenotype definitions for motility; leakage–resistant splits; and ablation panels whose failure mandated rejection.

Beyond these technical safeguards, the system introduces epistemic risks. Formalizing every hypothesis with metrics, baselines, and ablations enforces discipline but risks over–constraining inquiry, privileging prior framings, and codifying biases into decision ledgers. AI scaffolding can also create asymmetries of authority: once a criterion is ledgered, it acquires normative weight regardless of whether its assumptions remain contested. Human accountability is preserved through explicit roles—the PI proposing hypotheses, the Method Architect specifying tests, and the Critical Reviewer surfacing risks—but responsible use requires recognition that procedural rigor does not guarantee epistemic validity.

Beyond methodological safeguards, the broader impact of this work lies in demonstrating how AI can enforce transparency and reproducibility at the ideation stage, potentially raising the epistemic standards of collaborative science. At the same time, enforcing discipline too rigidly may suppress exploratory creativity or entrench flawed priors, so responsible deployment requires balancing constraint with openness. By situating AI as a scaffold rather than a replacement for scientific judgment, our approach aims to amplify accountability and transparency without displacing human expertise or creativity. We affirm that this work adheres to the NeurIPS Code of Ethics.

Reproducibility

The transcripts codify reproducibility through structural safeguards rather than post hoc claims. Immutable holdouts—LOIO and LOLO splits—were mandated to prevent information leakage. Model exploration was constrained by fixed hyperparameter grids for ridge and lasso regressions, avoiding opportunistic tuning. Calibration was bounded by explicit thresholds, with ECE required to remain ≤ 0.05. Ablation panels, including shuffled labels, negative wells, batch drops, and gene–set removals, were preregistered as decisive: failure on any ablation necessitated rejection of the hypothesis. Our analysis drew exclusively on these transcripts, ensuring that claims about reproducibility reflect the constraints embedded in the deliberative process rather than external assumptions.

Limitations and Sensitive–Use Cautions

The decision logs themselves note unresolved risks: batch correction may erase biology despite validation, motility proxies remain unstable even under refined definitions, and over–correction or new biases can emerge from ablations not yet anticipated. These limitations mean that biological conclusions cannot be read directly from the transcript; they are conditional on acceptance criteria and must be re–validated experimentally. More broadly, AI–mediated discipline can prematurely stabilize fragile constructs, converting them into apparently robust criteria before their validity is established. For this reason, sensitive applications such as biomarker discovery require independent replication and orthogonal validation, and claims.

References

[1] Zou Group. virtual-lab GitHub repository. Available at: `https://github.com/zou-group/virtual-lab` (MIT License, commit from February 2025). Accessed August 28, 2025.

[2] T. Saito and M. Rehmsmeier. The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):e0118432, 2015.

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017. Also available at: arXiv:1706.04599.

[4] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[5] J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

[6] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[7] Y. Zhang, G. Parmigiani, and W. E. Johnson. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, 2(3):lqaa078, 2020.

[8] J. H. Won et al. Rotenone-induced impairment of mitochondrial electron transport chain complex I leads to ATP depletion and loss of mitochondrial membrane potential. *International Journal of Molecular Sciences*, 16(7):16610–16624, 2015.

[9] A. L. Misko, S. Jiang, I. Wegorzewska, J. Milbrandt, and R. H. Baloh. Mitofusin 2 is necessary for transport of axonal mitochondria and interacts with the Miro/Milton complex. *Journal of Neuroscience*, 30(12):4232–4240, 2010.

[10] W. J. H. Koopman et al. Partial complex I inhibition decreases mitochondrial motility and velocity. *Biochimica et Biophysica Acta (BBA) – Bioenergetics*, 1767(8):730–741, 2007.

[11] V. Henrichs et al. Mitochondria-adaptor TRAK1 promotes kinesin-1-driven transport in crowded environments. *Nature Communications*, 11:3123, 2020.

[12] M. van Spronsen et al. TRAK/Milton motor-adaptor proteins steer mitochondrial trafficking to axons and dendrites. *Neuron*, 77(3):485–502, 2013.

[13] E. E. Glater, L. J. Megeath, R. S. Stowers, and T. L. Schwarz. Axonal transport of mitochondria requires Milton to recruit kinesin heavy chains and is light-chain independent. *Journal of Cell Biology*, 173(4):545–557, 2006.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: Yes
   Justification: The paper's core claim is about disciplining ideation (not asserting biological truth). The abstract and Section 1 state scope and assumptions (AUPRC/ECE gates, role-structured agenda rules), and Sections 6 & 8 bound generalization and limitations.

2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: Yes
   Justification: Section 8 and the Disclosures discuss proxy instability, batch-correction risks, scope limits, and conditions for advancement; mitigations are also stated.

3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: N/A
   Justification: The paper presents no theorems or formal proofs; results are qualitative analyses of structured transcripts and decision logs.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper?
   Answer: Yes
   Justification: Reproducibility is addressed via preregistered agenda rules, baseline ladder (B0–B2), LOIO/LOLO splits, fixed hyperparameter grids, and decisive ablation panels; decision cards/ledger specify evidence, risks, and ownership.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main results?
   Answer: N/A
   Justification: No new dataset/model training is reported; core results are transcript- and protocol-based. The adapted virtual-lab framework is cited with MIT license and version.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details necessary to understand the results?
   Answer: N/A
   Justification: No training/evaluation experiments are reported; instead, Sections 2 & 5 specify decision architecture (roles, baselines, splits, ablations).

7. **Experiment statistical significance**
   Question: Does the paper report error bars or other appropriate information about statistical significance?
   Answer: N/A
   Justification: No quantitative benchmarking or statistical testing is presented.

8. **Experiments compute resources**
   Question: Does the paper provide sufficient information on compute resources needed to reproduce the experiments?
   Answer: N/A
   Justification: No compute-intensive experiments are reported; the contribution is a transcript-grounded analysis of an AI-mediated decision architecture.

9. **Code of ethics**
   Question: Does the research conform to the NeurIPS Code of Ethics?
   Answer: Yes
   Justification: The Responsible AI section affirms adherence to NeurIPS Code of Ethics and details risk awareness, human oversight, and mitigations.

10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts?
    Answer: Yes
    Justification: The paper discusses epistemic/organizational risks (over-constraint, ossification, Goodhart pressure) and mitigations (dynamic gates, orthogonal validation, role rotation/red-teaming), as well as positive impacts on transparency and auditability.

11. **Safeguards**
    Question: Does the paper describe safeguards for responsible release of data/models that could be misused?
    Answer: N/A
    Justification: No high-risk model or dataset is released; safeguards are not applicable.

12. **Licenses for existing assets**
    Question: Are the creators or original owners of assets used in the paper properly credited and are the licenses mentioned?
    Answer: Yes
    Justification: The adapted Zou-group virtual-lab repository is cited with URL, license (MIT), and version (commit from February 2025); usage respects the license.

13. **New assets**
    Question: Are new assets introduced in the paper well documented?
    Answer: N/A
    Justification: The paper releases no new datasets, models, or software artifacts.

14. **Crowdsourcing and research with human subjects**
    Question: For crowdsourcing experiments or research with human subjects, does the paper include instructions and details?
    Answer: N/A
    Justification: The study analyzes author-generated transcripts; no crowdsourcing or human-subjects data collection occurred.

15. **IRB approvals**
    Question: Does the paper describe participant risks and IRB approvals, if applicable?
    Answer: N/A
    Justification: No human-subjects research was conducted; IRB approval is not applicable.

16. **Declaration of LLM usage**
    Question: Does the paper describe the usage of LLMs if they are central to the methodology?
    Answer: Yes
    Justification: LLMs are central to the method (role-structured agents, agenda rules, adversarialization, ledgering); usage and safeguards are described in Sections 1–7 and in the Responsible AI section.