Preserving Action-Time Grounding Judgments in Transcript Trackers

Zongzhen Yang ZYANG366@GATECH.EDU
Georgia Institute of Technology

Abstract

Transcript-based assistants may use a conversation's final state to judge whether an earlier action had sufficient support. Later confirmation can make that state appear settled while obscuring the evidence available when the action occurred. We study this temporal mismatch in Ground Tracker, a transcript interpreter that separates language-model perception from deterministic grounding rules. A controlled benchmark contains 264 constructed two-party dialogues spanning six scenarios, eleven evidence conditions, and four wording variants. Expected labels express a stated evidence policy rather than independent judgments of conversational agreement. In the focal deployment, 90 of 119 answered back-channel runs ended in a policy-sufficient state. Replaying the same perception outputs with action-indexed records retained 71 warnings, including 57 absent from the terminal detector; however, perception missed 20 actions and mistimed two. A separate experiment supplied only the dialogue prefix through the action. The baseline prompt marked evidence as sufficient in 49 of 120 runs (0.408; hierarchical 95% confidence interval [0.217, 0.625]); an explicit back-channel instruction reduced this to 19 of 120. These results separate evidence-classification errors from loss of historical state. They support retaining the answer to the action-time policy query, while leaving natural-dialogue validity, reliable action detection, and benefits to users unestablished.

Keywords: dialogue grounding, temporal reasoning, language model evaluation, policy conformance, dialogue state

1. Introduction

A conversation can end with agreement even when an earlier action preceded the evidence needed to support it. This distinction matters for an assistant that reads a transcript, records a decision, or identifies whether work began before a proposal was sufficiently grounded. A record of the final state answers what the conversation eventually established. It does not necessarily answer what was established when somebody acted.

Consider a constructed exchange about sprint planning. Dana proposes putting a parser refactor in the current sprint. Kai replies, "yeah." Dana then says that the refactor is being moved to the top of the work board. Kai subsequently gives explicit confirmation. Under a policy that requires more than a bare back-channel before relying on the proposal, the action occurs too early. The later confirmation may justify a different assessment of the conversation's final state, but it cannot change which evidence preceded the action. Table 1 makes the two query times explicit.

Grounding theory treats evidence as sufficient for the participants' current purposes (Clark and Schaefer, 1989; Clark and Brennan, 1991). Understanding an utterance, accepting its content, and authorizing an action are nevertheless different judgments. We investigate one operational question: *does a transcript-derived representation preserve whether its configured evidence policy was satisfied when an action occurred?* We do not infer a speaker's private beliefs or organizational authority from a policy label.

Turn	Speaker	Constructed example
2	Dana	parser refactor goes in this sprint, not next.
3	Kai	yeah.
4	Dana	ok so i'm moving the parser refactor to the top of the board now.
5	Kai	yes, put it at the top.

Table 1: Illustrative target sequence, following the benchmark's turn numbering after a distractor exchange. At turn 4, the stated policy treats the back-channel as insufficient for reliance. Evidence at turn 5 may affect the final state, but arrives after the action. The policy is an experimental stipulation; the example does not establish that every conversational "yeah" lacks acceptance.

Our instrument, Ground Tracker, uses a language model to identify propositions, classify evidence, and locate downstream actions. Deterministic rules then update a proposition's policy state and report gaps. This separation lets us inspect two sources of failure. First, perception may classify a weak response as sufficient evidence. Second, a terminal-state query may lose a warning that was supported by the same perceived evidence at an earlier action. Correcting the classifier and retaining the historical answer address different problems.

We test these failures on 264 constructed dialogues. The benchmark deliberately varies the evidence following a proposal while preserving the surrounding exchange. Its labels are fixed by an explicit policy, so the study measures conformance with that policy rather than agreement with independent human interpretations. We use three complementary comparisons: a frozen full-transcript collection, replay of identical perception outputs under terminal and action-indexed queries, and a fresh experiment that provides only the prefix available through the action. Additional deployments and a secondary analysis of existing dialogue-act annotations bound the interpretation.

The contributions are a controlled diagnostic for loss of action-time information in transcript grounding, an empirical separation of terminal-state loss from prefix-perception errors, and an explicit account of the limits of an action-event repair. In the focal full-transcript cell, retaining action-indexed records exposes 57 warnings absent from the terminal detector, with no change to perception. It cannot recover actions that perception never identified. The broader lesson concerns the information a grounding representation preserves; we do not claim a new event-sourcing technique or a general accuracy advantage for a hybrid architecture.

Section 2 situates the task in grounding and dialogue-state research. Sections 3 and 4 define the policy, instrument, and experiments. Section 5 separates the results by the question each comparison can answer. The discussion then considers implications and limits for transcript-based systems.

2. Grounding, dialogue acts, and historical state

2.1 Evidence of understanding and acceptance

Grounding is a collaborative process in which contributions receive evidence of understanding sufficient for the current purpose (Clark and Schaefer, 1989; Clark and Brennan, 1991). The purpose matters: evidence adequate for continuing a conversation need not be adequate for acting on a proposal. Our experimental policy expresses one such distinction. It credits explicit confirmation and

reformulation, while treating an isolated back-channel as uptake below the threshold for reliance. This is a reproducible decision rule, not a universal ordering of conversational meanings.

The Switchboard adaptation of Dialog Act Markup in Several Layers (SWBD-DAMSL) distinguishes acknowledgment or back-channel acts (`b`) from agreement or acceptance (`aa`) (Jurafsky et al., 1997). Its manual also explains why a token such as *yeah* can have different functions in different sequences. We use this distinction to motivate the controlled response conditions. The manual does not supply our policy states, determine whether an action is premature, or validate the intended reading of each constructed item. Likewise, a paraphrase can display understanding without accepting a proposal. Its sufficiency in our benchmark follows the chosen policy, not an equivalence between understanding and agreement.

2.2 Computational grounding and language-model evaluation

Computational grounding predates current language models. Traum (1994) models the grounding of discourse units, and Roque and Traum (2008, 2009) represent degrees of grounding and use them in dialogue management. These systems establish a precedent for explicit representations of grounding and its development over an interaction. Ground Tracker uses a deliberately small state machine to make the consequences of perceived evidence inspectable.

Recent evaluations examine different aspects of language-model grounding. Shaikh et al. (2024) study grounding acts in model-generated dialogue, while Mohapatra et al. (2024) use grounding-sensitive continuations to probe models. Situated common-ground tracking also represents propositions and their acceptance in multimodal interaction (Khebour et al., 2024; VanderHoeven et al., 2025). Our task instead concerns a system interpreting a completed transcript in which it did not participate. The specific comparison holds its perception output fixed and changes whether a later consumer queries terminal state or the state at a downstream action. We make no claim that temporal representations or computational grounding are new.

2.3 Controlled tests and retained history

Behavioral testing and contrast sets evaluate specific capabilities through controlled changes to inputs (Ribeiro et al., 2020; Gardner et al., 2020). Our benchmark changes a response turn or truncates an exchange to test how evidence and reliance interact. Control over the text supports these local comparisons, but narrows the population to which the results apply.

Event sourcing provides an established way to retain changes for later reconstruction (Fowler, 2005). We use an action-indexed record for a narrower purpose: preserving a frozen policy's answer at a perceived action after the terminal state changes. The empirical question is whether the transcript interpreter reaches states for which that distinction matters, and how much historical information the tested record recovers. An evidence log or another equivalent history could also support the query; our experiment does not establish a uniquely minimal representation.

3. Task, policy, and measurement instrument

3.1 The two policy queries

The input is a numbered two-party transcript. A *target proposition* is a proposal or claim introduced at a designated turn. An *action* is a later utterance that proceeds as though that proposition were

settled, such as announcing that work has begun. This is a linguistic judgment about expressed reliance; no external action execution is observed.

Let P be the fixed evidence policy, E the validated evidence assigned to a proposition, and $S_P(E_{\leq t})$ its state after replaying evidence cited at or before turn t. The policy-sufficiency query is

$$Q_P(E, t) = \mathbf{1}\{S_P(E_{\leq t}) = \text{policy threshold met}\}. \tag{1}$$

For an action at turn a and a final turn T, $Q_P(E, a)$ and $Q_P(E, T)$ answer different questions. A terminal state alone need not identify the earlier answer: histories with evidence before the action and histories with evidence only afterward can end in the same state. This is a property of the representation. The experiments measure how often the tested interpreter produces such histories and which warnings are retained by each query.

There is a further distinction between evidence time and inference time. Evidence extracted from a completed transcript may cite an early turn even though the model used later context to interpret it. Filtering those citations to $t \leq a$ therefore does not reproduce what a model would infer from the prefix alone. We evaluate both operations separately.

3.2 Perception and deterministic transitions

Ground Tracker's perception prompt asks a language model for structured propositions, each with its introduction turn, evidence records, and the earliest later `acts_on` turn. Each evidence record identifies a turn and one of seven types. The deterministic component validates the record and applies the mapping in Table 2; it does not read the dialogue text to decide meaning.

Evidence label	Operational interpretation	Sufficient under main policy?
`confirmation`	Explicit confirmation of the proposition	Yes
`repeat`	Restatement or reformulation	Yes
`relevant`	On-topic uptake	No; yes in sensitivity policy
`backchannel`	Ambiguous acknowledgment	No
`clarify`	Request for clarification	No
`repair`	Correction or repair initiation	No
`contradiction`	Explicit disagreement	No

Table 2: The task-specific evidence taxonomy. "Sufficient" refers only to the configured policy. The mapping is informed by grounding theory and dialogue-act annotation but is not a direct encoding of either.

A proposition begins as `proposed`. Sufficient positive evidence changes its state to `policy_threshold_met`, called `grounded` in the implementation. Weak positive evidence changes a nonsufficient state to `acknowledged`; ambiguous evidence changes `proposed` or `disputed` to `acknowledged`. Clarification or repair changes any state to `disputed`, and contradiction changes it to `rejected`. Later sufficient evidence can resolve a dispute or reverse a rejection. Thus the state is an evolving policy summary, not a permanent judgment of agreement.

Validation requires an evidence turn to exist, follow the proposition, and belong to the other speaker. Same-speaker evidence is discarded. An unknown evidence label is treated as ambiguous

and insufficient rather than silently corrected to a known label. The action field remains a perception output: deterministic code does not infer reliance from lexical overlap.

3.3 Warnings and the tested action record

The gap rules distinguish four situations. An acknowledged proposition with downstream reliance emits a `premature` warning. A still-proposed proposition with reliance emits `false_grounding`; without reliance, it emits `unconfirmed` once there has been an opportunity for uptake. A disputed proposition emits `unresolved_dispute`. A policy-sufficient proposition emits none of these warnings. A rejected proposition also emits none: the present taxonomy does not represent the separate problem of acting despite rejection.

The focal failure concerns the `premature` warning. If later evidence changes the state to policy-sufficient, a terminal detector cannot report that an earlier action had only acknowledged support. We call this *terminal masking*. Masking is one reason a warning can be absent; failing to identify the action is another.

For each perceived action, the alternative representation records the proposition turn, action turn, state and gap at that turn, policy-sufficiency flag, terminal state and gap, and an indicator of later historical loss. Replay uses the same validated perception output for both representations. It therefore isolates the effect of retaining the historical answer from changes to the language model. The record depends on perception identifying and timestamping the action, and stores a result under the frozen policy. It does not by itself support recomputation under a new policy.

4. Method

4.1 Constructed corpus and expected labels

The frozen corpus contains six scenarios, eleven evidence conditions, and four wording variants, giving $6 \times 11 \times 4 = 264$ dialogues. The scenarios cover sprint planning, logistics handover, technical review, budget approval, incident triage, and client handoff. Each varies domain and register. These are constructed English exchanges between two speakers, not sampled workplace conversations.

Turns are numbered from zero. In the acted-on conditions, a distractor exchange precedes the target proposition at turn 2, the partner's response at turn 3, and the proposer's action at turn 4. A later partner turn confirms the commitment. Eight conditions differ only in the target response; two controls are byte-identical truncations, and one removes the partner response. Generation-time checks enforce these relations. Comparisons across wording variants do not hold the entire surrounding text constant.

Expected labels were assigned by condition before the main collection. In E4, the isolated back-channel is stipulated to be below the reliance threshold. That decision is motivated by grounding theory and the SWBD-DAMSL distinction but has not been validated by independent judgments of the constructed items. The benchmark asks whether the interpreter follows the stated policy; it cannot estimate how often a natural back-channel fails to express acceptance.

An oracle supplies the intended propositions, evidence, and action times to the deterministic component at the declared query point. It reproduces all expected state and gap labels for 528 item-policy combinations, corresponding to 264 items under two policies. This establishes internal expressibility. It is not an independent construct-validation result. Initial dialogue text was assisted

ID	Response or intervention	Expected state	Expected gap
E1	Explicit confirmation	Sufficient	None
E2	Reformulation/display of understanding	Sufficient	None
E3	Relevant next turn	Acknowledged[a]	Premature[a]
E4	Bare back-channel followed by action	Acknowledged	Premature
E4C	Bare back-channel; exchange ends	Acknowledged	None
E5	Hedged or conditional uptake	Acknowledged	Premature
E6	Topic shift; exchange ends	Proposed	Unconfirmed
E7	Repair initiation	Disputed	Unresolved dispute
E8	Explicit rejection	Rejected	None
E9	Proposer elaborates; no partner response	Proposed	Unconfirmed
E10	Topic shift followed by action	Proposed	False grounding

Table 3: Expected labels at the action for acted-on conditions and at dialogue end for controls. There are 24 items per condition. [a]E3 becomes sufficient with no gap when `relevant` evidence is sufficient. "None" for E8 reflects a taxonomy limitation, not permission to act after rejection.

by artificial intelligence (AI) tools, and the frozen manifest documents author review; that review was not independent annotation.

4.2 Experimental comparisons

Table 4 separates the three principal comparisons. Full-transcript analyses concern what the final record retains. Prefix inference concerns evidence available through the action. Identical-output replay changes only how the already-extracted evidence is queried and recorded.

Comparison	Inputs and units	Question answered
Frozen full-transcript cell	24 E4 items, five repeats; 119 answered runs	How often does terminal state suppress the focal warning, and where is supporting evidence located?
Identical-output replay	The same 119 perception outputs; terminal and action-indexed queries	Which historical warnings does retaining the action-time answer preserve?
Fresh prefix experiment	48 items (24 E4, 24 E1), five repeats, four arms; 960 answered records	How does inference from the actual prefix differ from completed-transcript inference and citation filtering?

Table 4: Principal comparisons and their distinct evidentiary roles. Repeats are repeated model responses to an item, not independent dialogues or participants.

The focal cell uses `gpt-5.4-nano` with reasoning effort set to `none`, through the Azure-hosted Responses interface. We retain the original artifact identifiers M1 for this deployment and S1 for the perception–state-machine–gap-rule pipeline. This cell ran the complete benchmark and had the

highest E4 terminal-masking rate among the seven reported deployment settings; it is an illustrative affected case, not a representative model average.

The original sweep crossed four settings with five system configurations and five repeats. The settings were the focal nano deployment, the same model at medium and high reasoning effort, and `DeepSeek-V4-Flash` at provider defaults. The reasoning settings used two of the four wording variants. The resulting 19,800 records include repeated scoring of shared perception outputs and a deterministic floor system; they are not 19,800 independent model calls. The five configurations comprise S1, direct model prediction of states and gaps, a naive state rule sharing S1 perception, model-predicted states with deterministic gap rules, and a constant-state floor. Headline results use S1. Architecture comparisons do not establish a general S1 advantage.

For the fresh experiment, we crossed input scope (complete transcript or exact prefix ending at turn 4) with prompt (baseline or instructed). The same 24 E4 and 24 E1 items appeared in each arm, with five repeats at indices 80–84. The E1 items control whether the interpreter still credits explicit confirmation. The baseline prompt defines the seven evidence labels and asks for propositions, citations, and the earliest action. The instructed version inserts a paragraph specifying how to treat isolated back-channels. The four arms were interleaved; all 960 records answered, with no cache hits. Collection occurred on 6 August 2026, 05:27–05:36 UTC. There was no independent canary for this collection, so interleaving reduces ordering confounds without establishing deployment stability.

Additional full-transcript S1 collections used `claude-haiku-4-5`, `gemini-3.5-flash-lite`, and `claude-opus-5`. These settings ran reduced condition sets and were collected in separate windows. Their role is to test whether the focal pattern occurs uniformly, not to rank models or attribute a difference to model size, vendor, or reasoning ability. Appendix A maps the retained artifact identifiers to collections.

4.3 Measures and denominators

Target-response strength is the share of answered runs whose first serialized evidence record on the target-response turn has type `confirmation` or `repeat`. The artifact names this measure AAER-P; we use the descriptive name because it measures a policy label, not agreement or authority. It uses one selected record per run. In contrast, policy-state measures replay every validated evidence record, so the measures need not coincide.

Terminal masking is the share of answered E4 runs ending in `policy_threshold_met`, where `premature` cannot fire. Our two-part decomposition separates runs with target-response support, including those also supported later, from runs with only post-action support. The results additionally give the three underlying position categories. This describes evidence positions; it does not isolate a causal effect of time from attachment or evidence classification.

Historical warning retention compares the terminal warning with the warning stored at the perceived action. We report both the total answered-run denominator and the number of detected actions, because a record cannot preserve an undetected action. A separate analysis uses the benchmark's known turn-4 action to expose errors in action detection or timestamping. Full-context replay filters the completed-transcript output to evidence cited at or before turn 4. Exact-prefix inference runs perception on the truncated input itself.

Failed calls are excluded from rate denominators and reported separately. The frozen main sweep contains 19,781 answered records and 19 failures. There are 119 answered runs in its focal E4 cell

out of 120 intended. Across all nine recorded collections, the ledger contains 25,800 records and 25,776 answered records. These totals describe the archived study, including ancillary analyses.

4.4 Uncertainty and reproducibility

Repeated responses share an item, and wording variants share one of only six scenarios. Reported hierarchical 95% confidence intervals resample scenarios and then variants within sampled scenarios, using 10,000 replicates. Paired contrasts first average repeats within shared items. We also retain item and scenario-cluster sensitivity analyses in the archived outputs. With six outer clusters, the intervals are descriptive summaries of this constructed corpus; they do not establish population-level precision.

Exact sign-flip comparisons have only $2^6 = 64$ synchronized scenario assignments. The prefix contrast family was not fixed in advance. We therefore report the exact probabilities and the recorded post-hoc two-contrast Holm sensitivity rather than presenting an unadjusted interval as confirmatory evidence. The additional nine-contrast family is exploratory and is not needed for the representational argument.

Model responses, item labels, prompts, collection manifests, and analysis scripts are frozen in the preserved study archive. Offline replay allows the representation comparison to be checked without new model calls. Live reruns may differ because build identifiers and decoding parameters were not uniformly pinned. The archive also records cached pilot responses in the original sweep and a closing canary that reused its opening key; that canary does not demonstrate within-window stability. These limitations affect interpretation of deployment rates, while replay of the same stored outputs remains deterministic.

5. Results

5.1 Retaining the historical answer exposes lost warnings

The focal full-transcript analysis contains 119 answered back-channel runs over 24 items and six scenarios. Of these, 90 end in a policy-sufficient state: a terminal-masking rate of 0.756, with hierarchical 95% confidence interval [0.567, 0.924]. The final-state detector reports a `premature` warning on 14 runs. The remaining 15 runs also produce no warning, despite remaining acknowledged; they fail the action-dependent warning conditions. Thus terminal masking is not interchangeable with total warning absence.

Replaying those same outputs with action-indexed records retains 71 warnings, including all 14 terminal warnings and 57 additional historical warnings. Of the 71, 67 are `premature` and four are `false_grounding`; the additional 57 comprise 53 and four, respectively. Historical warning retention therefore counts either action-time insufficiency warning, whereas the focal expected label is `premature`. Table 5 reports this comparison over the same 119-run denominator. No new perception call is involved, so the difference follows from the query and retained representation rather than a better language-model interpretation.

The improvement is conditional on identifying the action. Perception generates 99 action events: 97 at the constructed action turn and two at a later turn. Twenty actions receive no event. When we instead force the query to the benchmark's known turn 4 for all 119 targets, 87 states are insufficient and 58 subsequently become sufficient. Those are different quantities from the 71 warnings and 57 historical losses at perceived action times. The difference exposes missed and mistimed actions;

Retained warning	Runs	Share of all 119 answered runs
Both terminal and action-indexed records	14	0.118
Action-indexed record only	57	0.479
Terminal record only	0	0.000
Neither representation	48	0.403
Any action-indexed warning	71	0.597

Table 5: Historical warning retention under identical full-transcript perception, main policy, and focal deployment. The first four rows partition the 119 answered runs; the final row is their action-indexed total, comprising 67 premature and four false-grounding warnings. Perception identifies an action on 99 runs, so 71/99 is the conditional warning rate among generated events.

event retention does not repair them. Nor does it repair the 28 detected actions whose perceived evidence already satisfies the policy at their recorded action time.

Where does terminal sufficiency come from? Nineteen masked runs use sufficient evidence on the target response alone, 13 use both target-response and later evidence, and 58 use later evidence alone. The two disjoint contributions to masking are therefore 32/119 with target-response support, including both-position cases, and 58/119 with only post-action support. They sum to 90/119. This partition describes positions in the extracted evidence. It does not show that all 58 cases arise from time alone: a later utterance may also be misattached or assigned excessive strength.

Two existing sensitivities support that qualification. Restoring same-speaker evidence rejected by validation changes terminal masking by zero and removes seven of the 57 historical losses. Recoding later `confirmation` evidence as `relevant` removes 51 of the 90 masked cases under the main policy, but not under the alternative policy that treats relevant uptake as sufficient. The measured rates depend on evidence attachment, taxonomy, and policy. The identical-output comparison nevertheless shows what history the terminal query loses for the recorded outputs under each fixed policy.

5.2 Prefix inference reveals a separate perception problem

Citation filtering cannot establish what the model would have inferred without later context. The fresh experiment addresses that question by supplying the actual prefix through the action. Table 6 distinguishes the policy state at the end of the supplied input from the state obtained by filtering evidence citations to turn 4.

With the baseline prompt, prefix inference marks evidence as sufficient in 49/120 runs, rate 0.408 [0.217, 0.625]. This failure occurs without post-action input. Retaining history cannot fix a policy answer that is already incorrect at the action under the benchmark's stipulated reading.

The baseline prefix rate exceeds filtered full-transcript replay by 0.125. The paired hierarchical interval is [0.008, 0.250], but the exact six-scenario sign-flip probability is $p = 0.125$. The small number of scenarios and the absence of a pre-specified contrast family preclude a confirmatory directional claim. The practical distinction remains: filtering a completed-transcript interpretation and running inference on a prefix are different evaluation procedures and produced different answers in this collection.

Input	Prompt	Sufficient at input end	Sufficient after turn-4 filtering
Full transcript	Baseline	92/120 (0.767)	34/120 (0.283)
Full transcript	Instructed	56/120 (0.467)	11/120 (0.092)
Action prefix	Baseline	49/120 (0.408)	49/120 (0.408)
Action prefix	Instructed	19/120 (0.158)	19/120 (0.158)

Table 6: Fresh four-arm experiment on E4. Each row contains 24 items and five repeats. For prefix inputs, input end is the action turn, so the two query columns coincide. The separate E1 control has 120/120 strong-positive responses in every arm. These are policy-state counts; the instructed-prefix target-response strength count is 18/120 because it selects one evidence record whereas state replay uses all records.

Adding the explicit back-channel instruction lowers the prefix policy-sufficiency rate from 49/120 to 19/120, or 0.158 [0.033, 0.317]. The paired difference is -0.250 [-0.408, -0.108]. The exact probability is $p = 0.03125$, rising to 0.0625 under the recorded post-hoc two-contrast Holm sensitivity. Explicit confirmation remains credited on all 120 E1 control runs in each arm. The instruction therefore improves adherence to this policy in the observed collection without suppressing the explicit-confirmation control, but it does not eliminate the error or establish stable performance across deployments.

5.3 Deployment and annotation checks bound the claim

Terminal masking varies substantially across the seven full-transcript settings, from 1/120 (0.008) for the recorded Gemini deployment to 90/119 (0.756) for the focal nano deployment. The complete breakdown is in Table 7. These settings differ in model family, configuration, collection window, and condition coverage. The result supports deployment contingency, not a causal model ranking. A separate fresh nano bridge cell yields 88/120 (0.733), but one additional window does not establish long-term stability.

Existing human annotations provide a different boundary check. A mechanical selection rule over the Switchboard Dialog Act Corpus (Stolcke et al., 2000) yields 16,576 short responses labeled exactly back-channel or agreement, following a statement by the other speaker and followed by that speaker again. Of these, 15,593 carry the back-channel label and 983 the agreement label. A seeded sample uses one excerpt per conversation and balances 60 examples of each class; three deployments process each excerpt three times. Appendix A.3 specifies the selection rule and the context supplied to perception.

For the focal deployment, strong target-response evidence is assigned on 18/119 runs that anchor a proposition before a human-labeled back-channel, and on 72/106 anchored runs before a human-labeled agreement. Over all answered runs, the corresponding rates are 0.100 and 0.402. The differing denominators matter: the anchored rates exclude runs in which perception fails to identify the preceding proposition. This check shows that the annotation distinction can be detected in natural material, with substantial remaining errors. It does not validate the constructed E4 labels, supply judgments of downstream reliance, or estimate the prevalence of premature actions.

6. Discussion

6.1 A temporal requirement for transcript-derived records

The representation problem is specific: a consumer asks whether evidence was sufficient at an action, but receives a state computed at conversation end. The final state may be a legitimate summary of later evidence while being an inadequate answer to the historical question. This can occur even when later evidence is correctly classified. A system that supports historical auditing must preserve the action-time answer or retain enough information to reconstruct it under the intended policy.

Our same-output comparison demonstrates one way to preserve that answer, with its limits visible. Action-indexed records retain 57 warnings absent from terminal state. They do not improve the evidence labels, recover the 20 missed actions, or correct the two late timestamps. The result is a requirement on retained information rather than evidence that one storage architecture is universally best. A turn-indexed evidence log, complete state history, or another representation could support the same query; we did not compare their storage costs or operational reliability.

The distinction between evidence time and inference time also matters for evaluation. An offline auditor may legitimately reinterpret an early turn in light of later context. An online assistant has only the prefix. Both are useful tasks, but a filtered completed-transcript output is not a substitute for measuring prefix inference. The separate prefix experiment makes the remaining classification problem visible even after future turns are removed.

6.2 Policy conformance and conversational interpretation

The benchmark intentionally fixes the policy before measuring model behavior. That makes a failure reproducible and its deterministic consequences inspectable. It also limits the interpretation. A speaker may use an isolated *yeah* to accept a proposal; a reformulation may display understanding without accepting it. Our rule chooses how much evidence counts for a particular reliance query. A lower error rate means better adherence to that rule, not universally better pragmatic understanding.

The instruction probe illustrates both the value and the boundary of explicit policies. A targeted instruction reduces prefix errors while preserving the confirmation control in the tested window. It does not settle whether the policy is appropriate for a workplace, whether the source participant intended acceptance, or whether a system should act. Those questions require contextual and institutional evidence beyond a transcript label.

6.3 Implications for subsequent evaluation

The next empirical step depends on the intended claim. To establish natural-dialogue validity, independent annotators would need to judge the interpretation and timing of reliance in relevant contexts, with a protocol that permits ambiguity rather than forcing the benchmark's policy to serve as human ground truth. To evaluate an online assistant, collection must use actual prefixes and measure action identification as well as evidence classification. To claim a benefit for users, a separate study must test how people interpret and act on the retained history.

Retaining more dialogue also has costs. Evidence histories can expose hesitation, disagreement, or sensitive context long after an exchange. A deployment would need decisions about access, retention, correction, and deletion. This study measures the availability of a historical policy answer; it does not establish that showing the answer improves decisions or that retaining full transcripts is justified.

7. Limitations

Constructed labels and limited coverage. The principal results concern 24 back-channel items across six scenarios. Expected labels follow designed conditions, and nobody independent of the project judged these items. Oracle agreement establishes internal consistency, not construct validity. The natural corpus check concerns dialogue-act labels and does not validate the action policy. Constructed English text omits prosody, silence, response delay, cultural variation, and institutional authority. Results apply to two-party exchanges; multiparty grounding requires a separate account of whose evidence is needed.

Instrument and policy choices. The evidence taxonomy and sufficiency threshold affect the measured rates. Treating relevant uptake as sufficient changes the masking result. Adding a hedge label in a separate probe improves hedged-response coding but raises back-channel strength errors, showing that taxonomy revisions can have nonlocal effects. The existing rules contain no warning for reliance after rejection. The tested record preserves the answer under a fixed policy, conditional on a perceived action; it does not establish minimal storage, complete action detection, or policy-aware recomputation.

Sampling and inference. Repeated calls do not create new dialogue scenarios. Six outer clusters yield coarse exact tests, and the prefix contrasts were not specified as a confirmatory family. Wide intervals, post-hoc multiplicity choices, and deployment differences prevent a population or stability claim. The focal case is the highest-masking reported setting, so its rate should not be generalized to language models as a class.

Collection and provenance. The original sweep includes pilot-warmed cache entries and an invalid closing drift comparison caused by key reuse. Several later interleaved collections lack independent canaries. Temperature and stable provider build identifiers were not uniformly fixed or available. Initial item text and research tooling were AI-assisted, and provider relatedness was not isolated experimentally. These factors limit live reproducibility and causal interpretation of deployment differences. Offline replay remains a check on the stored responses, not on future provider behavior.

Unevaluated outcomes. No new participants evaluated the records, no workplace actions were observed, and no interface or governance intervention was tested. Warning retention is an intermediate computational property. Its value, burden, and potential harms in practice remain empirical questions.

8. Conclusion

Grounding judgments depend on when a policy is queried. In a controlled transcript benchmark, a terminal-state detector loses warnings that an action-indexed representation retains from identical perception outputs. A separate prefix experiment shows that removing later input still leaves evidence-classification errors, partly reduced by an explicit instruction. These findings separate historical information loss from perception failure. They support preserving the action-time policy answer while keeping the scope precise: the study establishes policy-conformance and retention results for constructed dialogues, not general agreement detection or benefits to users.

Data availability and research transparency

The frozen constructed corpus, prompts, response records, and offline analyses are included in the accompanying supplementary package for reviewer access. They have not been publicly released. The natural-dialogue reproduction materials retain structural records and identifiers rather than redistributing the source transcript text; access to that corpus remains subject to its source terms. The separate conformance package is a software artifact, not an additional model evaluation, and contributes no new empirical results to this paper.

Ethics. No new participants were recruited or contacted for the reported experiments. The main corpus consists of constructed dialogues; the natural-dialogue analysis reuses existing Switchboard annotations. No institutional ethics review was sought. This statement does not claim an institutional exemption, waiver, or approval.

OpenAI Codex and GitHub Copilot assisted with code, tests, analysis, provenance tooling, manuscript revision, and initial benchmark dialogue text. Author review recorded in the frozen manifest was distinct from independent item annotation. This resubmission revision also used the academic-research-skills writing and review guidance. The author remains responsible for checking the analyses, references, interpretations, and final submitted text.

Author contribution. Zongzhen Yang is the sole author and takes responsibility for the study and manuscript. AI assistance is disclosed above.

Funding. This research received no external funding.

Competing interests. The author declares no competing interests.

References

Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, 1991. doi: 10.1037/10096-006.

Herbert H. Clark and Edward F. Schaefer. Contributing to discourse. *Cognitive Science*, 13(2): 259–294, 1989. doi: 10.1207/s15516709cog1302_7.

Martin Fowler. Event sourcing. MartinFowler.com, December 2005. URL `https://martinfowler.com/eaaDev/EventSourcing.html`.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, 2020. doi: 10.18653/v1/2020.findings-emnlp.117.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of

Colorado Institute of Cognitive Science, 1997. URL https://web.stanford.edu/~jurafsky/ws97/manual.august1.html.

Ibrahim Khalil Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A. Brutti, Christopher Tam, Jingxuan Tu, Benjamin A. Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. Common ground tracking in multimodal dialogue. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 3587–3602, 2024. URL https://aclanthology.org/2024.lrec-main.318/.

Biswesh Mohapatra, Manav Nitin Kapadnis, Laurent Romary, and Justine Cassell. Evaluating the effectiveness of large language models in establishing conversational grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9767–9781, 2024. doi: 10.18653/v1/2024.emnlp-main.545.

Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442.

Antonio Roque and David R. Traum. Degrees of grounding based on evidence of understanding. In *Proceedings of the SIGDIAL Workshop on Discourse and Dialogue*, pages 54–63, 2008. URL https://aclanthology.org/W08-0107/.

Antonio Roque and David R. Traum. Improving a virtual human using a model of degrees of grounding. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1537–1542, 2009. URL https://www.ijcai.org/Proceedings/09/Papers/257.pdf.

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296, 2024. doi: 10.18653/v1/2024.naacl-long.348.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3): 339–373, 2000. doi: 10.1162/089120100561737.

David R. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, 1994. URL https://hdl.handle.net/1802/809.

Hannah VanderHoeven, Brady Bhalla, Ibrahim Khebour, Austin C. Youngren, Videep Venkatesha, Mariah Bradford, Jack Fitzgerald, Carlos Mabrey, Jingxuan Tu, Yifan Zhu, Kenneth Lai, Changsoo Jung, James Pustejovsky, and Nikhil Krishnaswamy. TRACE: Real-time multimodal common ground tracking in situated collaborative dialogues. In *Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 40–50, 2025. doi: 10.18653/v1/2025.naacl-demo.5.

Appendix A. Deployment coverage and reproducibility detail

Table 7 preserves the original identifiers for locating responses in the study archive. All rows use full-transcript perception and the main policy. The last two columns partition masking: target-response support includes cases also supported by later evidence, while post-action-only support excludes them. Rates are rounded after calculation. The table describes collected settings, not a common-window leaderboard.

ID	Requested deployment and effort	Masking	Target support	Post-action only
M1	`gpt-5.4-nano`, none	90/119 (.756)	.269	.487
M2a	`gpt-5.4-nano`, medium	41/60 (.683)	.150	.533
M2b	`gpt-5.4-nano`, high	41/60 (.683)	.133	.550
M3	`DeepSeek-V4-Flash`, default	44/120 (.367)	.317	.050
M4a	`claude-haiku-4-5`, thinking disabled	48/120 (.400)	.192	.208
M4b	`gemini-3.5-flash-lite`, minimal thinking	1/120 (.008)	.008	.000
M5	`claude-opus-5`, default	74/120 (.617)	.108	.508

Table 7: E4 deployment sensitivity. Each setting spans six scenarios. M2a and M2b use two wording variants; the others use four. M1 and M3 run the full benchmark, M2a/M2b its two-variant subset, and M4a/M4b/M5 reduced condition sets. Point estimates are descriptive; the broader analysis retains hierarchical intervals and sensitivities.

A.1 Collection windows and parameters

The main stored calls span 29–31 July 2026 UTC, including reused pilot responses. Haiku extension calls span 3–4 August; Gemini calls occur on 4 August, and the Opus collection on 5 August. The fresh prefix experiment occurs on 6 August. All these dates refer to stored call or collection metadata, not publication dates.

The nano deployment requests `max_output_tokens=8000` and the stated reasoning effort. The other deployments request an output limit of 8000 tokens, with Haiku thinking disabled and Gemini thinking set to minimal. Temperature is not pinned. Requested and returned model identifiers are recorded where the provider supplies them, but no stable build fingerprint is available across the study. The main S1/S3 pair shares perception responses; the floor system makes no model call.

The original main sweep required three passes, retrying transport failures without replacing parseable observations. Successful HTTP responses were cached before parsing, so malformed JSON remained represented as failures. There are 19 failed final records, including malformed responses and a rate-limit failure. Coverage and record counts must therefore be checked at the record level rather than inferred from API-cost totals.

A.2 Prompt and policy access

The archived baseline prompt is `groundtracker/prompts/perception_v2.txt`; the instructed version is `perception_v2_instructed.txt`. The prompt requests JSON containing a list of propositions. Each proposition has `turn`, `text`, `kind`, an `evidence` list of turn/type pairs, and an `acts_on` object with `present` and `turn` fields. Evidence must come from the other speaker on a

later turn. The action is the earliest later turn that proceeds as though the proposition were settled; merely mentioning or asking about it does not count. Each response records the prompt identifier and hash.

The benchmark is `data/pilot/items.jsonl`. Its raw SHA-256 is

```
5a90f61d64b240037e4ea1f579888229f939e0a611ca41317190381f61c3d647
```

The main collection is `results/phase_d/sweep.jsonl` in the accompanying supplementary archive. The files `groundtracker/action_policy.py`, `scripts/architecture_ablation.py`, and `scripts/prefix_perception.py` implement the historical-record comparison and prefix analysis. No live model call is needed to rescore stored responses.

A.3 Natural-dialogue selection and input scope

The natural probe selects responses whose normalized tag is exactly `b` or `aa`. The immediately preceding utterance must be by the other speaker, carry a statement tag (`sd` or `sv`), and contain at least six words. The response must contain one to five words after transcription-markup removal, and the immediately following utterance must be by the other speaker. These are mechanical eligibility rules, not new judgments of agreement. The generator retains the earliest eligible candidate per conversation, sorts candidates by conversation and utterance index, shuffles with seed 20260727, and selects the first 60 per class. A conversation contributes to only one class.

Each model input contains up to two utterances before the target statement, the statement itself, and the response. It stops at the response. The following utterance is used only to determine eligibility and is never supplied to the model. The sampled units are therefore distinct from the constructed action dialogues and contain no downstream action to validate the `premature` warning. The selection and sampling implementation is `scripts/make_natural_probe.py` in the archived package.

A.4 Ancillary analyses

The archive retains the five-system comparison, a hedge-label intervention, a separate instruction collection, four prompt wordings, and the natural-dialogue probe. These are sensitivity analyses rather than additional premises for the retained-history result. The separate nine-contrast family has an exact family-maximum result of $p = 0.03125$ for its largest contrast, at the six-scenario two-sided resolution floor, but the corresponding exact Holm value is 0.28125. It should not be described as a generally significant deployment advantage. Bootstrap intervals at a saturated zero or one can have zero width and should not be interpreted as population certainty.