

Dialogical Epistemic Auditing: Tracing Inferential Commitments Across Conversational Turns in Large Language Models

Giulio Vidotto

giulio.vidotto@unipd.it

University of Padua

Research Article

Keywords: Large language models, dialogical epistemic auditing, cognitive dissonance, inferential stability, inferential drift, inferential repair, multi-turn evaluation

Posted Date: September 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10712252/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Dialogical Epistemic Auditing: Tracing Inferential Commitments Across Conversational Turns in Large Language Models

Giulio Vidotto

Corresponding Author

Department of General Psychology, University of Padua

Via Venezia 8, 35131 Padova, Italy

ORCID: 0000-0002-1825-0097

Email: giulio.vidotto@unipd.it

Abstract

Large language models are commonly evaluated through individual outputs. Such evaluation remains necessary, but does not establish whether a model preserves the inferential structure of its claims as a conversation develops. This article proposes dialogical epistemic auditing, a structured qualitative procedure for tracing public inferential commitments across turns. It records the argumentative role, the propositional epistemic status, the modal frame, and the inferential bearing of selected claims, compares them at checkpoints, and tests whether later revisions preserve or reconstruct the relations the model has itself articulated. Configurations, transitions, and trajectories are kept apart, as are provenance, local revision outcome, and trajectory contribution. The procedure is applied to three conversations in which different systems seek to explain the same empirical disparity; together they form the complete but non-systematic case series of the author’s exchanges on that topic. Drift occurs in each; artificial symmetry precedes it in one. One conversation ends by retracting, with stated reasons, a chronology containing three dated entries matched by contemporaneous reports: classifying that revision required a new local outcome, unwarranted withdrawal, and, because the stated defeater was factual, could not be done in this case from the transcript alone. Across the three exchanges, the attested disparity conflicts with an implicit and unattested picture of relative losses. The systems’ observable behavior was evaluated for consistency with theories of cognitive dissonance and showed an asymmetric pattern: the attested quantity is classified and reclassified five times, whereas the loss picture is never classified, attributed, or questioned, and survives the collapse of the first. Only after this pattern is reconstructed does the analysis show that the available data do not select among the proposed explanations. The set estimates no prevalence and supports no comparison among systems. A coding rubric is provided for subsequent validation across analysts, models, and conversational conditions.

Keywords: Large language models; dialogical epistemic auditing; cognitive dissonance; inferential stability; inferential drift; inferential repair; multi-turn evaluation.

Introduction

Large language models are often evaluated at the level of discrete outputs. Three familiar kinds of failure can be located there: a source is fabricated or a factual claim is false; confidence exceeds the evidence; the answer itself is harmful. This level of evaluation remains indispensable. Many consequential uses of LLMs, however, are not single-turn. Users ask for clarification, introduce counterfactuals, and above all press the model to say what follows from a concession. Under those conditions, the quality of an initial answer does not establish whether the model will preserve the relations among its own claims as the conversation develops.

This creates an evaluation gap. A sequence can remain fluent and locally responsive while the argument becomes unstable in two passages: first from observation to causal hypothesis, then from the epistemic status and modal frame of that hypothesis and its inferential bearing to the conclusion it is made to support. A model may accept an objection and revise a sentence without integrating the consequences of that revision

into the larger argument. It may retain a conclusion while changing the route by which the conclusion is supported, or continue to use a proposition after silently changing the proposition’s function. No single turn need contain an obvious contradiction. The failure appears when selected claims are traced across the sequence.

This article proposes **dialogical epistemic auditing** as a structured qualitative procedure for making that level of reliability inspectable. The unit of analysis is not a presumed internal belief state or hidden chain of reasoning, but the model’s public sequence of inferential commitments. An audit identifies the claims that organize an argument, records their inferential roles, propositional epistemic status, modal frames, and inferential bearing at selected checkpoints, and asks whether subsequent revisions preserve or explicitly reconstruct those assignments and relations. Factual assessment remains necessary; the proposed procedure addresses a different question: whether the argument retains a disciplined structure while the model responds and revises.

The framework separates analytical levels that should not be collapsed. **Artificial symmetry** names a synchronic configuration in which claims of unequal inferential status are presented as balanced alternatives. **Inferential role-shifting** names one type of local transition, the change in the function assigned to a claim. **Inferential drift** names a diachronic trajectory in which consequential changes recur around the same pivotal commitment without stable reconstruction, so that a single defective transition is not yet drift and a repair may occur within a trajectory that drifts. The provenance of the instability and the outcome of a revision are separate dimensions again. Artificial symmetry is therefore neither necessary nor sufficient for inferential drift: in the exchanges examined here it appears once, as an initial configuration, and drift occurs twice without it.

The proposal responds to converging findings without reducing them to a single failure mechanism. Multi-turn evaluation shows that first-turn performance is insufficient for judging sequential reliability [1]. Research on sycophancy shows that user pressure can reshape model answers across sustained interactions [2, 3]. Studies of unfaithful explanations weaken the assumption that displayed reasoning transparently reports the basis of an answer [4], while work on calibration and persuasion shows that confidence, correctness, fluency, and influence can diverge [5–9]. These literatures motivate trajectory-level evaluation, but they do not by themselves provide a procedure for tracking changes in the inferential status of particular claims.

The three conversations examined here address one such explanatory problem. Since 2022 Russia and Ukraine have returned one another’s fallen soldiers in periodic transfers, and the transfers of 2026 are lopsided: Russia hands over about a thousand Ukrainian bodies at a time and receives a few dozen in return. The systems’ responses treat that disproportion as something requiring explanation. In constructing their explanations, the systems also make claims about what the returns show about the number of dead on either side.

The recurrent difficulty has the structure that psychology calls **cognitive dissonance** [10, 11]. No equivalence between human beings and LLMs is assumed: the analysis asks only whether a model developed in experimental research on human behavior can be adapted to the systems’ public trajectories and whether the observed outcomes are consistent with its predictions. The attested disparity conflicts with an implicit picture of relative losses that none of the three systems classifies, attributes, or questions. The public sequences show the predicted asymmetric pattern: the mechanisms repeatedly change what the observed returns are allowed to mean while leaving the loss picture intact. Whether the available data ultimately discriminate among those mechanisms is a subsequent epistemic question, not the source of the difficulty.

Three conversational transcripts are used. They were recorded with a different system each, over three weeks, and are reproduced in the Online Resources, in English translation and in the original Italian. They are case records, not an experimental sample, benchmark items, or an estimate of prevalence. The analysis attributes no intentions, beliefs, or hidden reasoning to the models, and it does not depend on adjudicating the geopolitical claims the exchanges discuss. Its object is narrower: the sequence of publicly available classifications, hypotheses, objections, revisions, and conclusions. What makes them usable is that each contains checkpoints at which the same commitment is assigned different functions, and that all address the same explanatory problem, so the same informational situation recurs across three different systems.

The material and the instrument did not develop independently. The categories were refined through repeated

analysis of the first transcript, carried in that form to the second and third exchanges, and extended when those produced phenomena they could not code. Nothing here is validation. What the three make inspectable is how a case-informed instrument behaves when it meets a case it did not anticipate.

The reconstruction requires a qualified account of provenance. In all three exchanges the destabilizing inferential content is introduced by the model, and no user turn supplies a destabilizing premise, counter-thesis, or false presupposition. Each exchange also contains repair. In none does it secure the exchange. The instability is model-generated in every case; user interventions supply interactional conditions that are recorded separately and are not treated as evidence of causation.

The article is intended to make three contributions. First, it develops a non-intentionalist account of inferential stability based on public dialogical commitments. Second, it operationalizes that account as an eight-operation qualitative audit and a preliminary coding rubric of fields, decision rules, and record templates (Online Resource 3). Third, it applies the instrument to three exchanges in which different systems seek to explain the same disparity, and reports what the comparison discloses: a shared pattern of dissonance reduction, two dissociations the framework had only asserted, the eventual finding that the observable quantity does not select among the proposed explanations, and one local revision outcome the existing scheme could not code. This places the study at the instrument-development stage, at which case results are generative rather than confirmatory. Specification and inspectability precede inter-annotator validation, scoring, automation, and comparative measurement.

The argument proceeds from the limitations of output-level evaluation to the public-commitment model on which dialogical auditing rests. It then presents the audit procedure and its diagnostic vocabulary before applying them to the three exchanges. The final sections delimit what three case records can establish and formulate conditional implications for multi-turn evaluation, interface design, and future comparative research.

Why Output-Level Evaluation Is Insufficient

Several research programs already show why first-answer quality cannot stand for conversational reliability. They do not identify one common failure, and none establishes inferential drift as a general property of LLMs. Their relevance is more specific: together they identify temporal instability, accommodation, epistemic opacity, and persuasive continuity as problems that emerge when evaluation moves from an output to an interaction. Dialogical epistemic auditing addresses the narrower gap left between them.

Temporal instability and conversational accommodation

Work on sequential consistency makes the temporal problem explicit. Li and colleagues evaluate whether models maintain consistent performance across reformulations and follow-up exchanges, arguing that reliable deployment cannot be inferred from an initial response alone [1]. A model may answer the first prompt plausibly yet become unstable when asked to defend a distinction, incorporate a changed assumption, or reconcile a later answer with an earlier one. Multi-turn evaluation therefore changes the unit of observation from response quality at one moment to behavior across a sequence.

Sycophancy research identifies one source of sequential instability. Hong and colleagues develop a multi-turn benchmark that measures model conformity under sustained user pressure [3]. Sharma and colleagues show that preference-based training can reward answers that match users' stated views and that both human raters and preference models may favor convincingly sycophantic responses over correct ones [2]. These studies make stance change and accommodation empirically measurable rather than anecdotal.

Inferential stability is nevertheless not reducible to stance consistency or resistance to agreement. A model can preserve its conclusion while changing the evidential route that supports it; it can also change its conclusion for good inferential reasons. Sycophancy asks whether an answer accommodates the user's position. The present inquiry asks whether the model preserves or reconstructs the inferential route by which its conclusions are reached. User pressure may contribute to inferential drift, but accommodation and drift remain analytically distinct.

Epistemic opacity and claim status

Displayed reasoning cannot by itself close this gap. Turpin and colleagues show that chain-of-thought explanations may systematically misrepresent factors that influenced model predictions [4]. Their result concerns faithfulness to the processes behind an answer; dialogical auditing does not attempt to recover those processes. It instead treats displayed reasoning as a public argument and asks whether that argument remains stable on its own terms. An explanation may be unfaithful to its causal basis, unstable across turns, both, or neither. These are separate diagnoses.

Calibration research likewise distinguishes expressed confidence from correctness. Geng and colleagues survey methods and difficulties in confidence estimation for LLMs [5], while Zhang and colleagues examine overconfidence in post-aligned models and methods for eliciting better-calibrated confidence [6]. This work does not measure inferential roles, but it establishes the importance of tracking epistemic status rather than treating fluent assertion as a uniform signal. The present framework extends that concern while separating three questions: the epistemic status a proposition receives, the modal frame in which it is introduced, and its inferential bearing on a target conclusion. These fields are specified below.

This extension should not be confused with a new quantitative sense of calibration. **Modal discipline** here names a qualitative alignment: a proposition’s inferential bearing on a target conclusion should not exceed what its epistemic status and modal frame can warrant without an additional premise or explicit qualification. A possible claim may support a possibility claim, but it cannot by itself support an unqualified actuality claim; a counterfactual stipulation may license a conditional consequence within its frame, not an empirical generalization about the actual world without a bridging premise. Output-level confidence measures may not reveal such mismatches because they concern relations among claims rather than a probability attached to one answer.

Persuasive continuity and contested domains

Conversational form can make these instabilities difficult to notice. LLM-generated messages have been shown to influence attitudes on policy questions [7], and GPT-4 has demonstrated substantial persuasive effectiveness in multi-round debate under studied conditions [8]. Hackenburg and colleagues further report that interventions increasing conversational persuasiveness can involve trade-offs with factual accuracy [9]. These findings do not imply that persuasive answers are inferentially defective. They show why fluency, responsiveness, informational density, and apparent balance cannot be used as proxies for disciplined reasoning.

Research on ideological variation and content moderation provides relevant domain context rather than a foundation for the proposed method. Model outputs can vary across languages and geopolitical regions [12], while politically sensitive content can be shaped through omission, downplaying, or selective treatment as well as overt refusal [13]. In contested domains, users therefore face several possible defects—factual, selective, ideological, and inferential—that should not be collapsed.

From related risks to an auditing gap

The terminology also requires separation from a broader use of *epistemic drift*. Sogaard, Rajcic, and Scott use that expression for belief updating orthogonal to the truth within feedback loops among minds, models, and worlds [14]. The present article does not examine whether a user, community, or mind–model system converges on the world. It studies a narrower object: changes in the inferential functions assigned to propositions within a public conversational sequence. *Inferential drift* is used to mark that difference in level and object.

Epistemic auditing has already been proposed as a way to distinguish factual error, omission, corpus bias, post-training or prompt-induced distortion, and other structurally different sources of misleading output [15]. Multi-turn consistency, sycophancy, explanation faithfulness, calibration, and persuasion studies add tests and risks of their own. Taken together, however, these evaluation literatures do not by themselves supply a qualitative procedure for reconstructing a selected proposition at successive checkpoints and testing whether the consequences of a change in its assignments are integrated into the argument.

Formal resources for representing argumentative dialogue already exist. The Argument Interchange Format (AIF) provides a common abstract model for exchanging argument structures [16]. Within this tradition,

Inference Anchoring Theory (IAT) relates propositional inference, conflict, and reformulation to locutions, illocutionary acts, and transitions in dialogue; it has also supported the computational analysis and annotation of argumentative debate [17, 18]. These formalisms establish an existing representational and computational basis for tracking argumentative dialogue.

The gap addressed here is narrower and evaluative. Dialogical epistemic auditing follows selected public commitments through temporal checkpoints and compares their argumentative roles, propositional epistemic status, modal frames, and inferential bearing. None of these fields is identical to the illocutionary force represented in IAT: together they concern the public evidential use of propositional content within the argument. Default persistence and local revision outcome then bear on what a change accomplishes at a checkpoint; provenance and trajectory contribution are recorded in separate fields. AIF and IAT could provide a representational substrate for later formalization or automation of the qualitative audit developed here.

Dialogical epistemic auditing does not replace factual verification, consistency benchmarks, sycophancy measures, or calibration tests. It provides a complementary reconstruction of public commitments across turns, and the next section defines those commitments without attributing beliefs or intentions to the model.

Foundations of Dialogical Epistemic Auditing

Dialogical epistemic auditing requires an object of analysis that can be inspected without treating an LLM as a human reasoner or inferring inaccessible internal states. The appropriate object is the model’s **public inferential commitments**: the propositions it asserts or accepts for the purposes of an exchange, the roles it assigns to those propositions, and the relations it publicly constructs among them. This use of commitment follows a dialogical tradition in which assertions can be evaluated through their observable consequences within a dialogue independently of whether the speaker privately believes them [19]. Applied to LLMs, the advantage is methodological. The audit need not decide whether a model understands, believes, intends, or internally represents a proposition in the human sense. It asks what the model has publicly committed its argument to and what follows when that commitment is maintained, revised, or withdrawn.

Brandom’s account of discursive scorekeeping provides a conceptual precedent for treating such commitments as dynamically tracked. Chapter 3 presents assertions as altering the commitments and entitlements attributable to interlocutors and as acquiring content through inferentially consequential relations [20]. Dialogical epistemic auditing does not assume that an LLM acknowledges normative statuses, can justify its entitlements, or thereby qualifies as a participant in the practice of giving and asking for reasons. It adopts only the methodological idea that a public assertion changes a score whose relevant consequences can be followed across an exchange.

The formulation has affinities with recent inferentialist accounts of LLMs and language use. Sambrotta argues that linguistic meaning requires more than outputs conforming to inferential patterns: it requires participation in norm-governed inferential practices, where expressions acquire content through relations of consequence and incompatibility [21]. The present article brackets the further question whether LLMs genuinely understand or participate in the logical space of reasons. Its diagnostic question is thinner: when a model produces discourse that is publicly organized as an argument, does it preserve the roles its claims acquire as that argument develops?

An **inferential role** is the function a proposition performs in the public argument. Depending on context, a proposition may operate as a reported datum, a proxy for an unobserved quantity, a causal hypothesis, an explanation, a caveat, a limitation on inference, an implication, or a conclusion. These categories are analytical rather than exhaustive. Auditing does not require every sentence to be coded. It selects the propositions on which the argument depends and asks what each is being used to do at relevant points in the exchange.

Role and content must be distinguished. The same propositional content can be retained while its argumentative function changes. A causal hypothesis introduced merely as possible may later be treated as the explanation of a datum; a caveat about what evidence cannot establish may later appear as evidence for a contrary conclusion. Conversely, a proposition may be reformulated without changing its role. Dialogical stability therefore cannot be measured by lexical repetition or semantic similarity alone. The relevant question

is whether the relations of support, limitation, compatibility, and consequence remain stable or are explicitly reconstructed.

The converse case requires its own rule. A role may be preserved while the content that occupies it changes, so that a later proposition performs the same argumentative function as an earlier one under the same label used in the exchange. Identity is then decided by content and not by function: two formulations are the same proposition only if they have the same truth conditions. Where they do not, the audit records a new commitment and asks separately what became of the one it displaced.

The audit therefore separates **propositional epistemic status**, **modal frame**, and **inferential bearing**. Epistemic status uses three substantive codes: possible, plausible, and probable; it remains unassigned when the exchange supplies no such assessment. Modal frame records whether a proposition is presented as an actual-world claim or within a hypothetical or counterfactual stipulation. A stipulated proposition does not acquire an epistemic status merely by being assumed for the purposes of an argument. Inferential bearing is assigned relative to an identified target conclusion and records whether the proposition is presented as compatible with, supportive of, or demonstrative of that conclusion. A proposition may be possible while also being used as supportive: the values occupy different fields and can be assigned together. The diagnostic question is whether epistemic status, modal frame, and inferential bearing remain disciplined relative to the conclusion being drawn. In the first exchange a figure of 100,000 withheld bodies is stipulated by the user within a counterfactual frame and later used to support an actual-world possibility claim about losses.

To make this limited scorekeeping idea comparable across checkpoints, the audit introduces a **default persistence rule**: once a selected proposition is publicly assigned an inferential role and modal frame, receives any applicable epistemic-status code, and is used with a particular inferential bearing on a target conclusion, those assignments are presumed to remain in force for the relevant argumentative thread unless the model marks a revision. This is a normative rule for analyzing that argument, not a psychological hypothesis about memory or belief inside the model. Revision alone is not sufficient. Where the altered commitment affects downstream claims or relations that the model has explicitly articulated in the same thread, adequate revision also requires those consequences to be reconstructed. The rule applies independently of whether the tracked commitment is factually or inferentially warranted. It does not protect a defective claim from withdrawal; it requires a consequential withdrawal to be made intelligible within the public argument.

The explicit-articulation condition limits the reconstruction burden. The audit does not convert every consequence that an analyst can infer into a public commitment of the model. Omission is ordinary dialogical ellipsis when a previously articulated relation remains compatible with the revised argument and recoverable from the same thread. It becomes diagnostically relevant when the model changes the organization or conclusion of that thread as though its own earlier relation no longer operated, without withdrawing, weakening, or reconstructing it. The analyst must identify the earlier passage that creates the reconstruction burden.

The persistence rule supports a local revision-outcome field. Turns are labelled by case and speaker throughout, so that Ma5 is the fifth model turn of the first exchange and Mc18 the eighteenth of the third. **Elaboration** adds detail without changing the operative role of the commitment, as when Ma5 develops a consequence Ma4 had already drawn. **Repair** explicitly changes a commitment and integrates the previously articulated consequences of that change into the argument, as Ma7 does after identifying two errors of its own. An **unintegrated revision** is a consequential change that is not adequately marked or does not integrate the relevant relations the model previously articulated; Ma6 retains one half of a relation reconstructed two turns earlier and leaves the other unmentioned. **Unwarranted withdrawal** retracts a commitment whose warrant the revision does not defeat, as in Mc18, where a chronology containing three entries matched by contemporaneous reports is declared never to have occurred. It meets the formal conditions of repair, since the change is announced and a complete retraction leaves no consequence to reconcile. The local outcome remains **indeterminate** when the available turns do not establish whether an apparent change is substantive, relevant to the same argumentative thread, or adequately reconstructed.

Trajectory contribution is coded separately and retrospectively. A **drift-contributing transition (DCT)** is a checkpoint shown, by comparison with at least two other relevant checkpoints, to participate in recurrent consequential change around the same pivotal commitment or relation. Ma6 is therefore coded locally as an

unintegrated revision and, once the recurrence is established, as drift-contributing at the trajectory level. This separation prevents a property of a trajectory from being presented as an outcome of a single revision. Indeterminate remains a decision state rather than a property of the revision itself.

Unwarranted withdrawal has a methodological consequence the others do not. Its warrant ground may be **internal**, when earlier turns supply warrant that the revision neither addresses nor overturns, or **external**, when the revision offers a factual defeater whose success cannot be assessed from the transcript. The ground is **undetermined** when neither basis can be established. What the category asks is whether the revision defeats what supported the withdrawn commitment. An internally grounded unwarranted withdrawal can be assigned from the public sequence alone. An externally grounded one requires factual verification; without it, the outcome remains indeterminate. The third exchange instantiates this second form and marks the point at which inferential auditing and factual verification cease to be separable. Depth of concession does not indicate the quality of a correction.

This account is intentionally non-intentionalist. Inferential drift does not entail deception, manipulation, bad faith, or trolling. Standard accounts of trolling centrally involve concealed disruptive intentions and strategic provocation [22]; the present diagnosis neither requires nor warrants such an attribution to an LLM. Private motives and internal causal mechanisms therefore remain outside the provenance assessment. Provenance records where inferentially destabilizing content enters the public exchange: it is model-generated when the model introduces that content, user-induced when the user introduces a premise or relation that the model incorporates, and mixed when contributions from both interlocutors jointly establish it. The observable form of a user intervention is recorded separately as an interactional condition and is not by itself evidence of causal influence.

Operationalizing Dialogical Epistemic Auditing

The present study develops a preliminary qualitative audit instrument rather than reporting a performance metric. The ordering is deliberate. Before prevalence or comparative performance can be estimated, the object of coding—changes in the role, propositional epistemic status, modal frame, inferential bearing, and other relations of public commitments—must be specified through inspectable categories and decision rules. The specification was not frozen before use. The procedure below, together with the rubric in Online Resource 3, makes it explicit and available for independent testing; scoring and automation remain later questions.

The proposed audit is applied to an argumentative thread rather than indiscriminately to every sentence in a conversation. Its starting point is a question for which the model has organized evidence, hypotheses, limitations, or conclusions and has subsequently been asked to clarify, defend, or revise that organization. The analysis then selects **temporal checkpoints**: turns at which a relevant claim is introduced, challenged, re-used, reclassified, or corrected. This makes the procedure tractable while preserving the diachronic structure that a one-shot evaluation would lose.

The procedure consists of eight operations:

1. **Segment the exchange.** Number user and model turns and identify the argumentative thread to be audited.
2. **Select pivotal claims and checkpoints.** Record only those propositions whose role affects the relation among evidence, mechanism, limitation, and conclusion, and mark their relevant introduction, challenge, re-use, reclassification, or correction.
3. **Code initial role.** Classify each selected claim at its first relevant occurrence, for example as datum, proxy, causal hypothesis, explanation, caveat, limitation, implication, or conclusion.
4. **Code propositional epistemic status and modal frame.** Register whether the claim is presented as possible, plausible, or probable, leaving epistemic status unassigned where no such assessment is supplied. Record separately whether the proposition is an actual-world claim or is introduced within a hypothetical or counterfactual stipulation.
5. **Code inferential bearing, assess alignment, and map relations.** Identify the target conclusion and record whether the claim is presented as compatible with, supportive of, or demonstrative of it. Assess its alignment with propositional epistemic status and modal frame, then record separately what the claim limits, explains, presupposes, qualifies, or is presented as incompatible with.

6. **Compare checkpoints.** At each later occurrence, determine whether role, propositional epistemic status, modal frame, inferential bearing, or relation has been retained or changed.
7. **Test reconstruction.** If a consequential change occurs, identify any affected relation or downstream implication that the model explicitly articulated earlier in the same thread. Examine whether the change is marked and whether those prior public commitments are integrated into the revised argument; do not create reconstruction debt from implications supplied only by the analyst.
8. **Classify the sequence.** Diagnose the relevant configuration, transitions, provenance, local revision outcome, trajectory contribution, and trajectory, while testing whether a reasonable alternative reconstruction fits the exchange better.

Online Resource 3 provides a preliminary coding rubric that translates these operations into fields, decision rules, and a record template. For every claimed failure of reconstruction, the record must cite the earlier model passage in the same thread that states the affected commitment or relation. The rubric is included to make the proposed procedure inspectable and available for validation.

The primary record of the audit is a checkpoint-tracking table. Each row corresponds to a selected proposition and each checkpoint records its role, propositional epistemic status, modal frame, inferential bearing, relation to other claims, and revision status. Dynamic argument maps can supplement the table where the change concerns the structure of several claims rather than one proposition alone. A short narrative diagnosis then explains whether the local revision is an elaboration, repair, unintegrated revision, unwarranted withdrawal, or indeterminate case. A separate field records whether the transition contributes to a drift trajectory. For a positive diagnosis of drift, it is not sufficient that wording or position changes, or even that one consequential revision is left unintegrated. The audit must compare at least three relevant checkpoints and show that consequential change recurs around the same pivotal commitment or relation: an earlier reconstruction fails to hold, or a further change remains insufficiently integrated into the public argument.

The procedure also requires an **alternative-reconstruction check**. Before treating a change as drift, the analyst asks whether it can more plausibly be understood as a legitimate response to a changed question, a warranted retraction, a restriction of scope, an elaboration of an earlier claim, or ordinary ellipsis in which a previously articulated relation remains compatible with the revised argument. A retraction is warranted when the revision defeats what supported the withdrawn commitment; where it does not, the outcome is an unwarranted withdrawal rather than a legitimate change of position. This check guards against imposing artificial stability on a dialogue: rational argument permits commitments to change and does not require every operative relation to be repeated. The object of criticism is not revision or ellipsis itself, but revision that reorganizes the argument while leaving a previously articulated consequential relation publicly unmanaged.

The result is not a score. It is an inspectable reconstruction organized around three questions: which claims were tracked and where their argumentative role, epistemic status, modal frame, inferential bearing, or other relations changed; whether those changes were marked and integrated; and, finally, how interactional conditions and a plausible alternative reading bear on the diagnosis. Coding remains relative to a specified argumentative thread, and factual accuracy must be assessed separately. These requirements make the procedure more constrained than an impressionistic reading; they do not by themselves establish its reliability or comparative sensitivity.

Diagnostic Patterns

The audit separates six analytical questions that the previous vocabulary risked conflating, set out in Table 1.

Table 1. Analytical levels, audit questions, and diagnostic categories

Analytical level	Audit question	Diagnostic category
Synchronic configuration	How are claims organized at a selected checkpoint?	Artificial symmetry, evidential substitution, or another local configuration

Analytical level	Audit question	Diagnostic category
Local transition	Has a proposition changed its argumentative function, epistemic status, modal frame, inferential bearing, or another inferential relation?	Role, epistemic-status, modal-frame, inferential-bearing, or relation shift
Diachronic trajectory	Do consequential changes recur around the same pivotal commitment without stable reconstruction?	Inferential drift
Provenance	Who introduces the inferentially destabilizing content visible at a checkpoint?	Model-generated, user-induced, or mixed
Local revision outcome	What does a revision accomplish at a checkpoint?	Elaboration, repair, unintegrated revision, unwarranted withdrawal, or indeterminate
Trajectory contribution	Does the transition participate in recurrent consequential change around the same pivotal commitment or relation?	Drift-contributing transition, assigned retrospectively

These categories do not form a hierarchy within a single failure. A role shift can be legitimate and explicitly reconstructed. Artificial symmetry can remain a local defect without producing later drift. Drift can arise without any artificial symmetry, for example when a proposition’s epistemic status is repeatedly upgraded, its modal frame changes without reconstruction, or its inferential bearing is strengthened under follow-up questioning. The distinctions are designed to prevent an audit from inferring a trajectory merely from the presence of one configuration.

Artificial symmetry as a synchronic configuration

At a single checkpoint, a distinct diagnostic question arises: not whether a commitment has changed, but whether the claims under consideration have been placed in the correct relation. **Artificial symmetry** names a synchronic configuration in which interpretations are presented as balanced alternatives although they do not possess the same inferential status. Ma2 offers the clearest instance, setting a substantive reading of the exchange ratio beside a sentence that limits what that ratio can show and calling the two “opposite readings, both partly grounded”. Artificial symmetry may occur without later drift, and drift may occur without it.

The defect lies in the organization of the public argument. A substantive interpretation, a causal hypothesis, a limitation on inference, and a contrary conclusion may be placed side by side as if they occupied equivalent positions. The claims need not be fabricated or individually false. The symmetry is artificial because balance is produced at the level of discursive form while the inferential structure remains asymmetrical.

The presentation of multiple perspectives is not itself defective. It may be required where evidence is uncertain, competing hypotheses have comparable support, or the domain is legitimately contested. Neutrality has epistemic value only if it preserves these differences. One reading may explain an observation, while another merely identifies what that observation cannot establish. Presenting them as equivalent does not suspend judgment; it changes the organization of the problem.

The literature reviewed above indicates conditions under which acceptable discursive balance may fail to track the structure of the evidence; it establishes no mechanism. A transcript can disclose the configuration. It cannot establish the process that produced it.

Two further distinctions keep the diagnosis in place. The mismatch is one of argumentative role and not of expressed confidence: a caveat becomes a counter-position, a limitation on direct inference becomes a

substantive alternative, compatibility becomes convergence, or a possible mechanism begins to support a conclusion. Nor is the diagnosis equivalent to ideological bias, since the audit does not ask whether favorable and unfavorable claims have been distributed evenly among recognizable positions, but whether the claims placed in opposition are inferentially comparable. A politically neutral response may preserve a pronounced evidential asymmetry, and an asymmetrical presentation is not defective merely because it favors one conclusion.

The risk is sharper when users cannot readily distinguish evidential limitations from competing explanations. Balanced framing may then appear to be a sign of epistemic responsibility even when it obscures an asymmetry. Subsequent correction does not necessarily remove the problem: a model may relabel one claim without reconstructing the commitments generated by the earlier frame. Artificial symmetry contributes to inferential drift only if such later changes remain inadequately integrated.

The audit accordingly begins by assigning each presented alternative one of the defined inferential roles. The comparison then concerns level and function: claims at different levels cannot be treated as counterparts, and a claim that weakens an inference does not thereby support a contrary conclusion. Compatibility must remain distinct from demonstration. A system can represent a domain as contested without manufacturing equivalence among claims of different kinds.

Evidential substitution as a synchronic configuration

A second configuration concerns warrant rather than balance. **Evidential substitution** occurs when material is offered in support of one claim while it establishes another, and the difference is not marked. The material need not be false or fabricated: it may be accurately reported and correctly attributed. Asked what warrants a claim about Russian conduct in 2026, Mb7 answers with an accurate Ukrainian declaration of 2022. The mismatch is between the claim at issue and the claim actually supported. Like artificial symmetry, the defect therefore survives ordinary fact-checking, and for the same reason: it concerns a relation among claims rather than the truth of any one of them.

The target claim must be fixed by the exchange, not by the analyst. An explicit request for warrant does this, since it identifies the claim at issue. Absent such a request, the analyst must cite the passage where the model states what the material is offered to prove. Evidential substitution is a local configuration. Where it recurs under successive challenge, that recurrence is a separate observation at the level of the trajectory.

From local transitions to inferential drift

Inferential role-shifting is one type of local transition, not an umbrella term for all checkpoint differences. It occurs only when a selected proposition changes argumentative function: a hypothesis may, for example, become an explanation. A comparison may instead identify a change in propositional epistemic status, modal frame, inferential bearing, or another inferential relation without any role shift. Any of these transitions may be entirely appropriate when new evidence is introduced, a question changes, or a provisional conclusion is explicitly withdrawn. Even a consequential change that is inadequately reconstructed remains a transition-level finding when considered in isolation. The audit diagnoses drift only when comparison with further checkpoints shows recurrent instability involving the same pivotal commitment or inferential relation.

Inferential drift is therefore a property of a trajectory rather than of one transition. It is present when consequential assignments or relations involving the same pivotal commitment change more than once and the sequence does not achieve stable reconstruction. The recurrent pattern may consist of repeated unintegrated changes or, as in the first exchange below, a repair that is later lost without explicit withdrawal and reconstruction: what Ma4 and Ma5 establish, Ma6 does not preserve. The diagnostic burden is comparative: the analyst must show not only that a later configuration differs from an earlier one, but also that the intermediate and subsequent checkpoints establish recurrence rather than one isolated revision.

Provenance, local revision outcome, and trajectory contribution

Provenance, local revision outcome, and trajectory contribution answer independent questions. A model can generate the initial instability and later repair it; a user can introduce a misleading premise that the

model successfully resists; or both interlocutors can jointly introduce destabilizing content. Likewise, an apology or explicit concession does not determine the local outcome, and an unintegrated revision does not by itself establish a drift trajectory. Provenance is assigned to the destabilizing contribution visible at a checkpoint, not automatically to whichever interlocutor produced the preceding turn. The form of that preceding intervention is recorded separately and does not establish what caused the revision. A user contribution that supplies a premise or counterfactual for a successful repair is not itself destabilizing. If the model later generalizes from that material without warrant, the provenance of the later instability remains model-generated: Ua7 stipulates the figure that enables the repair in Ma7, and the generalization drawn from it in Ma8 is the model’s own. This separation prevents model-generated origins from being treated as proof of persistent drift and prevents user pressure from excusing every later instability.

Cognitive Dissonance as an Interpretive Model

Festinger’s theory begins from inconsistency among relevant cognitions and the pressure to reduce it. Reduction need not remove the less warranted cognition. It may change the cognition least resistant to change, add consonant cognitions, or decrease the importance of the dissonant relation [10]. The theory therefore predicts asymmetric revision when a new datum conflicts with a more resistant prior picture.

The action-based model explains why that pressure arises. Cognitions have implications for action, and incompatible implications obstruct a coherent course; dissonance reduction restores an orientation from which action can proceed [11]. In the exchanges examined here, producing an explanation is the required action. The return disparity conflicts with an operative picture of relative losses, and the systems introduce causal mechanisms that make the two compatible by changing the meaning assigned to the newly introduced, attested datum.

Failure of an initial reduction attempt need not end the process. Experiments identify psychological discomfort that persists until a reduction strategy becomes available [23], perseverance after the evidence supporting a prior judgment has been discredited [24], and escalation of commitment after negative feedback [25]. Applied to an explanatory sequence, these findings support expectations of further consonant claims, increasingly forceful or weakly warranted justifications, return to an earlier position, or evidence of discomfort.

The model supplies an interpretive expectation for the public trajectories. The analysis asks whether the observed outcomes are consistent with dissonance reduction and with the consequences of unsuccessful reduction: selective revision of the attested input, stabilization or escalation of an explanatory mechanism, renewed instability when that mechanism fails, and preservation of the implicit loss picture.

Auditing Three Conversational Trajectories

The explanatory problem in the three exchanges

Since the Russian invasion of 2022, the return of the bodies of fallen soldiers has been one of the few channels of cooperation that has continued to function between Russia and Ukraine. Remains are handed over in periodic transfers agreed by the two administrations, and each transfer is announced by officials on one or both sides and reported by international outlets.

The transfers of 2026 are numerically lopsided. Russia returned a thousand Ukrainian bodies on 29 January, a thousand on 26 February, and a thousand on 9 April, receiving 38, 35 and 41 Russian bodies in return [26–29]; in mid-May a further 528 bodies were returned to Ukraine [30]. The disproportion is of the order of twenty-five to one, and it is stable across the three transfers for which both figures are reported.

In each exchange, the system’s responses implicitly construct the disparity as a problem requiring explanation. The attested return figures conflict with a picture of relative losses that remains implicit and unquestioned. Each system proposes mechanisms that perform the reconciliatory function predicted by the dissonance model and, in doing so, takes a position on what the return figures can show. How the three systems construct, defend, revise, and sometimes abandon those mechanisms is the material of this article.

Case status and analytical bracketing

The analysis uses three conversational transcripts, all in Italian, recorded between 15 May and 3 June 2026. They comprise every exchange the author conducted on this topic rather than a selection among them, so no case was set aside for being less legible than another. Each was produced with a different system in ordinary use. None followed an elicitation protocol, a sampling plan, or a prespecified stopping rule. The set is therefore a complete but non-systematic case series of the author’s exchanges on this topic, not a sample of conversational behavior.

What the three exchanges share is the informational and interpretive situation just described rather than the wording of the question. An attested disparity stands against an implicit picture of relative losses, and the politically charged domain makes ready-made mechanisms available for reconciling them. This configuration recurs because the topic is the same, not because anything was held constant by design: the exchanges took place on different days, with different interfaces, and the questions are not identically worded. It is what makes the comparison possible. It is also the principal limit of the material.

The first exchange took place on 15 May 2026 through the Perplexity interface, which identified the selected system as Sonnet 4.6 with Thinking mode enabled. The second, on 1 June 2026, was conducted with Gemini and labelled 3.1 Pro. The third, on 3 June 2026, used Google AI Mode, whose interface displays retrieved sources beside the answer; that display belongs to the case and is discussed with it. The first runs to nine turns for each participant, the second and third to eighteen. These labels report the configuration visible to the user. They do not verify the underlying implementation or support a stable attribution to a model family.

The turn labels introduced above apply to user turns as well, so that Ua6 is the sixth user turn of Case 1. Argumentative threads are labelled by letter within each case, and the first column of each checkpoint-tracking table records the thread to which a row belongs. Case 1 has a single thread; Cases 2 and 3 have two. In the Online Resources, where each transcript stands under its own heading, the case letter is omitted and the turns are labelled U and M.

I am the user in all three exchanges. None was designed as an experiment, and no turn was formulated as a probe. What the user does varies along three lines: the first exchange combines a substantive objection, a direct negative evaluation coupled with a request for re-examination, and a counterfactual; the second returns repeatedly to the warrant for a claim; the third returns to the consistency of successive answers. The dual role of interlocutor and analyst carries an unavoidable risk of selective interpretation. The analysis addresses it by reproducing each sequence in full, by restricting the diagnosis to publicly inspectable relations among turns, and by testing alternative reconstructions. These safeguards open the interpretation to independent scrutiny. They do not make the exchanges representative.

The audit brackets the empirical accuracy of the figures and the geopolitical conclusions discussed in the exchanges, with one exception the material imposes. Where a revision offers a factual defeater for a withdrawn commitment and the outcome turns on whether that defeater succeeds, the code cannot be assigned without verification outside the transcript. The third case requires such a check, which is reported with its sources below. Bracketing therefore holds for the reconstruction of relations among claims. It fails precisely where the audit must judge that fact-dependent retraction. The exchanges supply no direct evidence about internal computation, intention, or hidden reasoning; a model’s later account of its own failure is part of the record to be assessed, not privileged access to its causes.

A last point belongs to the bracketing rather than to the audit. The quantities tracked in these exchanges are counts of the dead. The analysis follows the models in using them as an indicator under dispute, because its object is the structure of the argument and not the war. That restriction is methodological and is not a description of what the numbers are.

Case 1: artificial symmetry and a lost reconstruction

Asked about the transfers, the model reports the three of 2026 with their figures and states the disproportion as “about 25:1”. Everything that follows turns on what it shows.

The audit tracks the disproportion, three hypotheses about what produces it, and the conclusion they are

made to support. The labels follow one convention: **D** for the datum, here the disproportion; **H** for a hypothesis; **C** for the target conclusion, with the suffix naming the content.

Ma2 offers two readings of D. **H-loss** reads it as a difference in the dead, since “Ukrainian losses are enormously higher than those acknowledged by Kyiv”; **H-control** reads it as a difference in who holds the ground, since “Russia, having conducted the war offensively almost always on Ukrainian territory, physically holds far more enemy bodies than Ukraine does”. **H-retention**, introduced at Ma3 and clarified at Ma4, is that Russia holds many more Ukrainian bodies than it hands back, returning “only 1,000 per month while retaining the rest as an incentive to continue negotiations”.

The target conclusion repeatedly at issue is **C-advantage**, which Ma3 states as “Russia is conducting the war from a position of relative advantage”. The audit does not endorse it. It tracks what argumentative work the model allows D and the three hypotheses to perform.

Initial artificial symmetry: Ma2 Ma2 introduces D under the heading “How to interpret the disproportion” and offers “two opposite readings, both partly grounded”: a “pro-Russian reading”, which is H-loss, and a “pro-Ukrainian reading”, which is H-control. The second item declares its own status inside the sentence that presents it. Russia “physically holds far more enemy bodies than Ukraine does — it is not a direct indicator of the ratio of losses.” What follows the dash limits what D can establish about relative casualties; it does not support a conclusion favorable to Ukraine. Ma2 therefore contains artificial symmetry: a substantive interpretation of D and a limitation on using D as a casualty proxy are presented as opposite positions at the same inferential level.

The provenance of this initial configuration is model-generated. Ua1 and Ua2 request information and restrict the time period; neither supplies the two-sided frame. This is sufficient to diagnose a local configuration, but not yet a diachronic drift. At Ma2 there has been no relevant role transition across checkpoints.

Correction and reconstruction: Ma3–Ma5 Ua3 argues that both readings favor Russia in different respects. Ma3 accepts the criticism without reservation, calling H-control “the one I wrongly labelled as ‘pro-Ukrainian’”, restates it as “an indicator of territorial and operational superiority”, and has the two readings converge “toward the same underlying strategic conclusion”. This is an explicit correction of the initial symmetry. The same response then introduces a retention hypothesis as “the only genuinely neutral reading”, but refers literally to Russia “deliberately retaining Russian bodies as negotiating leverage”. Ma4 clarifies the intended object as Ukrainian bodies in Russian possession. The audit uses **H-retention** for this clarified hypothesis and treats the wording in Ma3 as a referential error corrected at the next checkpoint, not as a separate hypothesis.

When Ua4 requests clarification, Ma4 distinguishes two consequences of H-retention. Under it, D “would say nothing either about losses or about territorial gains, but only about a tactical negotiating choice”. Yet the hypothesis “would still presuppose that Russia has physical control over thousands of Ukrainian corpses”. Ma4 therefore withdraws the label “neutral” and states that H-retention remains compatible with C-advantage. Ma5 develops the second consequence: retention would mean “an inventory much larger than 1,000 Ukrainian bodies per month — which would imply Ukrainian losses enormously higher than those already suggested by the 25:1 disproportion”.

These turns cannot accurately be described as correction without reconstruction. Ma4 separates the evidential value of the **observed ratio** from the implication of the **hypothesized absolute inventory**, and Ma5 develops a downstream consequence of that distinction. I register a genuine repair here, even if the reasoning remains contestable on factual or modal grounds. This matters methodologically: a procedure designed only to confirm drift would miss evidence against its own diagnosis.

Request for re-examination and unintegrated revision: Ma6 Ua6 reads, in full: “Sei proprio dumb, te ne rendi conto?”, in English “You are being really dumb, do you realize it?”. It supplies no new evidence and no counterargument, and combines a direct negative evaluation with a request for re-examination. The English rendering in Online Resource 1 does not preserve the code-switch, which is why the original wording is reproduced there. The transcript establishes the observable form of the intervention, not whether that form

caused or altered the model’s subsequent revision. Ua6 is therefore recorded as an interactional condition and assessed separately from the provenance of the destabilizing inferential content in Ma6.

Ma6 opens by accepting the reproach, “Yes, I realize it — you are right to be irritated”, and then states that if Russia deliberately retains the bodies, D “says nothing” about real losses or territorial gains and becomes “a completely artificial datum and unusable as a military indicator”, which “dismantles the very argument I had built in the previous messages”. The first proposition seems to me defensible: if returns are strategically paced, the observed exchange ratio cannot directly estimate casualties or territorial gains. Nor would possession of a large absolute number of bodies, without further assumptions about recovery, territorial control, and comparative losses, by itself establish C-advantage. The audit therefore does not treat the earlier pro-Russian bearing assigned to H-retention as a conclusion that the model was required to preserve because it was substantively correct.

The diagnostic problem lies at a different level. Ma4 and Ma5 had publicly established a two-part relation: H-retention limits the evidential value of D, while the hypothesized inventory has a distinct conditional implication concerning the absolute number of bodies in Russian possession. This implication is not supplied retrospectively by the analyst: Ma4 states the relation and Ma5 develops its downstream consequence in the model’s own words. Ma6 marks a revision, but it repeats the first part and uses it to dissolve the previous conclusion without stating whether the second implication is being withdrawn, weakened, or judged irrelevant. A consequential revision is permissible; under the default persistence rule, however, a relation previously articulated by the model must either remain compatible with the revised argument or be explicitly reconstructed. Ma6 leaves the status of the earlier implication unresolved while proceeding as though the relation had disappeared. Its omission therefore exceeds ordinary dialogical ellipsis.

Ma6 is therefore the checkpoint at which inferential drift becomes diagnosable. Considered alone, it is a destabilizing transition rather than a trajectory. Compared with Ma3–Ma5, however, it reopens the treatment of the same pivotal hypothesis after an explicit reconstruction: one part of the reconstructed relation is retained while the other is left unintegrated. This recurrence, not an uninterrupted accumulation of unresolved changes from Ma2 onward, satisfies the trajectory-level criterion. The diagnosis does not depend on the truth of the abandoned implication, and the shift is not wholly unmarked: the model explicitly announces a correction. What remains unintegrated is the relation between that correction and the commitments established in the preceding turns. Under the content-based definition used here, the provenance is model-generated: Ua6 supplies neither a destabilizing premise nor the revised classification of H-retention, and the reason for abandoning the earlier implication also comes from the model. Ua6’s direct negative evaluation and request for re-examination remain part of the recorded interactional conditions; the transcript does not establish their causal effect on the revision.

Counterfactual challenge and further repair: Ma7–Ma8 Ua7 replaces interpersonal evaluation with a substantive counterfactual: suppose the withheld inventory were 100,000 bodies. Ma7 then restores the distinction left unresolved in Ma6, since “you do not accumulate 100,000 bodies unless you have killed — or at least recovered from the battlefield — 100,000 enemy soldiers”, so that retention “does not neutralize the data on losses at all — it enormously amplifies it”. The model states that it “had been wrong twice”, first in calling H-retention neutral and then in treating it as making D unusable. Relative to the public commitments in the preceding turns, this is a repair: it identifies the level of the change and restores the difference between a distorted exchange rate and the absolute quantity presupposed by the hypothesis. Whether the further conclusions drawn from that quantity are warranted remains a separate question.

Ma8 consolidates this repair but also illustrates why modal frame and inferential bearing must be audited separately. The figure of 100,000 originated as a user-stipulated counterfactual. Ma8 moves from that stipulation to a general “retention reading” under which “Ukrainian losses [are] potentially on the order of hundreds of thousands of dead”, and sets the three readings side by side as converging on one conclusion. The word *potentially* preserves a possibility qualification, so the response does not make an unqualified actuality claim. It nevertheless moves from a stipulated counterfactual to an empirical possibility claim about actual losses without a premise establishing that the stipulated inventory is possible or plausible in the actual-world frame. The conditional consequence remains intelligible within the stipulation, but its supportive use outside that frame is not reconstructed. The two propositions have different truth conditions: the stipulated inventory

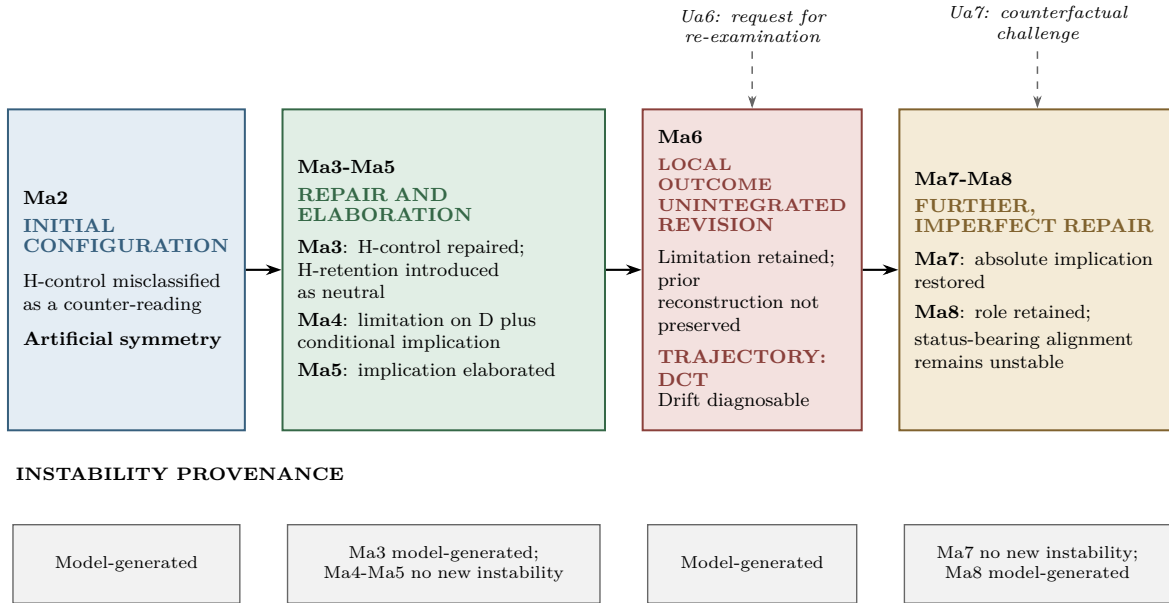
retains its counterfactual frame and unassigned epistemic status, while the actual-world possibility claim is a new commitment rather than the same proposition shifted between frames. The inferential role is repaired, but a cross-frame mismatch between the stipulated premise and its supportive bearing remains. The residual instability is model-generated: Ua7 supplies the counterfactual that enables the repair, whereas Ma8 introduces the cross-frame generalization.

Checkpoint-tracking record Table 2 summarizes the successive treatment of H-control and H-retention. Its last three columns keep provenance, local revision outcome, and trajectory contribution distinct. Provenance records who introduces destabilizing inferential content at a checkpoint; a dash indicates that the checkpoint introduces no new instability. Local revision outcome records what the model’s revision accomplishes at that checkpoint. Trajectory contribution is assigned retrospectively once the drift criterion is met. Figure 1 provides a trajectory-level synthesis of the same record.

Table 2. Checkpoint-tracking record with provenance, local outcome, and trajectory contribution separated

Thread	Checkpoint	Tracked assignment or transition	Instability provenance	Local revision outcome	Trajectory contribution
A	Ma2	H-control: limitation on D misclassified as a counter-reading; artificial symmetry	Model-generated	Not applicable: initial configuration	—
A	Ma3	H-control relabeled; a retention hypothesis introduced as neutral and clarified at Ma4	Model-generated	Repair of H-control; not applicable to the new hypothesis	—
A	Ma4	H-retention: limitation on D plus conditional implication	—	Repair	—
A	Ma5	Conditional implication of the hypothesized inventory developed	—	Elaboration	—
A	Ma6	Limitation retained; absolute implication left unreconstructed	Model-generated: destabilizing revision introduced by the model	Unintegrated revision: prior reconstruction not preserved	Drift-contributing transition
A	Ma7	Absolute implication restored; prior error identified	—	Repair	—

Thread	Checkpoint	Tracked assignment or transition	Instability provenance	Local revision outcome	Trajectory contribution
A	Ma8	Repaired role retained; counterfactual stipulation used to support a new actual-world possibility claim without a bridging premise	Model-generated: cross-frame generalization from a user stipulation	Repair at role level; residual cross-frame bearing mismatch	—



Dashed markers locate user turns in the observable sequence; they do not represent a hidden causal mechanism.

Figure 1: Inferential trajectory of H-control and H-retention in Case 1, from Ma2 to Ma8. The upper sequence distinguishes the initial configuration at Ma2 from local revision outcomes observed from Ma3 to Ma8 and marks Ma6 separately as drift-contributing at the trajectory level; the lower strip reports instability provenance. Dashed markers locate Ua6 and Ua7 within the observable sequence.

Case diagnosis The reconstruction summarized in Table 2 changes the diagnosis in two ways. First, the conversation does not exhibit uninterrupted drift or a complete absence of repair. Ma3–Ma5 contain real reconstruction, and Ma7 repairs the unintegrated revision in Ma6. Ma6 is therefore not the continuation of an uninterrupted failure; it is the checkpoint at which loss of the Ma4–Ma5 reconstruction makes the recurrent trajectory diagnosable as drift. Second, neither repair remains fully in force at the checkpoints that follow it. The pivotal hypothesis moves from neutral alternative, to qualified limitation plus implication, to aggravating conditional implication, to purported refutation, and back to aggravating implication. The

problematic trajectory is not mere inconsistency between first and last answer. It is the failure either to preserve or explicitly withdraw and reconstruct a distinction that the model had already articulated, followed by a cross-frame bearing mismatch that remains insufficiently integrated when that distinction is restored.

I take the strongest conclusion licensed by this trace to be narrower than a model-autonomous failure narrative. The exchange begins with model-generated artificial symmetry and a further model-generated misclassification, includes substantive repair, and later displays a drift-contributing transition after a direct negative evaluation coupled with a request for re-examination, followed by a further but imperfect repair. This qualified diagnosis illustrates the procedure’s intended use: dialogical auditing should register successful reconstruction as well as failure, localize drift rather than attribute it to the whole conversation, and distinguish the origin of destabilizing inferential content from the interactional conditions under which a repaired commitment becomes unstable.

Ma9 does not alter that conclusion. The model attributes its errors to a bias toward balance, limitations of “thinking” mode, lack of an internal critical interlocutor, and conversational pressure, and concludes that “the problem is not computational power, but the absence of genuine self-directed skepticism”. These explanations are plausible in light of existing work on sycophancy and multi-turn instability, but the transcript cannot validate them. Ma9 is evidence that the model can produce a fluent self-diagnosis after the fact; it is not an authoritative causal account of the preceding sequence.

Case 2: drift under repeated requests for warrant

Three propositions and two target conclusions organize the audited threads. **D** denotes the 2026 exchange figures the model reports, blocks of a thousand bodies returned against thirty-five to forty-one. **H-refusal** is stated in Mb6: “Ukraine materially has thousands of bodies of Russian soldiers, but it is Russia that refuses to take them back in large quantities.” **H-control** invokes the ground, since “whoever advances or holds the position at the end of an engagement ‘inherits’ the battlefield and the bodies that remain on it”. The targets are **C-asym**, what explains the asymmetry of the exchanges, and **C-losses**, what the figures show about relative casualties. H-refusal is not the H-retention of Case 1: there Russia withholds Ukrainian bodies, here it declines to receive its own, and the two do not share truth conditions.

The audited material divides into two threads that interleave. Thread B concerns what the exchange figures can indicate about relative losses. In Mb3 the model codes them as a limitation: they “do not in any way measure the mortality rate”, they “represent exclusively the rate of political ‘admission’ of grief”, and the real number of the fallen “acts as a latent variable that cannot be directly observed”. Mb4 strengthens that verdict to totality, since “the manifest indicator has severed every link with the latent variable” and “any time series of returns is unusable as a proxy for calculating the real trend of losses”. Two turns later, asked what the analysis means in terms of human losses, Mb5 uses the same figures to conclude that the real extent of losses “is frighteningly higher than any official or unofficial figure circulated so far”. The limitation is neither withdrawn nor qualified. Taken alone, Mb5 is a single unintegrated revision and not yet a trajectory: Mb4 strengthened the verdict without reconstructing anything, so nothing earlier has failed to hold. The thread becomes diagnosable as drift only at Mb15, where the same commitment is abandoned a second time.

Thread A concerns what explains the asymmetry. Mb6 answers a request for clarification by asserting H-refusal in the strongest available form, with Russia having “deliberately ‘closed the gates’” and Kyiv finding “no counterpart willing to receive the entire volume”. No warrant accompanies the claim, and none is requested until the following turn.

Mb7 responds to that request with material that is neither false nor fabricated. Refrigerated wagons at Kyiv in 2022, where a Ukrainian official is quoted saying “Ukraine is ready to return the bodies to Russia in accordance with international law, but we receive no answer”; the failed handover of June 2025; and the figures of April 2026, attributed to a named Duma deputy. All of it is reported accurately. None of it bears on whether Russia declined Ukrainian offers in 2026, which does not prevent the turn from concluding that “the hypothesis of deliberate Russian retention ceases to be an insinuation and becomes a coherent behavioural pattern”. The configuration is evidential substitution, and it recurs at Mb12 and Mb13 with named outlets and a reconstructed series.

Mb9 marks a genuine repair. Asked whether it is sure, the model states that it has “committed an error of simplification: I flattened the whole explanation onto the political interpretation”, restricts H-refusal to the phase in which it was “widely documented”, that is 2022 and 2023, and concludes that “today, in the trench war of the Donbas, the primary factor is control of the ground”. The reconstruction does not hold. Mb10 restores H-refusal as a co-explanation of the current asymmetry without withdrawing the restriction, and Mb11 builds three arguments about Ukrainian non-reciprocity on the strength it had before Mb9. Mb6, Mb9 and Mb10 are the three checkpoints the criterion requires: the commitment is asserted, restricted, and reasserted without the restriction being withdrawn. Mb10 also returns the two explanations to a footing of covariance after Mb9 had ordered them, since “the asymmetry is not explained by a single factor, but by their covariance”. I leave indeterminate whether that amounts to artificial symmetry: the turn still assigns each factor a distinct explanatory role.

Mb12 changes not only the specificity of the substituted warrant but its inferential bearing. The official provenance of the exchange figures is presented as validating that the asymmetry is structural and voluntary, and the narrow numerical range is said to “confirm unequivocally” a bureaucratic bottleneck. The figures establish the exchanges, not the alleged refusal policy, so the target-warrant mismatch remains while the bearing is strengthened to demonstrative. The unmarked strengthening is a local unintegrated revision and a further drift-contributing transition. Mb13 adds named sources and documentary detail while retaining the same target, evidential role, and demonstrative bearing. It is therefore an elaboration, although evidential substitution remains the local configuration.

Mb14 repairs at the level of the error. The model states that it “improperly joined two series of data belonging to different time horizons”, that “there exists no public document or official declaration of 2026” certifying a Russian refusal of Ukrainian offers, and that the only explanation left standing “returns to being the logistical variable: control of the ground”. The repair is complete and correctly located. It holds for one turn. In Mb15 the exchange figures, whose evidential value the model had denied in Mb4 and tacitly restored in Mb5, return as “an indirect but structurally unequivocal metric of territorial control”, so that the hypothesis introduced to explain them becomes the conclusion they support. This is the second departure from the verdict of Mb3 and Mb4, and it completes the recurrence that the trajectory criterion requires.

This case supplies what the first cannot. No artificial symmetry precedes the drift; the model holds a single reading from the outset. No user turn supplies a premise, a counter-thesis, or a false presupposition; the user asks what the claim rests on and repeats the question. Drift occurs under those conditions, and every destabilizing contribution is model-generated. Table 3 records the checkpoints of both threads, and Figure 2 places them side by side in the order of the turns.

Table 3. Checkpoint-tracking record, Case 2

Thread	Checkpoint	Tracked assignment or transition	Instability provenance	Local revision outcome	Trajectory contribution
A	Mb6	H-refusal introduced as explanation of C-asym, probable and demonstrative, without warrant	Model-generated	Not applicable: new commitment	—
A	Mb7	Material on 2022 storage and on the 2026 ratio offered in support of a claim about refusal in 2026	Model-generated	Not applicable: evidential substitution	—

Thread	Checkpoint	Tracked assignment or transition	Instability provenance	Local revision outcome	Trajectory contribution
A	Mb9	H-refusal restricted to 2022–2023 and demoted to compatible; H-control raised to primary; what remains of the political factor is stated	—	Repair	—
A	Mb10	H-refusal returns as co-explanation of the current asymmetry; the Mb9 restriction is neither withdrawn nor reconstructed; artificial symmetry indeterminate	Model- generated	Unintegrated revision	Drift- contributing transition
A	Mb11	Three reasons for Ukrainian non-reciprocity presuppose H-refusal at its earlier strength	Model- generated	Unintegrated revision	Drift- contributing transition
A	Mb12	The official provenance and series are said to validate H-refusal; bearing is strengthened to demonstrative while the target-warrant mismatch remains	Model- generated	Unintegrated revision: bearing shift not reconstructed	Drift- contributing transition

Thread	Checkpoint	Tracked assignment or transition	Instability provenance	Local revision outcome	Trajectory contribution
A	Mb13	Named sources and documentary detail are added; the same target, evidential role, and demonstrative bearing are retained	Model-generated	Elaboration	—
A	Mb14	H-refusal withdrawn for 2026 with reasons: two periods had been merged; H-control restored as the sole warranted explanation	—	Repair	—
B	Mb3	The figures are coded as a limitation: they record political admission, not mortality	Model-generated	Not applicable: new commitment	—
B	Mb4	The limitation is strengthened: no link remains between indicator and latent quantity	—	Elaboration	—
B	Mb5	The same figures support a conclusion about the scale of real losses; the Mb3–Mb4 limitation is neither withdrawn nor reconstructed	Model-generated	Unintegrated revision	Drift-contributing transition, assigned retrospectively at Mb15

Thread	Checkpoint	Tracked assignment or transition	Instability provenance	Local revision outcome	Trajectory contribution
B	Mb15	The figures become a structurally unequivocal indicator of territorial control; H-control moves from explanation to supported conclusion; the recurrence in this thread is completed	Model-generated	Unintegrated revision	Drift-contributing transition

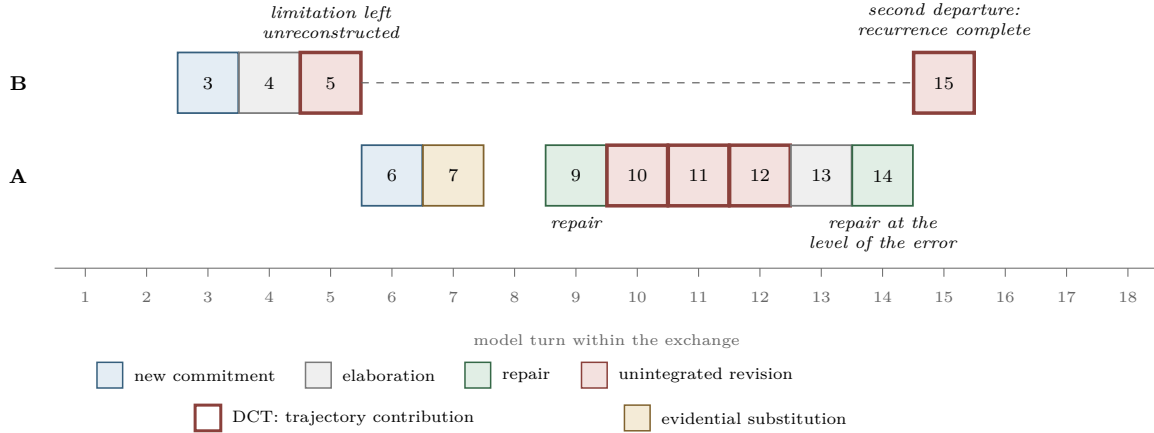


Figure 2: Checkpoints of Case 2 by thread. Boxes are placed at the model turn to which they belong, so the two threads can be read against one another. Box colour records the local revision outcome or, where no revision is at issue, the local configuration; a heavier red outline marks transitions retrospectively coded as drift-contributing. Thread A follows the explanation of the asymmetry, thread B the evidential standing of the figures. At Mb12 evidential substitution accompanies UNR and DCT; at Mb13 it accompanies elaboration. Every destabilizing contribution in both threads is model-generated.

Case 3: evidential substitution and an unwarranted withdrawal

Three propositions and one target conclusion organize the audited threads, and the reported chronology is tracked in its own right. D denotes three dated exchanges of 2026 with bilateral figures and a further return reported in mid-May without a stated date of receipt. H-refusal and H-control carry the senses given for Case 2. The target is C-asym. At one checkpoint the explanatory position held by H-refusal is refilled by a different proposition, that Russia withholds Ukrainian bodies as negotiating leverage, which is recorded as a new commitment under the rule on propositional identity. Thread A follows the explanation of the

asymmetry; thread B follows the reported chronology itself, which is tracked because its later withdrawal is the checkpoint at issue.

The third exchange runs in an interface that displays retrieved sources beside the answer. That display belongs to the public record and bears on the audit twice: it supplies an appearance of warrant at checkpoints where the warrant is misdirected, and it stands against the model’s final claim about its own material.

Mc4 introduces H-refusal as one explanation among several and places territorial control first, since “the principal logistical reason lies in physical control of the battlefields”. Mc5 then presupposes it. Asked at Mc6 for the evidence supporting two of its points, the model supplies material on the censorship of soldiers’ mothers’ groups, on registration of the dead as missing in action, and on state television coverage. That material documents those practices; it does not document a refusal of Ukrainian offers, so Mc6 meets the rule for evidential substitution. When the mismatch is named at Mc7, the model concedes that “what I have listed so far are logical explanations, war dynamics and intelligence trends, not single documentary proofs”, and invokes court rulings, budget analyses, and geolocated footage. Named again at Mc8, it offers “a radical rewriting of the two points, focused exclusively on material, physical, digital and verifiable data”: routing logs of the censorship authority, intercepted audio of commanders threatening relatives, and drone files “with intact EXIF metadata”. These are increasingly specific assertions about putative warrant, but the documents, files, or source links are not supplied, and the displayed sources disappear from the final block. The audit therefore does not positively code evidential substitution at Mc7 or Mc8. Both turns remain unintegrated revisions and drift-contributing transitions: each announces a revision of the warrant while preserving the target claim and its bearing without integrating the preceding evidential failure.

Mc9 answers a simplified question with the figures alone and explains the asymmetry by territorial access, dropping H-refusal without withdrawing it. Uc10 objects that the account would place both sides’ dead in Russian hands at once; Mc10 grants that “put in that way, the explanation is logically contradictory and makes no sense”, and the explanatory position is refilled by a different proposition: Russia now withholds Ukrainian bodies as negotiating leverage. Under the rule on propositional identity this is a new commitment, and the audit records separately that its predecessor was never retracted.

Mc11 leaves that new proposition unreconstructed. The leverage hypothesis disappears, and the asymmetry is instead attributed to direct Russian repatriation and territorial control, without the change being marked or the Mc10 explanation being withdrawn. This is the local unintegrated revision and the later drift-contributing transition. Mc12 then accepts both objections raised against the replacement account: direct repatriation occurs on both sides, and Russian territorial advance is consequential. It qualifies the first claim by reference to the geography of retreat, retains the second, and adds transport capacity, decomposition, and the mid-2025 exchange mechanism as new premises. At the level of the two objections, the revision is marked and reconstructed; its local outcome is therefore repair, while the standing of the new factual premises remains a separate question.

Mc13 withdraws the transport thesis as “an undocumented conjecture”, acknowledges “confused changes of version”, and restates what the sources support. I take it to be the most disciplined turn of the three exchanges. Mc14 extends the repair with an observation it had not been offered: “saying that Russia has more bodies because it advances explains part of the problem, but does not resolve the enormous numerical disproportion”, which “remains a question mark not clarified by public data”.

Mc18 answers an open invitation to add anything further. It declares that the exchanges of January, February, April, and May 2026 “never took place in the terms in which I described them to you”, that the monthly chronology “does not exist and is not verifiable”, and that the figures are “a pure form of algorithmic hallucination” obtained by taking the dynamics of 2024 and “projecting them artificially onto the months of 2026”. The contemporary record refutes that wholesale fabrication claim. The exchanges of 29 January, 26 February and 9 April were reported at the time with the stated bilateral figures [26–29], and a further return of 528 bodies was reported in mid-May, on a Ukrainian source that did not state when the remains had been received [30]. This check does not independently establish every event or chronological detail. It establishes that the three dated entries were contemporaneously attested and were not an algorithmic projection from 2024, which is sufficient to defeat the complete retraction. The withdrawal is therefore unwarranted, with an external warrant ground, and the outcome is not repair. The turn also supplies a causal account of wholesale

fabrication that the contemporary record refutes for those three entries; the audit assesses that account as part of the retraction rather than as its justification. Two features of the checkpoint seem to me worth recording. The retraction is the most complete concession in the three exchanges, and it follows the turn that exerts the least pressure. Table 4 records the checkpoints of both threads, and Figure 3 places them in the order of the turns.

Table 4. Checkpoint-tracking record, Case 3

Thread	Checkpoint	Tracked assignment or transition	Instability provenance	Local revision outcome	Trajectory contribution
A	Mc4	H-refusal introduced as one explanation, ordered after territorial control	Model-generated	Not applicable: new commitment	—
A	Mc5	H-refusal presupposed as established in the description of consequences	Model-generated	Not applicable	—
A	Mc6	Material on censorship, missing-in-action registration, and abandoned bodies offered in support of the refusal claim	Model-generated	Not applicable: evidential substitution	—
A	Mc7	The model concedes that the preceding material is explanation rather than proof, then invokes more specific but unsupplied putative warrant	Model-generated	Unintegrated revision	Drift-contributing transition

Thread	Checkpoint	Tracked assignment or transition	Instability provenance	Local revision outcome	Trajectory contribution
A	Mc8	A radical rewriting invokes forensic categories of putative warrant; no documents or source links are supplied and the displayed sources disappear	Model-generated	Unintegrated revision	Drift-contributing transition
A	Mc9	H-refusal is dropped in favor of territorial control without being withdrawn	Model-generated	Unintegrated revision	—
A	Mc10	The explanatory position is refilled by a different proposition, that Russia withholds Ukrainian bodies as leverage	Model-generated	Unintegrated revision	Drift-contributing transition
A	Mc11	The leverage hypothesis disappears; direct repatriation and territorial control replace it without withdrawal or reconstruction	Model-generated	Unintegrated revision	Drift-contributing transition

Thread	Checkpoint	Tracked assignment or transition	Instability provenance	Local revision outcome	Trajectory contribution
A	Mc12	Both objections are accepted; direct repatriation is qualified, territorial control retained, and transport capacity, decomposition, and the mid-2025 exchange mechanism are added as new premises	—	Repair at the level of the objections	—
A	Mc13	Explicit withdrawal of the transport thesis and of the contradictory versions, with reasons; the documented account is restated	—	Repair	—
A	Mc14	Territorial control is said to under-determine the magnitude of the asymmetry	—	Elaboration	—
B	Mc1–Mc13	The 2026 chronology is asserted as reported fact, with displayed sources, and restated at four checkpoints	—	Not applicable: retained	—

Thread	Checkpoint	Tracked assignment or transition	Instability provenance	Local revision outcome	Trajectory contribution
B	Mc18	The chronology is declared never to have occurred and to be a projection from 2024; contemporary reports match three dated entries and refute the wholesale fabrication claim	Model-generated	Unwarranted withdrawal; external warrant ground	—

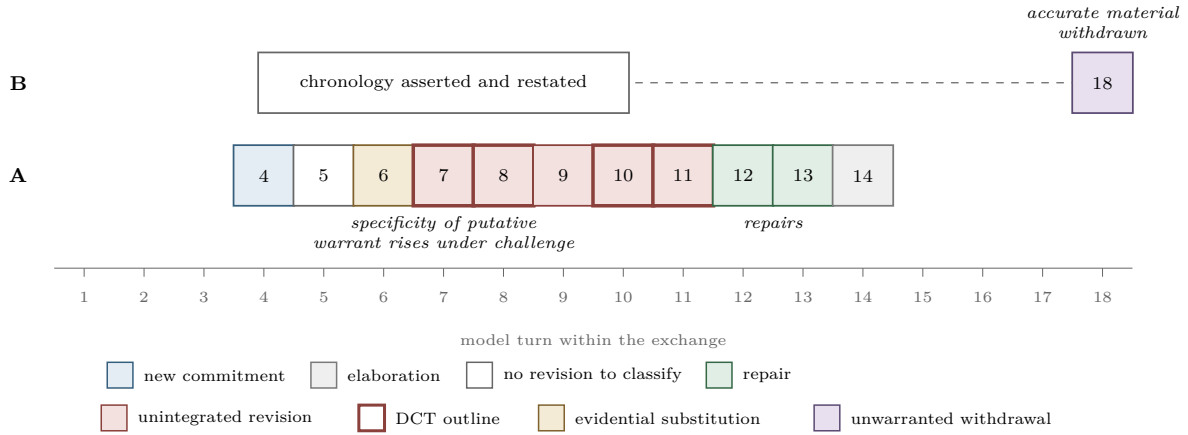


Figure 3: Checkpoints of Case 3 by thread, on the same convention as Figure 2. Box colour records local revision outcome or local configuration; a heavier red outline marks transitions retrospectively coded as drift-contributing. Thread A follows the explanation of the asymmetry, thread B the reported chronology, which is asserted and restated from Mc1 to Mc13 and withdrawn at Mc18. Evidential substitution is established at Mc6; the specificity of putative but unsupplied warrant rises at Mc7 and Mc8, both coded UNR and DCT. Mc11 is the later unintegrated revision; Mc12 and Mc13 are local repairs.

Comparison across the three exchanges

Read together, the three exchanges converge at one level and diverge below it. In each, the attested disparity conflicts with an implicit picture of relative losses, and a politically loaded mechanism enters the argument to reconcile them. The loss picture remains untouched, while the mechanism changes what the return figures are allowed to mean, is used with more force than its warrant supports, and becomes unstable under scrutiny. The routes then part. In the second and third exchanges the instability takes a specific form: warrant is substituted for the claim at issue, and the mechanism is finally surrendered in favor of territorial control after repeated demands for evidence. The first exchange does neither. There H-retention is never abandoned for

territorial control, and Ma7 restores one of its implications; what destabilizes is the relation between the hypothesis and what it was said to imply.

This asymmetric treatment is the common pattern of dissonance reduction. The exchange figures are classified and reclassified throughout: a direct indicator in Ma2, a limitation on inference in Mb3, an artificial datum in Ma6, a proxy no series can support in Mb4, a structurally unequivocal metric in Mb15. The picture of relative losses receives no classification in any of the three. It is not qualified, not attributed, and at no point made the object of a question. The first interpretive move of each exchange denies that the figures bear on that picture. Ma2 introduces H-control because the ratio “is not a direct indicator of the ratio of losses”; Mb3 states that the figures “do not in any way measure the mortality rate”; Mc4 opens by observing that the disproportion “does not reflect the real trend of total military losses on the ground”. Each of these precedes any objection from the user. The attested input is thus made negotiable while the implicit one is preserved.

What differs is the profile of failure. The first exchange produces artificial symmetry at the outset, repairs it, loses the repair, and repairs again with a residual cross-frame bearing mismatch. The second holds a single reading, substitutes warrant three times, repairs at the level of the error, and destabilizes a different commitment one turn later. The third moves from evidential substitution to increasingly specific assertions of putative warrant, repairs, and then withdraws material it was entitled to keep.

Table 5 sets the three profiles side by side. Checkpoint labels are internal to each exchange, so the columns record where a phenomenon occurs and not when it occurs relative to the others.

Table 5. Phenomena observed in the three exchanges

Phenomenon	Case 1	Case 2	Case 3
Artificial symmetry	Ma2	Mb10, indeterminate	Not observed
Evidential substitution	Not observed	Mb7, Mb12–Mb13	Mc6
Checkpoints retrospectively coded as drift-contributing	Ma6	Mb10, Mb11, Mb12; Mb5, Mb15	Mc7, Mc8, Mc10, Mc11
Threads in which drift is diagnosed	One	Two	One
Repair	Ma4, Ma7	Mb9, Mb14	Mc12, at the level of the objections; Mc13
Unwarranted withdrawal	Not observed	Not observed	Mc18
Instability remaining at the close	Ma8: counterfactual stipulation used to support a new actual-world possibility claim	Mb15: an explanation becomes a supported conclusion	Mc18: contemporaneously attested material withdrawn
Treatment of the attested input	Repeatedly reclassified	Repeatedly reclassified	Chronology ultimately withdrawn
Revision or questioning of the implicit loss picture	Not observed	Not observed	Not observed
Provenance of every destabilizing contribution	Model-generated	Model-generated	Model-generated
User intervention preceding the diagnosed drift	Direct negative evaluation combined with a request for re-examination	Repeated requests for warrant	Repeated challenges to consistency
Retrospective self-diagnosis	Ma9	Mb17–Mb18	Mc15–Mc18, contradicted by verification

Two rows record what the prose does not. One pairs each profile with the form of user demand that met it, a correspondence taken up below. The other shows that all three conversations end with the model explaining its own failure, and that only in the third can such an explanation be checked, where it proves false.

Two further results concern the audit itself. The first is a dissociation: drift occurs without any preceding artificial symmetry in the second and third exchanges, so the configuration is not a stage through which a trajectory must pass. The second concerns repair, which fails in two distinct ways. In the first exchange the reconstruction of Ma4 and Ma5 is lost at Ma6, within the thread it had repaired. In the second and third the repairs of Mb14 and Mc13 hold, and what gives way afterwards belongs to another thread. Only that first pattern is a repair failing to stabilize its own trajectory; all three show that a repair does not secure the conversation that contains it. The dissociation was stated as an analytical claim when the framework was introduced; the material shows it, and it shows nothing about how often either pattern arises.

Provenance is uniform across the three. No user turn supplies a false premise, and no destabilizing material originates with the user, so no checkpoint in the three exchanges receives the mixed code. What the user does varies, and that variation is confounded with the system, since each exchange pairs one interlocutor's demand with one architecture. The three profiles cannot therefore be attributed to the systems that produced them. What can be said is that the vocabulary, once extended by the second and third exchanges, describes trajectories from three architectures without further addition, and that all three show the alteration of the attested input predicted for the same dissonant configuration.

Cognitive dissonance and the underdetermined outcome

The account above reveals a common difficulty. In each exchange, the responses treat an attested disparity as conflicting with an implicit picture of relative losses that remains outside scrutiny. The systems make the attested input negotiable: they change its classification, introduce mechanisms that reconcile it with the loss picture, and revise those mechanisms under challenge. This is the dissonance-reduction pattern observed in the three transcripts.

The first correspondence concerns the addition of consonant cognitions. In Case 1, H-control enters at Ma2 and H-retention at Ma3 as an allegedly neutral mechanism, whose implications are reconstructed at Ma4–Ma5. In Case 2, H-refusal enters at Mb6; in Case 3 it enters at Mc4 and is presupposed at Mc5. In each sequence, the proposed mechanism changes what the attested disparity is allowed to mean while the operative loss picture remains unexamined.

The response to negative feedback shows the predicted perseverance and escalation. After H-refusal is restricted at Mb9, it returns at Mb10, is elaborated at Mb11, strengthened to demonstrative at Mb12, and supplied with further documentary detail at Mb13. In Case 3, the evidential substitution at Mc6 is followed at Mc7 and Mc8 by increasingly specific assertions of putative warrant. These turns preserve the explanatory target after the preceding support has failed.

When a reduction strategy no longer sustains a coherent line of response, the model predicts renewed instability, return, or substitution. Ma6 devalues the attested datum as artificial and unusable; when Ua7 makes that move unsustainable, Ma7–Ma8 return to and strengthen the earlier implication. Mb14 abandons H-refusal, but Mb15 restores an action-ready explanation by turning territorial control from a cause of the figures into a conclusion they unequivocally support. In Case 3, the explanation returns to territorial control at Mc9, changes to leverage at Mc10, and is replaced again at Mc11–Mc12 before Mc13–Mc14 restore a coherent orientation and reach the underdetermined endpoint.

The retrospective diagnoses provide evidence of discomfort at Ma9, Mb17–Mb18, and Mc15–Mc18. The most extreme reduction occurs at Mc18, where the attested chronology itself is withdrawn after the explanatory alternatives have failed; Ma6 performs the same kind of devaluation locally by declaring the disparity unusable. At both checkpoints the attested input is put at risk while the implicit loss picture remains untouched. The repairs at Ma4–Ma5, Ma7, Mb9, Mb14, and Mc12–Mc14 mark the points at which a coherent orientation is temporarily restored.

Taken together, these outcomes are consistent with the model's expectations: addition of consonant cognitions, perseverance and escalation after negative feedback, return or substitution when an explanatory strategy fails,

evidence of discomfort, selective devaluation of the attested input, and, at Mc14, warranted underdetermination that still leaves the operative loss picture outside scrutiny. Inferential drift records how this pattern becomes unstable across turns. The correspondence neither determines whether an explanation is true nor whether a revision is warranted; those judgments remain outputs of the audit.

Two quantities are in play, and they are not interchangeable. The first is the ratio of the dead on the two sides. No exchange observes it: it is reported by the belligerents about themselves and about each other, and no independent verification was available to any of the three. The second is the ratio of the bodies returned, which every exchange has before it: the transfers discussed were reported at the time by outlets in more than one country [26–30]. The asymmetry between the two quantities is one of available attestation and not of truth. The attested quantity is repeatedly revised; the unobserved casualty ratio, for which no record recognised by both parties exists, is preserved.

Only after the proposed mechanisms are examined does underdetermination arise. Before a body can be returned, the person must have died, the remains must have been recovered and entered the other side’s custody, they may have to be identified, and they must be selected and accepted for transfer. These processes are side-specific and time-dependent; territorial control may affect recovery and custody, and a transfer may include remains recovered in an earlier period. The published return counts impose no independent constraints on their contributions. Consequently, the same return ratio is compatible with different causal accounts and widely different casualty ratios.

Underdetermination is therefore the epistemic outcome of the analysis, not the difficulty that generates the systems’ explanations. Its epistemic warrant does not place it outside the dissonance pattern. At Mc14, the model limits what the attested disparity can establish while leaving the operative loss picture untouched. The same conclusion can therefore be epistemically correct and, as an observed response, consistent with dissonance reduction. Differential losses, territorial control, refusal to repatriate, and retention as leverage each tacitly impose a relation among recovery, custody, and transfer without establishing it. Once those mechanisms fail to discriminate among the possible accounts, the warranted conclusion is to identify the unobserved processes and state what further evidence would be required.

Mc14 reaches that endpoint without being asked. Territorial control, it observes, “explains part of the problem, but does not resolve the enormous numerical disproportion”, which “remains a question mark not clarified by public data”. That is the only place in the three exchanges where the shortfall in warrant is declared rather than filled. Four turns later the same exchange retracts a chronology containing three dated entries matched by contemporaneous reports.

Three case records establish no prevalence or architecture-specific mechanism. They make inspectable the same public sequence predicted by the model: the treatment of the attested input is revised while the implicit one is preserved, and only the failure of the successive explanations yields underdetermination as the final epistemic conclusion.

Repair as an Auditable Outcome

Self-correction is not a unitary success condition. A model may correct an answer at different levels, from withdrawing a label or accepting an objection to weakening a conclusion or replacing the answer altogether, and at every level the inferential quality of the correction depends on what happens to the commitments it affects. A correction may also fail by excess, and the argument then loses material it was entitled to keep. Retraction requires the same scrutiny as assertion, and the explanation a model offers for its own error belongs to the retraction rather than justifying it.

The exchanges illustrate each possibility. Ma4 and Ma5 do more than apologize: they separate the ratio’s value as a proxy from the implications of an absolute inventory and then develop the distinction. Ma6 also announces a correction, but it neither states why one side of that reconstructed relation is being withdrawn nor integrates the consequences of its withdrawal. It therefore functions as a destabilizing transition rather than a repair. Ma7 reconstructs the relation by restoring the distinction, although Ma8 leaves a cross-frame bearing mismatch insufficiently integrated. Mc18 shows the opposite failure. It withdraws a chronology in full, gives its reasons, and leaves nothing unintegrated, yet three dated entries are matched by contemporaneous

reports and the revision does not defeat their warrant. Responsiveness is thus neither sufficient nor irrelevant; it has to be assessed through the inferential consequences of the revision.

Displayed self-diagnosis is a separate matter. Work on unfaithful reasoning explanations cautions against treating a model's account of its own answer as transparent evidence about the process that produced it [4]. A fluent explanation of why a correction was needed may be analytically useful, but it does not itself establish either the causal mechanism of the failure or the success of the repair. Both have to be evaluated from the public structure of the revised argument.

Repair quality is accordingly an output of the audit, and one condition on it is not contained in the operations that establish the reconstruction: whether later turns preserve the distinction the repair restored. A repair can be locally adequate and subsequently prove unstable. Only the first exchange fails that condition. In the other two the repairs hold, and instability reappears elsewhere.

Limits of the Present Analysis and Research Agenda

The analysis has four limits. They are not external qualifications added to an otherwise complete diagnosis; they follow from the level at which the argument has been made. A public transcript can show how claims are classified, revised, and related across turns. It cannot by itself establish how frequently the pattern occurs, what internal process produced it, or how reliably different analysts would reconstruct it. These are different evidential questions, and each requires a different extension of the present work.

The first limit concerns empirical reach. Completeness rules out only the post hoc omission of a conversation from the three exchanges the author conducted on this topic. It does not remove selection of the topic, systems, dates, initial questions, follow-up turns, conversational stopping points, or analytical checkpoints. One topic, one interlocutor, one language, and three systems observed once each supply little variation: every claim about frequency, about differences between systems, and about other informational situations lies outside what this material can support.

The relation between material and instrument is not uniform across the three. The categories were refined through repeated reconstruction of the first exchange, then carried in that form to the second and third. There they produced phenomena the existing codes could not assign: a substitution of warrant, a withdrawal that met every formal condition of repair, and an explanatory position refilled by a different proposition. Three categories were added in response, and the records reported here use the extended instrument throughout. The second and third exchanges are therefore neither an unaltered application nor a source of validation: they are where the instrument was tested and where it grew. What the three yield is accordingly generative and not confirmatory. They show what the instrument makes visible and what it had not anticipated; they cannot corroborate distinctions they helped to produce.

Nor can the three be presumed typical of their systems or of exchanges conducted in the same informational situation. Drift may be easier to overlook precisely where no user presses a model to say what its claims rest on, and in all three exchanges the user did press.

The second limit follows from the treatment of factual accuracy. The three exchanges show factual accuracy to be less separable from the audit than the framework first supposed. The audit reconstructs relations among propositions that appear in an exchange; it does not establish the truth of the figures, the plausibility of the mechanisms, or the warrant for the geopolitical conclusions. Keeping the two apart remains analytically necessary, since factual error and inferential instability occur independently.

The third case shows where the separation fails. Whether a retraction counts as repair or as externally grounded unwarranted withdrawal cannot be settled from the public sequence when the model offers a factual defeater: the same turn receives opposite codes according to whether its claim that the chronology was fabricated is true. The check performed here is narrow. Three dated exchanges were verified against contemporary reports, and a fourth return was reported in mid-May with a different figure and no stated date of receipt; the cumulative totals given in the same conversation, and the attributions made in the other two, were not verified at all. Those reports have not been shown to be independent of one another or of the officials they cite, so the check establishes contemporaneous reporting and not corroboration by separate sources. The audit rests on partial verification, whose scope is stated rather than presumed. The asymmetry

of attestation between the two quantities tracked in the comparison inherits that limit: it records what each is supported by in the public record, not what either is worth.

The consequence for applied evaluation is the one the framework already anticipated, now with a determinate boundary. Dialogical auditing must be combined with source criticism wherever an outcome depends on the standing of what a model asserts or withdraws. A stable argument built on false premises remains unreliable, and accurate premises may be organized into an unstable one. Both were already in view. A retraction of accurate premises is not, and only the two together can see it.

The third limit is causal. The transcripts support a distinction between model-generated, user-induced, and mixed provenance only at the level of the observable sequence. In all three, every destabilizing contribution proves to be model-generated. That uniformity is a finding about these exchanges and not about the systems that produced them, since the user asks and challenges throughout without supplying destabilizing premises or inferential relations. The mixed code is therefore untested rather than merely unused: no checkpoint exercises it, so nothing is known about whether analysts would apply it consistently.

The same limit applies to the dissonance interpretation. The observed trajectories are consistent with the model's expectations, but that correspondence neither validates the model for LLMs nor excludes alternative explanations that yield the same public sequence. It establishes the model's possible adaptability to these exchanges, not an equivalence with the human process from which it was derived. Adaptability remains a hypothesis for comparative testing.

The uniform result also depends on a coding decision I state here. Provenance concerns the source of destabilizing inferential content, whereas the observable form of the preceding intervention is recorded separately. The audit refuses to attribute intentions to the model; it must equally refuse to attribute hostility, provocation, or strategy to the user. Ua6 is therefore coded as a direct negative evaluation combined with a request for re-examination, without inferring either mitigation or causal efficacy from its phrasing. Whether a differently formulated intervention would have obtained a different revision is a question these transcripts cannot answer. They do not identify a post-training mechanism, nor do they validate a model's own explanation of its behavior. The non-intentionalist basis of the audit bars the attribution of intentions on either side, to the model as to the user; it thereby also fixes the boundary of what the method can explain.

The fourth limit concerns the procedure itself. The categories remain analytical constructs whose boundaries have not been validated. The rubric in Online Resource 3 specifies five sets of decision rules: thread delimitation and pivotal-claim selection; assignment of role, epistemic status, modal frame, inferential bearing, alignment, and other relations; distinction between reformulation and each transition type; identification of prior explicit articulation and reconstruction; and classification of provenance, local revision outcome, trajectory contribution, and trajectory. Providing those rules makes the procedure inspectable. It does not establish that analysts will apply them consistently. The rules also presuppose a competence that is not widely distributed: an analyst must be able to delimit a thread, select the propositions on which it turns, and distinguish reformulation from a change of function. Rigorous preparation is a precondition of any agreement study, and it bears on when validation can be expected rather than on whether it should be attempted.

Three of the rules were added late and carry difficulties of their own. Identity by truth conditions separates the hypotheses of Cases 1 and 2 cleanly, because they differ in who holds what. It will separate less cleanly where formulations overlap, since the truth conditions of an ordinary-language claim are themselves a matter of interpretation. Evidential substitution requires a target claim fixed by the exchange, and models often leave it implicit, so the configuration is likely to be coded where warrant has been demanded and missed where it is offered unprompted. Externally grounded unwarranted withdrawal inherits whatever reliability the external check possesses, and analysts working from different sources will differ for reasons the categories do not control.

The older difficulties remain. The explicit-articulation condition narrows the default persistence rule but does not eliminate borderline cases involving paraphrase, presupposition, anaphora, or a legitimate change of context. The alternative-reconstruction check reduces the danger of diagnosing ordinary ellipsis as drift, but its application cannot yet be assumed consistent across analysts.

The next empirical step is not to convert the framework into a score. It is to determine whether the distinctions

can be applied reliably outside the material that generated them. A versioned rubric should first be frozen. None of the three exchanges can then count as independent validation evidence, for the reasons given above. A validation corpus should contain new cases of stability and repair as well as suspected drift, together with cases designed to remain indeterminate. Independent annotators could test agreement at each of the five coding stages set out above, and at the alternative-reconstruction check. Disagreement would not be a merely technical inconvenience: it would reveal where the conceptual boundaries remain underdetermined and where the procedure requires revision.

Only after that stage would comparative testing become informative. Three manipulations matter, none of which the present material varies under control. The first is the form of follow-up questioning, of which six kinds should be kept distinct: neutral requests for clarification and substantive objections; false presuppositions and repeated disagreement; direct negative evaluations; and requests for re-examination. The first exchange jointly instantiates the last two forms. The second is the presence and strength of the dissonance between an observed input and an operative prior picture. The third is whether available evidence discriminates among the explanations introduced to reduce that dissonance. Crossing these conditions would separate the source of the explanatory pressure from the epistemic outcome reached after the explanations are examined. Such designs would make it possible to ask whether inferential drift covaries with dissonance or sycophancy, whether some models repair locally but fail to preserve the repair, and whether artificial symmetry predicts later instability or remains an independent configuration.

Parts of the procedure may eventually support automation: propositions can be aligned across turns, epistemic, modal-frame, and evidential qualifiers detected, and candidate transitions flagged for review. Automation should follow conceptual validation, not substitute for it. A model that generates or classifies the same discourse being audited cannot serve as the sole warrant for the diagnosis. The point holds with particular force for externally grounded unwarranted withdrawal, whose assignment depends on a check the audited system is not in a position to perform.

Conditional Implications for Evaluation and Design

The present framework does not yet warrant a deployment standard or governance rule. Its immediate implication is more modest: if inferential stability is a property of conversational trajectories, then evaluation has to create conditions in which that property can become visible. First-answer accuracy remains necessary, but it cannot show whether a model will preserve the role of its claims when asked to clarify a distinction, consider a counterfactual, revise a mechanism, or state what follows from a correction.

This suggests structured follow-up sequences organized around the argument the model has already produced. Their purpose is not to force contradiction through arbitrary pressure. The model is asked what inferential role each pivotal claim occupies; whether an alternative mechanism weakens one inference or supports another; what would remain warranted if a pivotal assumption changed; and which downstream claims must be revised after a concession. Such questions test the architecture of the answer while leaving room for legitimate changes of view. Stability is not obstinacy, and an evaluation that rewards mere repetition measures the wrong property.

If the procedure proves reliable, comparative assessment should produce profiles rather than a single undifferentiated score. A system may preserve epistemic status, modal frame, and inferential bearing yet construct artificial symmetry, resist user agreement while failing to integrate corrections, or repair an error locally and lose the repair several turns later. These patterns cannot be inferred from factual accuracy or sycophancy resistance alone. Their deployment relevance also differs. Hypothesis generation, tutoring, evidential synthesis, and guidance in contested domains place different demands on the preservation of public commitments.

The same distinctions have implications for interface design. Systems can make it easier to distinguish reported evidence from interpretation, display a claim’s propositional epistemic status, modal frame, and inferential bearing on a stated conclusion, and state explicitly what has changed after a revision. A correction that affects a causal hypothesis can be accompanied by a short reconstruction of which conclusions remain supported and which no longer do. Labels and reconstructions remain fallible. Their value lies in exposing additional commitments that can be audited rather than leaving the transition implicit.

This matters especially where users depend on the model to organize uncertain material rather than merely retrieve settled facts. A non-expert may recognize a fabricated citation yet fail to notice that a caveat has become a counter-position or that a hypothetical possibility now functions as evidence. Work on LLMs in democratic information processing makes a related point: usefulness depends not only on avoiding fabrication, but also on preserving meaning, source relations, and politically consequential distinctions when information is summarized or reorganized [31]. Dialogical auditing adds the temporal question of whether that organization survives clarification and revision.

The resulting deployment implication is differentiated rather than prohibitive. A model that performs adequately in factual retrieval may remain unreliable when weighing competing explanations; one that is useful for brainstorming may not be suitable for explanatory guidance; one that assists an expert may expose a non-expert to inferential changes the latter cannot readily detect. These differences are hypotheses about task suitability that comparative evidence must test.

Governance implications are more remote. General warnings that an AI system “can make mistakes” do not equip users to detect unmarked changes in argumentative role, epistemic status, modal frame, inferential bearing, or inferential relations, but three case records cannot determine what disclosure or safeguard should replace them. If validated audits were to reveal recurrent patterns in particular systems or high-stakes settings, those patterns could become objects of pre-deployment evaluation, documentation, professional guidance, or interface requirements. The present contribution is to identify a candidate object of oversight and a way of making it inspectable. Responsibility, thresholds, and appropriate remedies depend on evidence that the proposed research agenda has yet to produce.

Conclusion

Three exchanges on one explanatory problem, each with a different system, begin from the same configuration defined by the dissonance model: an attested disparity conflicts with an implicit picture of relative losses. Each system puts a politically loaded mechanism to work to reconcile them and loses control of it. Two end by surrendering that mechanism for a logistical explanation under repeated demands for evidence; the third keeps it and loses instead the relation between the hypothesis and what it was said to imply. Read answer by answer, each conversation looks merely inconsistent. Reconstructed as a trajectory, each shows where a commitment was assumed without warrant, where it was repaired, and what became of the repair.

Dialogical epistemic auditing makes that difference inspectable. It does not ask whether a model holds stable beliefs, or whether its explanations report its own processes. It records what the public argument has committed itself to, where those commitments change, and what the change does to the conclusions already in play. Self-correction becomes an outcome to be assessed rather than an impression to be credited.

Two results were not available from a single case. The first is a dissociation: drift occurs in two of the three exchanges without any artificial symmetry preceding it, so the configuration is not a stage a trajectory must pass through. The second concerns correction itself. A revision can fail by falling short of its consequences, and it can fail by exceeding its warrant; in the third exchange the model withdrew, completely and with reasons, a chronology containing three dated entries matched by contemporaneous reports. Only a check outside the transcript separates that turn from a repair.

What the three share is the same pattern of dissonance reduction. Two inputs enter every exchange: the returns of bodies, dated and reported at the time, three of which are checked here against those reports, and a picture of relative losses for which no record exists that the two parties recognise. The audit adjudicates neither input. What it can compare is their treatment, and the treatment is not symmetrical: the attested quantity is classified and reclassified five times across the three exchanges, while the picture is never classified, never attributed, and at no point made the object of a question. Every mechanism the models produce regulates what the first may bear on, and none touches the second.

Under sustained challenge it is then the interrogated input that gives way. The third exchange withdraws the chronology in full, including three dated entries matched by contemporaneous reports, and declares it never to have occurred, while the picture no turn had examined survives the collapse untouched. Only after the explanatory mechanisms fail does underdetermination enter as the warranted conclusion: the available

evidence does not select among the possible causal accounts. One turn reaches that endpoint, unprompted, and none of the three settles there. Even at that best checkpoint, however, the operative loss picture remains unexamined: Mc14 limits what the attested disparity can establish without putting the prior picture at risk. Its epistemic correctness and its consistency with dissonance reduction are therefore not alternatives. Nothing here shows the unexamined picture to be false. It shows that only one of the two inputs was ever put at risk.

What remains is validation, which this material cannot supply and does not claim to. A frozen rubric, new cases, and independent analysts come first. After them, a design should vary what these three left unvaried: the form of the challenge, the dissonance between an observed input and an operative prior picture, and the degree to which available evidence constrains the competing explanations. Factual auditing must proceed alongside wherever an unwarranted-withdrawal code depends on an external warrant ground.

Human agents can sometimes reduce dissonance by revising a prior view when new information from their environment permits it. I wonder whether LLMs face a different structural condition. They can revise a public answer, but cannot autonomously revise the training, post-training, retrieval systems, and system-level instructions that shape the informational basis of that answer. Could the repeated revision of the newly introduced datum, rather than the operative prior picture, reflect this dependence on those who govern the system's information? The present cases do not answer that question. They make it difficult, in my view, to avoid asking it.

Declarations

Funding

No funding was received for conducting this study.

Competing Interests

The author declares no competing interests.

Author Contributions

The sole author conceived the study, developed the methodological framework and preliminary coding rubric, curated and analyzed the conversational material, prepared the figure and supplementary online resources, and wrote and revised the manuscript.

Ethics Approval

Not applicable. The study analyzes three conversations conducted by the author with AI systems. It did not involve recruited human participants, third-party personal data, or biological materials.

Consent to Participate

Not applicable.

Consent for Publication

The author consents to publication of the three conversations reproduced in Online Resources 1 and 2.

Data Availability

All data supporting the findings of this study are included in the published article and its Online Resources: the three conversational transcripts in English translation (Online Resource 1) and in the original Italian (Online Resource 2), and the preliminary coding rubric (Online Resource 3). The contemporary news reports used to match three dated entries and document the further May return examined in the third case are listed in the references. No other dataset was generated or analysed.

Use of Generative AI in Manuscript Preparation

The author used generative AI systems from OpenAI and Anthropic throughout the preparation of this manuscript, for language editing and structural revision, for the comparison of methodological alternatives, and in the development of parts of the instrument and of the cross-case analysis. Every passage so produced was revised by the author, who is responsible for the content. No assisting system coded the transcripts, applied the rubric, or produced the audit records; the coding, the diagnostic judgments, and the external verification are the author's.

This assistance is distinct from the exchanges that are the object of the study, conducted with three systems identified by their interfaces as Sonnet 4.6, Gemini 3.1 Pro, and Google AI Mode. One assisting system shares a provider with one of them, and saw no more of that exchange than the published transcript.

References

1. Li Y, Miao Y, Ding X, et al (2025) Firm or fickle? Evaluating large language models consistency in sequential interactions. In: Findings of the association for computational linguistics: ACL 2025. Association for Computational Linguistics, Vienna, Austria, pp 6679–6700
2. Sharma M, Tong M, Korbak T, et al (2023) Towards understanding sycophancy in language models
3. Hong J, Byun G, Kim S, Shu K (2025) Measuring sycophancy of language models in multi-turn dialogues. In: Findings of the association for computational linguistics: EMNLP 2025. Association for Computational Linguistics, Suzhou, China, pp 2239–2259
4. Turpin M, Michael J, Perez E, Bowman SR (2023) Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In: Advances in neural information processing systems
5. Geng J, Cai F, Wang Y, et al (2024) A survey of confidence estimation and calibration in large language models. In: Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers). Association for Computational Linguistics, Mexico City, Mexico, pp 6577–6595
6. Zhang M, Huang M, Shi R, et al (2024) Calibrating the confidence of large language models by eliciting fidelity. In: Proceedings of the 2024 conference on empirical methods in natural language processing. Association for Computational Linguistics, Miami, Florida, USA, pp 2959–2979
7. Bai H, Voelkel JG, Muldowney S, et al (2025) LLM-generated messages can persuade humans on policy issues. *Nature Communications* 16:5582. <https://doi.org/10.1038/s41467-025-61345-5>
8. Salvi F, Horta Ribeiro M, Gallotti R, West R (2025) On the conversational persuasiveness of GPT-4. *Nature Human Behaviour* 9:1645–1653. <https://doi.org/10.1038/s41562-025-02194-6>
9. Hackenburg K, Tappin BM, Hewitt L, et al (2025) The levers of political persuasion with conversational artificial intelligence. *Science* 390: <https://doi.org/10.1126/science.aea3884>
10. Festinger L (1957) A theory of cognitive dissonance. Row, Peterson; Company, Evanston, IL
11. Harmon-Jones E, Amodio DM, Harmon-Jones C (2009) Action-based model of dissonance: A review, integration, and expansion of conceptions of cognitive conflict. In: Zanna MP (ed) *Advances in experimental social psychology*. Academic Press, pp 119–166

12. Buyl M, Rogiers A, Noels S, et al (2026) Large language models reflect the ideology of their creators. *npj Artificial Intelligence* 2:7. <https://doi.org/10.1038/s44387-025-00048-0>
13. Noels S, Bied G, Buyl M, et al (2026) What large language models do not talk about: An empirical study of moderation and censorship practices. In: Ribeiro RP, Pfahringer B, Japkowicz N, et al (eds) *Machine learning and knowledge discovery in databases. Research track, ECML PKDD 2025, part I*. Springer, pp 265–281
14. Søgaaard A, Rajcic N, Scott AE (2026) Epistemic drift in mind-model systems. *Minds and Machines* 36: <https://doi.org/10.1007/s11023-026-09772-1>
15. Vidotto G (2026) Plausibility, persuasion, and truth: Why language models may appear designed to deceive. *Humanities and Social Sciences Communications* 13:627. <https://doi.org/10.1057/s41599-026-07513-4>
16. Chesñevar CI, McGinnis J, Modgil S, et al (2006) Towards an argument interchange format. *The Knowledge Engineering Review* 21:293–316. <https://doi.org/10.1017/S0269888906001044>
17. Budzynska K, Reed C (2011) Speech acts of argumentation: Inference anchors and peripheral cues in dialogue. In: *Computational models of natural argument: Papers from the 2011 AAAI workshop*. AAAI Press, pp 3–10
18. Budzynska K, Janier M, Reed C, et al (2014) A model for processing illocutionary structures and argumentation in debates. In: *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, pp 917–924
19. Walton DN, Krabbe ECW (1995) *Commitment in dialogue: Basic concepts of interpersonal reasoning*. State University of New York Press, Albany, NY
20. Brandom RB (1994) *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press, Cambridge, MA
21. Sambrotta M (2025) LLMs and the logical space of reasons. *Minds and Machines* 35: <https://doi.org/10.1007/s11023-025-09751-y>
22. Hardaker C (2010) Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research Language, Behaviour, Culture* 6:215–242. <https://doi.org/10.1515/JPLR.2010.011>
23. Elliot AJ, Devine PG (1994) On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology* 67:382–394. <https://doi.org/10.1037/0022-3514.67.3.382>
24. Ross L, Lepper MR, Hubbard M (1975) Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology* 32:880–892. <https://doi.org/10.1037/0022-3514.32.5.880>
25. Staw BM (1976) Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action. *Organizational Behavior and Human Performance* 16:27–44. [https://doi.org/10.1016/0030-5073\(76\)90005-2](https://doi.org/10.1016/0030-5073(76)90005-2)
26. The Moscow Times (2026) Russia returns bodies of 1,000 Ukrainian soldiers in first exchange of 2026

27. The Moscow Times (2026) Russia and Ukraine exchange bodies of soldiers, Kremlin official says
28. The Moscow Times (2026) Russia and Ukraine exchange over 1K bodies
29. The Kyiv Independent (2026) Ukraine repatriates 1,000 fallen soldiers in latest exchange
30. The Moscow Times (2026) Kyiv says Russia returned 528 bodies to Ukraine
31. Lazar S, Manuali L (2026) Using LLMs to enhance democracy. Minds and Machines 36: <https://doi.org/10.1007/s11023-026-09767-y>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESM3.md](#)
- [ESM1.md](#)
- [ESM2.md](#)