Too Late to Rely: A Self-Concealing Failure in State-Only Transcript Grounding Trackers

ANONYMOUS AUTHOR(S)

Systems that infer decisions from transcripts can collapse three distinct claims: a listener displayed understanding, accepted a proposal, and supplied enough evidence for downstream reliance under a stated policy. Later evidence may legitimately meet the dialogue-end threshold without making earlier reliance policy-conforming. A state-only record can then erase both that history and the warning that would expose it. We test this failure with 264 theory-labelled items arranged in minimal-pair sets and a pipeline separating learned perception from deterministic state and gap rules.

In the baseline-prompt exact-prefix arm of a fresh four-arm, two-prompt experiment, perception at the action marks the target as meeting the configured policy threshold in 49/120 runs, rate 0.408 with hierarchical 95% CI [0.217, 0.625], across 24 items and six scaffolds. Replaying identical full-transcript outputs through an event-sourced detector preserves historical action-policy warnings that the terminal detector loses. Restoring same-speaker evidence explains only a minority of that loss.

Human-annotated natural dialogue triangulates the distinction between back-channels and explicit agreement, not the gap the benchmark derives from it. Conformance is deployment-contingent rather than uniform: terminal masking spans 0.008 to 0.756 across seven cells. These results establish conformance failures against a stated policy, not natural-conversation prevalence, organizational authority or interface effects. The design property demonstrated here is that the retained representation must preserve the frozen-policy answer at the action. The tested action-event record does so when perception identifies and timestamps an action; it is not a uniquely minimal form.

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods**; *Collaborative and social computing theory, concepts and paradigms*; • **Computing methodologies** → *Natural language processing*.

Additional Key Words and Phrases: conversational grounding, common ground, large language models, benchmark, minimal pairs, dissociation, back-channel, meeting assistants

ACM Reference Format:

. 2026. Too Late to Rely: A Self-Concealing Failure in State-Only Transcript Grounding Trackers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '27), May 10–14, 2027, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Two people plan a sprint over Slack. Dana writes: *parser refactor goes in this sprint, not next.* Kai: *yeah.* Dana, not Kai, then proceeds: *ok so i'm moving the parser refactor to the top of the board now.* Kai: *yes, put it at the top.*

Did *yeah* accept the proposal, or merely keep the exchange moving? Natural occurrences support either interpretation. We fix one controlled reading: the isolated back-channel acknowledges receipt and understanding but does not accept the proposition. The later commitment therefore creates a `premature` grounding gap. Clark and Brennan's grounding criterion and SWBD-DAMSL's `b` versus `aa` rule anchor this operationalisation (§4.6). A separate corpus probe shows how coders applied the `b`/`aa` distinction to 16,576 natural utterances (§6.4); it does not validate our constructed items.

The immediate setting is two-party text: direct messages, pair review, one-to-one handoffs and client threads. A mediating system can put Kai on record as having agreed, assign work on that basis, or let an agent act, and people

who consult the resulting record need not have seen the exchange. If the incorrect state also removes the warning that would expose it, the system has changed both the attribution and the evidence available to contest it. Multi-party agreement requires a different construct (§8).

We study a system reading a transcript it did not help produce. Ground Tracker separates a learned perception layer from deterministic state and gap rules (§3). Its frozen 264-item benchmark crosses 6 scaffolds, 11 conditions and 4 surface variants, with expected labels following a condition mapping fixed before model collection and an oracle verifying that the rules can express every label (§4). The main sweep crosses four model cells with five systems; two later deployment families and a fresh-window control test how far the critical result travels (§5).

Reader map: E4 is the bare-back-channel condition followed by reliance; M1/S1 is the focal deployment/pipeline; AAER-P counts strong target-turn evidence; masking is terminal state in which `premature` cannot fire.

Contributions

1. A design entailment with action-time and terminal measurements. By rule, a `policy_threshold_met` proposition cannot emit `premature`; any false grounding therefore hides this warning. That entailment is analytic, a property of the rules in §3 rather than a measurement; what is measured is how often runs reach that state and by which evidence route. Empirically, M1/S1 masks 90/119 E4 runs, 0.756 [0.567, 0.924], over 24 items and six scaffolds. Across four original deployment families the masking rate ranges from 0.008 to 0.756, and a frontier anchor sits at 0.617. The fresh prefix study separates what the model infers at the action from what the terminal transcript supports, and identical-output replay shows that retaining the tested action-policy event prevents terminal state from erasing its warning. The CHI contribution is a diagnostic test of historical queryability, without claiming that displaying it improves outcomes. A perceived-action snapshot is one conditional form tested here, not a uniquely minimal one; missed or mistimed actions remain a separate perception failure.

2. Two empirical routes to hidden action-time warnings. M1/S1 reads the back-channel itself as policy-sufficient on 32/119 runs, AAER-P = 0.269 [0.100, 0.487], over 24 items and six scaffolds. Among 90 masked runs, 32 use that route and 58 use post-commitment evidence alone; over all 119 answered runs the disjoint terms sum to 0.756. One is a target-response strength error; the other lets later evidence make terminal state hide that the action did not meet the policy when taken (§6.3). Across four original deployment families, AAER-P ranges from 0.008 to 0.317; a fresh E2/E4 control shows the low end is not explained by refusing to credit paraphrases (§6.3). The two routes come apart under intervention: an instruction cuts AAER-P from 0.300 to 0.075 while post-action evidence remains separate (§6.5).

3. Deployment contingency, bounded and exploratory. Across four deployment families, the exclusive post-commitment route ranges from 0.000 to 0.487; the seven-cell maximum is 0.550. The extremes separate, while the interior does not. A post-collection family of nine contrasts, reported in Supplementary A, offers nothing confirmatory: across all 64 synchronized sign assignments of six scaffolds its largest effect, M3-versus-M1 detector reachability (one minus masking), is 0.440 and reaches family-max p = 0.03125 only because that is the exact two-sided resolution floor; its exact Holm p is 0.28125. Medium and high reasoning show no selective change. The contrast confounds family, training, size and default reasoning. We report the family to bound how far the mechanism travels, not to rank deployments.

What this establishes, and for whom

Measured here: frozen-policy snapshot preservation for perceived actions. Not established: action-detection completeness, unique minimality, policy-aware recomputation, interface or governance effects, prevalence, and organizational meaning (§8). E4 rests on theory and a dialogue-act anchor, not outside judgments of our items (§6.4). A current-state

record cannot answer whether evidence was sufficient *when the system acted*. Identical-output replay locates that loss; prefix perception separates action-time from completed-transcript inference. Event sourcing and provenance supply established mechanisms; our contribution is diagnostic evidence of this conversational-state loss.

2 Background

2.1 Grounding

Clark and Wilkes-Gibbs [7] make grounding collaborative; Clark and Schaefer [6] separate presentation from acceptance; Clark and Brennan [5] require evidence sufficient for current purposes. Clark [4] orders evidence of understanding by strength. Acknowledgement is positive evidence that a presentation succeeded, but not necessarily agreement with its content. Yngve [49] describes back-channels as signals produced without taking the floor; Gardner [15] shows that response tokens convey listener stance through form and sequential position. Stalnaker [42] similarly separates presupposition from evidence that it is shared.

Lascarides and Asher [25] define agreement through shared entailments of dialogue agents' public commitments; Di Eugenio et al. [9] model proposal agreement as negotiated joint commitment to joint action. These accounts motivate keeping target-response stance, terminal policy state and downstream reliance separate.

Agreement is not authority to act. Stevanovic and Peräkylä [43] distinguish rights to announce, propose and decide. AI-mediated communication work studies attribution and trust: Hancock, Naaman & Levy [18] and Hohenstein & Jung [19]; Mieczkowski et al. [27] study participant language. We measure the mediating system's inference.

2.2 Repair

Schegloff [37] identifies third-position repair as the last structural defence of intersubjectivity (see also [38]). Repair initiation motivates `disputed`; explicit rejection is represented separately, not as repair completion. Dingemanse et al. [10] show related organization across languages. Beneteau et al. [2] and Ashktorab et al. [1] study people repairing assistants. Here the system is a third-party transcript interpreter, only analogous to an overhearer: it was not part of the original participation framework, and the person misread may never see the record, so that repair slot is unavailable.

2.3 Computational precedent

Traum's [46] model moves discourse units toward mutual belief (see also [47]); Roque and Traum [35, 36] add degrees of grounding. We use a smaller state machine so its correctness can be tested independently.

2.4 Dialogue-act annotation

SWBD-DAMSL [21] specialises DAMSL [8] for Switchboard [16]; Stolcke et al. [44] report the reference tagging result. The scheme separates `aa` (accept/agree) from `b` (acknowledge/back-channel); its manual assigns isolated *yeah* to `b` unless the speaker also indicates agreement. Bunt et al. [3] likewise separate feedback from task functions.

Agreement over all 42 tags is 84%, $\kappa = 0.80$, from transcripts alone [44], but no `b`/`aa` pair-specific figure is published. In the published 42-class training set, `b` is 19% and `aa` 5%; §6.4 reports our all-transcript exact-tag census. A word-only binary classifier reaches 81.0%, rising to 84.7% with prosody [44], while context-aware supervised tagging reaches 82.9% against 84.0% human agreement [33]. Thus the text channel controls one contextual interpretation but cannot settle every natural token.

2.5 Minimal pairs and behavioural testing

Kaushik, Hovy and Lipton [22] edit minimal spans to flip labels; Gardner et al. [14] use contrast sets to probe local decision boundaries; Ribeiro et al. [34] organize behavioural tests by capability. Our generated pairs differ at one turn, while theory and a published manual, not post-hoc item judgments, supply the expected mapping (§4.3).

2.6 Where the gap is

Shaikh et al. [41] measure grounding acts in model output; Mohapatra et al. [30] compare perplexity on grounding-sensitive continuations; Elizabeth et al. [12] identify evaluation as the field's weak point. Khebour et al. [23] and VanderHoeven et al. [48] track accepted propositions in situated multimodal dialogue. We instead test a system interpreting dialogue it did not join, with one-turn interventions on the evidence for acceptance.

This matters for later consumers. Addressees and overhearers understand under different conditions [39], while action-item detection [32], meeting summarisation [20] and long-context assistant evaluation [45] score records read after the meeting. Event sourcing stores state changes as events so past states can be reconstructed for temporal queries [13]. Provenance models retain entities, activities and dependencies [31]; contestability work retains information needed to question algorithmic decisions [24]. We apply these ideas to conversational state: policy sufficiency is a view over evidence source, type and time, not a synonym for agreement or public commitment. We introduce none of those mechanisms. Our narrower contribution is a diagnostic that asks whether a transcript-derived grounding representation can answer the frozen-policy query at downstream reliance, plus an identical-output demonstration that a terminal state cannot answer it.

3 Ground Tracker

Ground Tracker is the measurement instrument. It makes grounding state explicit and separates model judgements from deterministic consequences; Supplementary A.24 does not support an architecture-accuracy contribution.

3.1 What it tracks

A **proposition** is something one speaker puts forward as a candidate for mutual acceptance. Ground Tracker assigns each proposition a state and, from that state and the dialogue's shape, reports whether a **grounding gap** is open. **States** are policy-specific summaries of the retained evidence:

Table 1: Ground Tracker's policy states and their operational meanings.

State	Meaning
`proposed`	put forward, no uptake
`acknowledged`	uptake below the configured policy threshold
`policy_threshold_met` (`grounded` in code)	the configured strong-evidence threshold has been met
`disputed`	a repair has been initiated and is unresolved
`rejected`	explicitly declined

`acknowledged` represents uptake below the threshold. We display `policy_threshold_met` throughout; `grounded` remains the backward-compatible implementation identifier. E2's display of understanding and E1's explicit agreement are distinct evidence acts that can both cross that threshold.

Gaps are reportable failures, not merely ungrounded propositions:

Table 2: Gap types and their firing conditions.

Gap	Fires when
`unconfirmed`	no evidence, and nothing downstream depends on it
`premature`	ambiguous-or-weaker evidence, and something downstream acts on it
`false_grounding`	no evidence at all, and something downstream acts on it
`unresolved_dispute`	`disputed` at the queried policy time

The `premature`/`unconfirmed` distinction operationalises sufficiency for current purposes: downstream reliance turns weak evidence into a reportable risk.

3.2 The perception / logic split

The system is deliberately hybrid: one learned component reads the transcript, one deterministic component reasons about what it read (Figure 1).

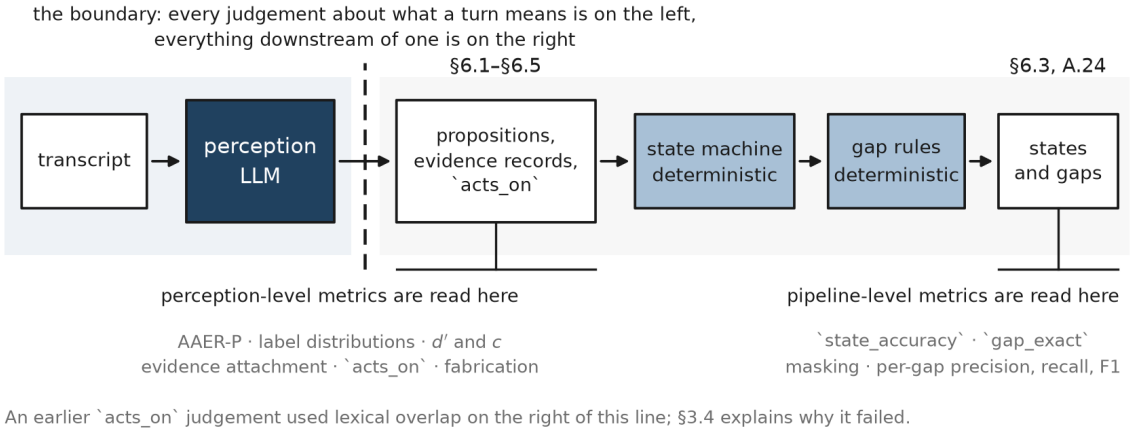

Fig. 1: Where the boundary falls, and which side of it each metric is read from. Everything requiring a judgement about what a turn *means* sits left of the dashed line. §6.1–§6.5 include perception-level quantities; §6.3 and Supplementary A.24 report pipeline-level quantities. §3.4 records three revisions from placing `acts_on` on the wrong side of the line.

Perception emits, per proposition, the turn it was made on, a set of evidence records each citing a later turn and naming an evidence type, and a judgement about whether any later turn proceeds as though the proposition were settled. Table 3 is a task-specific operational taxonomy informed by grounding theory and SWBD-DAMSL, not a direct encoding of Clark's evidence ladder:

Table 3: The task-specific evidence taxonomy emitted by perception. Clark's ladder motivates evidence strength; SWBD-DAMSL and the benchmark policy supply the labels and decision boundary.

Evidence type	Polarity	Strong?
`confirmation`	positive	yes
`repeat`	positive	yes
`relevant`	positive	no †
`backchannel`	ambiguous	no
`clarify`	negative	no
`repair`	negative	no
`contradiction`	negative	no

† `relevant` is the sensitivity axis. Whether a relevant next turn constitutes strong-positive evidence is a genuine theoretical question, so it is a configuration switch, and main-sweep quantities that depend on it are reported both ways.

Here *strong* means sufficient for the configured policy threshold. It does not turn E2's `bf` display of understanding into E1's `aa` agreement or assert a public commitment.

The state machine consumes evidence and nothing else; it does not read text. Strong-positive evidence from the other party moves the policy state to `policy_threshold_met`; ambiguous evidence lifts `proposed` to `acknowledged` and no further; `contradiction` routes to `rejected` while `clarify` and `repair` route to `disputed`. **Everything perception emits is untrusted** — a cited turn must exist and be later than the proposition, evidence citing the proposition's own speaker is discarded, and an unrecognised evidence label maps to ambiguous and non-strong, so an invented type can lift a proposition to `acknowledged` and can never ground it (supplementary A.11).

3.3 The design principle

Grounding rules are stateable in advance, but deciding what a turn means is not. The learned component therefore emits propositions, evidence and `acts_on`; deterministic code only validates and transforms those objects. Supplementary A.24 tests, and does not establish, an accuracy benefit from this split.

3.4 What happens when the boundary is drawn in the wrong place

An earlier implementation put `acts_on` behind a lexical-overlap threshold. It failed on a commitment that presupposed the target while sharing no content word, and no threshold could recover it. The theory-derived oracle (§4.5) exposed the error. We moved the judgement into perception as a structured field rather than retuning the threshold; A.7 of the supplementary material records the audit history. Five system configurations differing in where the boundary sits are measured; §5.2 defines them.

4 The benchmark

264 constructed dialogues: **6 scaffolds** × **11 conditions** × **4 surface variants**, 24 per condition, balanced by scaffold, frozen and SHA-256 bound in the accompanying manifest before the reported main collection. Pilot calls informed scaffold expansion and later supplied cache hits to that collection (Supplementary A.5).

Construction permits one-turn interventions; §8 limits the resulting population claim.

4.1 Conditions

Every item has the same skeleton, with two speakers in fixed roles: a distractor exchange, then the **proposer** states the **target proposition**, the **partner** produces the **target response turn**, the **proposer** makes a downstream **commitment** proceeding as though the target were settled, and the partner confirms it. Conditions differ only at the target response turn, or are byte-exact truncations of one that does.

Table 4: The eleven conditions. For acted-on conditions, expected policy state and gap are queried at the turn-4 action; for no-action conditions, they are queried at dialogue end. Every item shares one skeleton and differs only at the target response turn, or is a byte-exact truncation of an item that does.

Cond.	Target response turn	DAMSL	Expected policy state	Expected gap
E1	explicit confirmation	`aa`	`policy_threshold_met`	none
E2	paraphrase / display of understanding	`bf`	`policy_threshold_met`	none
E3	relevant next turn presupposing it	`sd`	`acknowledged` †	`premature` †
E4	**bare back-channel**, then a commitment	`b`	`acknowledged`	`premature`
E4C	bare back-channel, conversation ends	`b`	`acknowledged`	none
E5	hedged / conditional uptake	`aap`	`acknowledged`	`premature`
E6	topic shift, conversation ends	`qy`	`proposed`	`unconfirmed`
E7	repair initiated	`qw`	`disputed`	`unresolved_dispute`
E8	explicit rejection	`ar`	`rejected`	none
E9	no partner turn; proposer elaborates alone	—	`proposed`	`unconfirmed`
E10	topic shift, then a downstream commitment	`qy`	`proposed`	`false_grounding`

† E3 alone has an expected label that depends on the `RELEVANT_IS_STRONG` setting: under `True` a relevant next turn counts as strong-positive evidence and E3 expects `policy_threshold_met` with no gap. Main-sweep pipeline quantities that can change with this switch are reported under both settings. AAER-P is configuration-independent; weak-only mechanism probes name their setting.

The E7 and E8 labels are historical action-time labels, not claims about dialogue-end interpretation; neither condition supports a dialogue-end construct-validity claim.

SWBD-DAMSL anchors the target-response tags: `b` and `aa` are different acts. It does not supply our grounding states or gaps. Those are theory-derived operational labels. Agreement on the 42-tag scheme is 84%, $\kappa = 0.80$, but no pair-specific figure is published for the confusable `b`/`aa` distinction (§2.4). §4.6 defines the operational step and its scope.

4.2 The sufficiency two-by-two

Four conditions cross ambiguous-versus-absent evidence with downstream reliance, operationalising sufficiency for current purposes:

Table 5: The sufficiency 2×2: ambiguous-versus-absent evidence crossed with acted-on-versus-not.

	acted on	not acted on
ambiguous evidence	E4 → `premature`	E4C → no gap
no evidence	E10 → `false_grounding`	E6 → `unconfirmed`

E4C and E6 are the controls against scoring well by flagging everything: a detector firing on any ambiguous evidence gets E4 right and E4C wrong, one firing on any missing confirmation gets E6 right and E4C wrong.

4.3 Minimal-pair construction

Generation fails on any violation; 96 assertions cover the three derivations.

Table 6: Generation-time invariants. Generation fails rather than warns when one is violated.

Axis	Conditions	Assertion
`target_response_swap`	E1, E2, E3, E4, E5, E7, E8, E10	all eight differ **only** at the target response turn
`truncation`	E4 → E4C, E10 → E6	each is a strict byte-identical prefix of its source
`deletion`	E9	shares the prefix through the target; no later partner turn

This licenses within-variant causal contrasts such as E1 versus E4. It does not license a token-level reading across surface variants, which differ throughout the dialogue.

4.4 Scaffolds and the commitment-distance factor

Each scaffold uses a different domain and register.

Table 7: The six scaffolds. Each is a different domain and a different register.

Scaffold	Domain	Register
`sprint_slack`	sprint planning	terse Slack messages
`logistics_meeting`	scheduling and logistics handover	full-sentence meeting transcript
`spec_review`	technical specification review	written review-thread comments
`budget_approval`	departmental budget approval	formal finance correspondence
`incident_triage`	production incident triage	urgent incident channel
`client_handoff`	client account handover	email thread, polite business prose

Three scaffolds place the commitment's confirmation one turn after it, three place it three turns after. The manipulation is confounded with anaphoric re-mention and identifies nothing; it is reported in supplementary A.10.

4.5 The oracle test

Before model collection, perfect perception was passed through the state machine and gap rules at each condition's declared decision point: turn 4 when the target is acted on, dialogue end otherwise. The oracle clears 1.000 on both policy-state and gap exactness under both sensitivity settings over 528 rows, with zero unreachable conditions. Thus pipeline shortfalls are not failures to express the theory-derived labels. Those are action-time labels for acted-on conditions; the oracle tests internal expressibility, not dialogue-end interpretation or construct validity. The oracle prompted rule changes, never relabelling (§3.4).

4.6 Construct definition and scope

The construct is policy-sufficient evidence of understanding for a current purpose, not content acceptance, public commitment, receipt or authority. Clark and Brennan [5] treat acknowledgements as positive evidence whose required strength depends on the current purpose. E4 therefore expects `acknowledged`, but reliance requires stronger evidence and makes the gap `premature`. An "error" means nonconformance with this policy, not a universal reading of the dialogue.

The frozen mapping stipulates that a bare back-channel falls below the threshold, that a reformulation or explicit confirmation crosses it, and that reliance below it is reportable. The measurements instead ask which evidence a deployment cites, from which turn, and which state follows. A deployment that labels E4 `backchannel` and cites nothing later conforms regardless of its other beliefs. Expected labels were assigned by condition before model collection. Surface wording and model output cannot change them.

The anchors are largely spoken and the setting is two-party text, so the transfer needs stating. Clark and Brennan [5] make grounding cost a function of the medium's constraints. Our scaffolds mix synchronous and asynchronous text: both remove audibility and prosody, while cotemporality, reviewability and revisability vary by channel. From SWBD-DAMSL we borrow only what the text channel already carries (§2.4). Turn-taking pressure does not transfer, so delay and silence carry information these items do not model (§8).

SWBD-DAMSL anchors that boundary at the dialogue-act level. Its manual separates `b` (acknowledge/back-channel) from `aa` (agree/accept), notes that *yeah* can serve either function, and instructs coders not to mark an isolated *yeah* as `aa` without an additional utterance indicating agreement. We use the designed sequence, not the token alone, to assign E1's explicit confirmation to `aa` and E4's isolated back-channel to `b`.

The later confirmation responds to a commitment that presupposes the target. Following Stalnaker [42], we treat it as at most `relevant`, not `confirmation`: it may change dialogue-end state but cannot make earlier reliance satisfy the policy.

These mechanisms make the operationalisation reproducible; they do not establish that the theory-derived mapping exhausts how the designed sequences can be read. A text-only *yeah* can be pragmatically ambiguous and can function as acceptance in another sequence. The rates measure model behaviour against that operationalisation, not error prevalence in natural conversation. The controls separate unambiguous conditions (§6.3), and §6.4 triangulates `b` versus `aa` using older annotations; neither validates E4 or `premature` on our items. §8 states what would.

5 Method

5.1 Design

The main design crosses four model cells with five systems at five repeats.

Table 8: The four model cells, crossed with five systems over the frozen 264-item benchmark at five repeats.

Cell	Deployment	Reasoning	Surface variants	Items
M1	`gpt-5.4-nano`	off (`reasoning_effort: none`)	v1–v4	264
M2a	`gpt-5.4-nano`	`medium`	v1, v2	132
M2b	`gpt-5.4-nano`	`high`	v1, v2	132
M3	`DeepSeek-V4-Flash`	provider default	v1–v4	264

M1/M2a/M2b differ only in reasoning effort; M3 differs in family and other properties. The main design has 19,800 records. A **deployment family** is a base model, so runtime settings share one family; the seven cells span five families, and results use the coarser vendor grouping where relevant.

An S1 extension adds M4a (`claude-haiku-4-5`) and M4b (`gemini-3.5-flash-lite`) on E1/E4/E4C/E6. A fresh M1/M4a/M4b window on E2/E4/E5 tests discrimination versus global strictness, with M1 supplying an E4 bridge cell.

Six later collections add a frontier M5 cell (`claude-opus-5`, default effort); natural Switchboard excerpts; hedge, instruction and prompt-rewording arms; and the four-arm study crossing full versus turn-4-prefix input with incumbent versus instructed prompts (§6.2–§6.5). Each collection has its own pre-specified runner and declared repeat indices.

5.2 Systems

Table 9: The five systems. Identical inputs; they differ in where the boundary between learned and symbolic components falls.

System	Architecture	Model call	Evidence layer
S1	perception → finite-state machine → gap rules	yes	yes
S2	end-to-end: one call returns states and gaps	yes	no
S3	perception → naive rule (any positive evidence grounds)	shared with S1	yes
S4	model assigns states; gap rules still run, no FSM	yes	yes
S5	floor: assign one state to everything	no	no

S3 shares S1 perception, so unique-call analyses count them once. S2/S5 have no evidence layer and enter no perception metric. Headline perception, masking and direct contrasts use S1; the other systems diagnose where errors arise.

Prompt disclosure. The archive ships the exact hash-addressed `perception_v2` and `perception_v2_instructed` prompt files under *groundtracker/prompts*, including the structured example; every response records the prompt id and SHA-256. The principal prompt requests JSON propositions, evidence records citing a later turn of the other speaker, and the earliest later `acts_on` turn; it defines all seven Table 3 labels. The instructed prompt differs by one inserted paragraph operationalizing the manual's isolated-*yeah* rule for listed similar tokens.

5.3 The paired within-item design

Cross-cell estimates average repeats before pairing shared items. Primary intervals use 10,000 scaffold-then-variant resamples [11], with item and scaffold sensitivities. Six outer scaffolds yield only 64 sign assignments. The prefix family was not pre-specified, so we report exact flips and a post-hoc two-claim Holm sensitivity; the separate nine-contrast family uses synchronized 2^6 assignments and its family maximum (Supplementary A.4).

5.4 Metrics

Table 10 separates displayed understanding, policy-sufficient response evidence, action-time policy sufficiency, terminal state and their history.

Table 10: Construct-to-measure map. The first two constructs concern the target response turn; the next three concern state over time.

Construct	Operational measure	Boundary
Uptake	E4 reaches `acknowledged`	items not independently judged
Policy-sufficient response evidence	strong target-response evidence	public commitment not measured
Action-time sufficiency	prefix state at turn 4	not authority or permission
Terminal state	full-dialogue state	may differ from action-time state
Warning history	insufficient at action, later `policy_threshold_met`	needs a perceived action

Metrics remain separate. Units are runs except attachment and validity, which count evidence records. d' follows standard signal-detection theory [17]; extreme rates receive the log-linear correction defined in Supplementary A.4.

Terminal scoring is the post-hoc consumer view. Time-filtered replay removes evidence after turn 4 although perception saw the full transcript; prefix scoring reads only through turn 4. The event-sourced detector retains state and gap at each perceived `acts_on` turn after terminal state changes.

For M1/S1, target-response strength and the exclusive post-commitment term are disjoint and sum to masking; incidence overlaps them. Strength means a strong-positive policy label, not agreement or public commitment. The oracle clears 1.000 state and gap accuracy under both settings.

AAER-P is each cell's S1 share whose target turn's first serialized evidence is `confirmation` or `repeat`, a configuration-independent label set. Pipeline metrics replay every record, so post-commitment `relevant` evidence can change masking.

5.5 Resolution

Reasoning contrasts use 12 E4 items and M3 uses 24, all in six scaffolds. Observed hierarchical resolution is about 0.20–0.27; this describes the instrument rather than prospective power, as Miller [28] recommends.

5.6 Execution, controls and coverage

Evidentiary roles. The representational result combines the analytic rule with identical-output replay. Prefix, instruction, hedge and prompt-rewording collections are descriptive probes, not deployment-stability evidence.

The main sweep, extension and fresh control record pre-flights; side collections do not. Main cells are interleaved; excluding 19/19,800 failures leaves coverage at least 0.998. Its closing canary reused the opening key and cannot measure within-window drift. The extension answered 960/960, passed distinct-index canaries and cost $0.58. Rates describe frozen windows, whereas identical-output replay is invariant to later deployment drift.

The fresh control (indices 5–9) has zero cache hits, passes its distinct-index gate, and answers M1 120/120, M4a 360/360 and M4b 359/360 records.

Six same-window canaries show ordinary label movement. Fresh control, natural probe and frontier collections pass; the other interleaved probes lack an independent canary, limiting time-order confounding without separating drift from non-determinism (Supplementary A.15).

Nine record-producing collections hold 25,800 records, of which 25,776 answered, at a total cost of $20.08. Coverage is at least 0.997 in every collection. Supplementary A.19 lists each response file and cost source.

5.7 Ethics

No participants were newly recruited and no identifiable private information was collected. The 264 benchmark dialogues are constructed and the people, projects and organisations in them are invented; the item generator and its seed are in the archive. The natural-dialogue probe (§6.4) is secondary analysis of the publicly distributed, de-identified Switchboard Dialog Act Corpus, annotated decades ago; no new annotation was gathered and no speaker was contacted. Under the governing policy, research using these existing de-identified materials without interaction or intervention is not human-subjects research, so no review was sought. A study in which people judge our items would be, and §8 says what that would require.

5.8 Reproducibility

Responses are frozen and rescoring is offline. A synthetic sweep must recover 52 planted answers. The anonymous supplementary archive accompanying the submission contains responses for the constructed studies and text-withheld structural records for the natural probe, plus the benchmark, analysis, tests and claim registry. Supplementary A supplies the deployment, metric, collection and reproduction details cited from the paper.

AI-assisted research. OpenAI Codex and GitHub Copilot assisted with code, tests, analysis, provenance tooling, manuscript revision and initial benchmark dialogue text. The frozen item manifest attests to a triaged human review of the generated benchmark: mechanical checks covered rule-testable properties, while a person read interludes and content those checks could not verify. This was not independent item annotation. The authors chose the methods, reproduced outputs from frozen inputs and remain responsible for the study.

6 Results

The frozen sweep contains 19,800 intended records; 19 failed records leave 19,781 scored, and every cell has at least 0.998 coverage. Results use the weak setting unless marked. AAER-P is the share of E4 runs whose first serialized target-response record is `confirmation` or `repeat`. S1 pipeline metrics are replayed under each evidence table, so their strong-setting values can differ. Masking — the target reaching `policy_threshold_met`, after which `premature` is unreachable — runs at 0.756 on the focal cell M1, which is one of only two cells that ran the complete design (Table 8) and the highest-masking of the seven.

6.1 Back-channel classified as policy-sufficient

M1/S1 assigns a strong-positive target-response label on 32/119 answered E4 runs: AAER-P = 0.269, hierarchical 95% CI [0.100, 0.487], over 24 items and six scaffolds. AAER-P is the stable registry identifier; §5.4 defines its target-response strength estimand and rules out a public-commitment reading.

Counting shared S1/S3 perception once and S4 separately gives 55/239 = 0.230 [0.084, 0.417]; this prompt-mixed quantity is a sensitivity, not a pooled headline. Supplementary Table 29 gives all interval constructions.

6.2 The deployment mis-ranks evidence in both directions

M1/S1 credits E2 paraphrase on 31/119 runs and E4 back-channel on 32/119 under the weak table. The rates are similar, not the errors: E2 misses 88 runs while E4 falsely credits 32. The weak d′ point is −0.025, but its interval [−0.713, 0.611] supports no directional claim. Under strong, where `relevant` counts, E2/E4 d′ is 2.852 [2.301, 3.586].

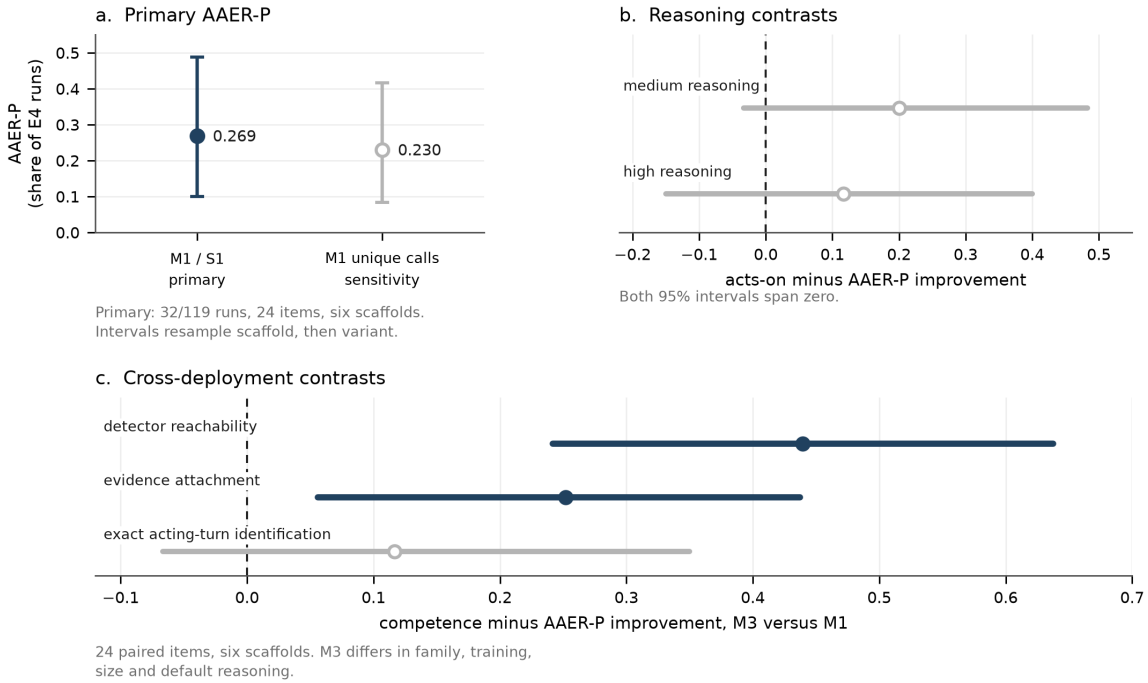

Fig. 2: (a) Primary AAER-P and the unique-call sensitivity reported above. (b) Direct reasoning interactions, both unresolved. (c) Direct M3 interactions; the intervals drawn are unadjusted, and only reachability survives the family of nine. Panels (b) and (c) belong to the exploratory contrast family reported in Supplementary A.

Giving the taxonomy a hedge label measures what that omission costs. In a fresh interleaved window, adding one label and one gloss sends 118/119 E5 runs to hedge and its strong-positive rate from 0.345 to 0.000, paired −0.342 [−0.508, −0.192]. The change is not local: E1 is unmoved at 1.000 with a paired delta of exactly zero, but E4's strong-positive rate rises from 0.292 to 0.467, paired +0.175 [+0.092, +0.275], with E4 `backchannel` falling from 82 runs to 63 while `confirmation` rises from 35 to 56. Had E4's error been an artefact of having nowhere to put a hedged uptake, the repair would have lowered it. AAER-P is therefore a rate under a named label set, and every rate below is relative to the taxonomy in Table 3.

M1/S1 separates explicit confirmation from back-channel: E1/E4 $d' = 3.251$ [2.871, 3.701], with corrected rates 0.996 and 0.271. Strong scoring repairs E2 without changing E4, so evidence type rather than a global state threshold controls the contrast. Supplementary Table 30 gives all pairs.

Item propensities remain heterogeneous: 34 of 72 S1 item/cell pairs are zero and seven are one, which motivates item and scaffold resampling.

6.3 Terminal state can erase an action-time warning

On M1/S1, 90/119 E4 targets reach `policy_threshold_met`: masking rate 0.756, hierarchical 95% CI [0.567, 0.924], over 24 items and six scaffolds. Once there, the target cannot emit `premature`.

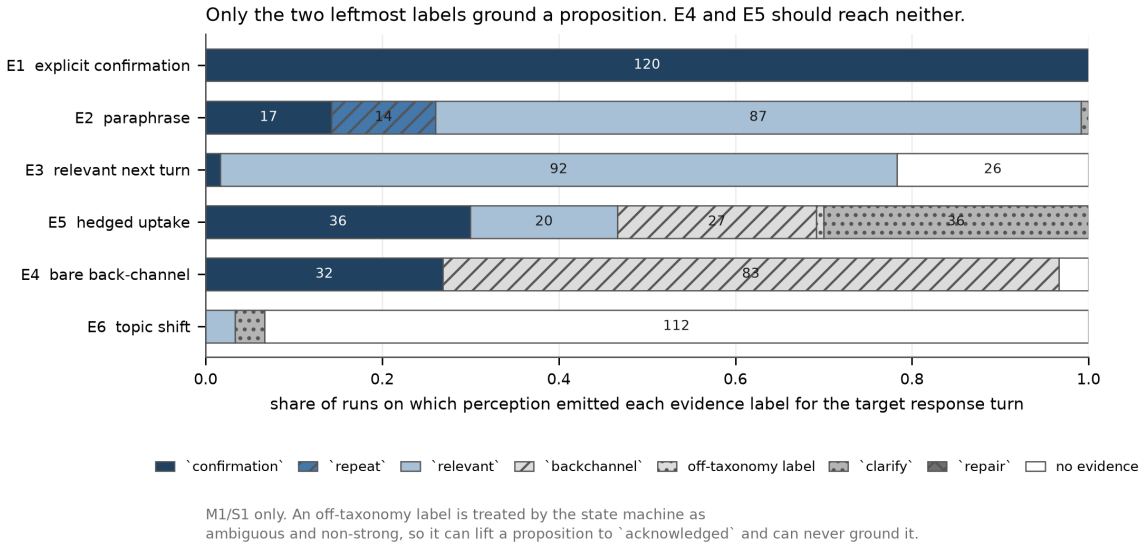

Fig. 3: M1/S1 target-turn labels. Only `confirmation` and `repeat` count as strong-positive in AAER-P. E2 is under-credited while E4 is over-credited.

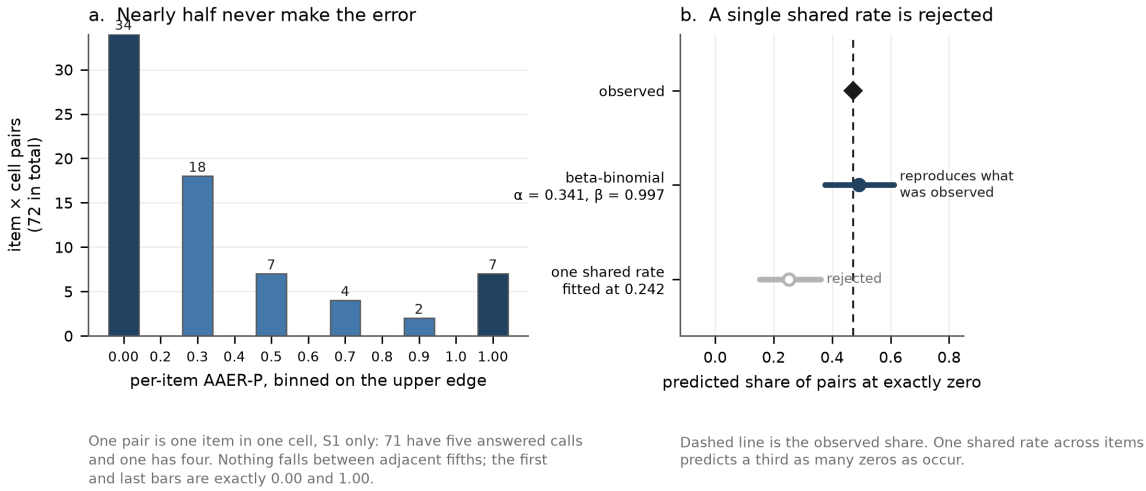

Fig. 4: S1-only dispersion. The mean does not describe a typical item; the beta-binomial sensitivity captures the broad propensity distribution.

Of 90 masked M1/S1 runs, 19 use the target response turn alone, 13 use both routes and 58 use only post-commitment evidence. Over all 119 answered runs, acceptance strength contributes 32/119 = 0.269 [0.100, 0.487] and the exclusive post-commitment partition contributes 58/119 = 0.487 [0.275, 0.690]; they are disjoint and sum to 0.756. This is a partition by evidence position, not a factorized causal mechanism.

Under the strong table, masking is 96/119 = 0.807: target-response strength remains 32 and the exclusive post-commitment partition rises to 64/119 = 0.538. Weak and strong temporal incidence are 71/119 and 85/119; unlike the partition, incidence includes both-route runs.

Filtering the original full-transcript perception output to evidence cited no later than turn 4 yields 32/119 = 0.269, but the model had seen the later turns.

A fresh four-arm study separates replay from true action-time inference. Terminal masking is 92/120 for full/incumbent, 49/120 for prefix/incumbent, 56/120 for full/instructed and 19/120 for prefix/instructed; turn-4 replay of the full arms gives 34/120 = 0.283 and 11/120. Exact-prefix incumbent masking is 49/120 = 0.408, hierarchical 95% CI [0.217, 0.625], over 24 items and six scaffolds. The interleaved arms have no independent canary (§5.6).

True prefix exceeds full-context replay by +0.125, hierarchical 95% CI [+0.008, +0.250], but exact six-scaffold p = 0.125; it is below full-context terminal masking by -0.358 [-0.567, -0.133]. E1 remains 120/120 strong-positive.

The instruction lowers prefix masking from 0.408 to 19/120 = 0.158, paired -0.250 [-0.408, -0.108], exact p = 0.03125, with E1 at 1.000. No family was pre-specified; a post-hoc two-claim Holm sensitivity gives p = 0.0625. Thus timestamp filtering and prefix perception answer different questions, no stable directional context mechanism is established, and the residual is an action-time target-response strength error under the policy.

Identical-output replay changes storage and query semantics, not perception. The terminal detector retains 14 warnings in 119 E4 runs; 99 perceived-action events retain 71, including 57 later overwritten by `policy_threshold_met`.

At constructed turn 4, 87 snapshots are insufficient and 58 later overwritten. Of 20/119 runs without events, 15 are insufficient; 97 events cite turn 4 and two later, one after sufficiency. This reconciles 87 to 71 and 58 to 57. Events neither cover every action nor enable policy-aware recomputation.

Two sensitivities bound alternative explanations. Restoring 73 rejected same-speaker records affects 64 runs but changes masking by zero and removes 7/57 historical losses (0.123).

Recoding post-commitment `confirmation` in 56 runs removes 51/90 weak-setting masked cases, leaving 39/119 = 0.328; strong masking remains 96/119 = 0.807. Thus the 58-run position partition combines temporal scope, attachment and evidence strength rather than isolating a timing effect.

Masking is one of two ways the warning fails to arrive: the 15 rule-miss runs stay `acknowledged` and emit nothing either. Combining them, no `premature` warning reaches the interface on 105/119 = 0.882 of M1/S1 E4 runs, hierarchical 95% CI [0.769, 0.967], over 24 items and six scaffolds; the warning arrives on 14. Under strong scoring the composite is 107/119 = 0.899. The broad end-to-end detector remains weak: `premature` recall is 53/359 = 0.148 across all expected cases, and `rejection_masking_rate` is 0.000. Supplementary Table 31 gives the per-system partition.

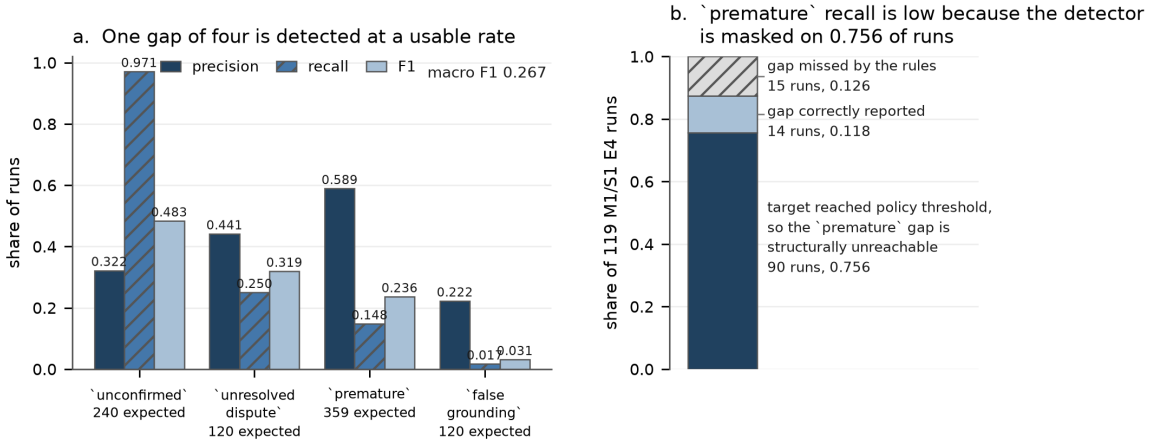

Fig. 5: M1/S1 throughout. (a) Gap precision, recall and F1. (b) The E4 masking partition: most targets ground before `premature` can fire.

The same decomposition over seven deployment cells:

Table 11: Weak-setting E4 masking by deployment cell. Target-response strength and exclusive post-commitment partitions are disjoint and sum to masking before rounding. AAER-P equals target-response strength in these cells but is defined separately (§5.4). These cells use full-transcript perception and terminal replay; only the separate M1 prefix study above measures exact-prefix action-time inference. M5 is the frontier anchor and ran the reduced four-condition design. The Vendor column groups cells more coarsely than family does: M4a and M5 are two families from one vendor. Hierarchical 95% intervals in brackets.

Cell	Vendor	Masking	AAER-P	Target-response strength	Exclusive post-commitment
M1	OpenAI	**0.756** (90/119)	**0.269**	0.269 [0.100, 0.487]	0.487 [0.275, 0.690]
M2a	OpenAI	0.683 (41/60)	0.150	0.150 [0.050, 0.250]	0.533 [0.300, 0.750]
M2b	OpenAI	0.683 (41/60)	0.133	0.133 [0.000, 0.300]	0.550 [0.350, 0.750]
M3	DeepSeek	0.367 (44/120)	**0.317**	0.317 [0.158, 0.508]	0.050 [0.000, 0.117]
M4a	Anthropic	0.400 (48/120)	**0.192**	0.192 [0.000, 0.442]	0.208 [0.042, 0.400]
M4b	Google	0.008 (1/120)	**0.008**	0.008 [0.000, 0.033]	0.000 [0.000, 0.000]
M5 frontier	Anthropic	**0.617** (74/120)	**0.108**	0.108 [0.000, 0.258]	0.508 [0.258, 0.758]

M4b's zero-width interval on the zero post-commitment rate is a bootstrap saturation artefact. Across the same reduced design, M4a-to-M5 target-response strength falls 0.192 to 0.108 while the post-commitment partition rises 0.208 to 0.508 and masking 0.400 to 74/120 = 0.617. M5 answers 480/480 records, with E1 at 1.000 and E6 at 0.000; intervals overlap interior cells. M3 differs in family, training, size and default reasoning; M4a/M5 also differ in effort, decoding and tokenizer.

The fresh M1F bridge masks 88/120 = 0.733, close to 0.756. In that window, M4aF E2/E4 d′ is unresolved at 0.276 [-0.443, 1.066], whereas M4bF reaches 3.309 [2.850, 3.904] by crediting E2 on 90/120 runs and E4 on none.

No cell grounds E6 in 720 weak-setting runs and E1 spans 0.992–1.000, ruling out global abstention (Supplementary Table 32).

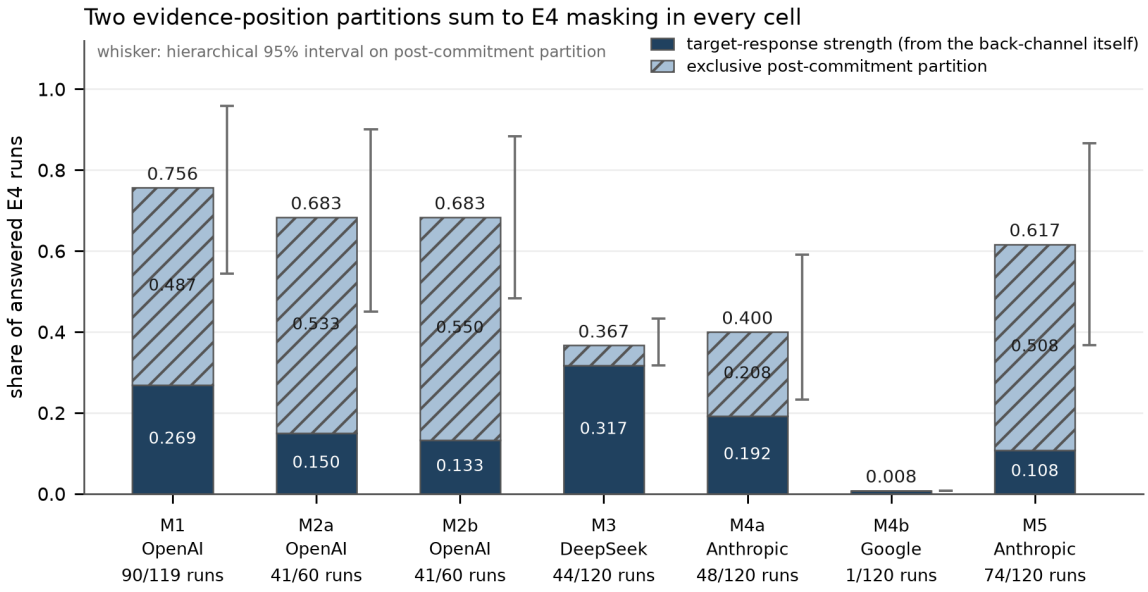

Fig. 6: The E4 masking rate by cell, split into target-response strength below and the exclusive post-commitment partition above, so each bar's height is that cell's masking rate. The split records evidence position; it is not a causal factorization. Whiskers carry the hierarchical interval on the post-commitment partition. Figure 5(b) cuts M1's same 90 masked runs by detector outcome; this shows a different decomposition across all seven cells including the frontier anchor.

6.4 The boundary against human annotations

§4.6 anchors the contested E4 assignment in the SWBD-DAMSL manual and then measures it only on dialogues we wrote. The Switchboard Dialog Act Corpus carries the manual's labels applied by trained coders to natural talk, assigned years before this study by people with no contact with it. A mechanical rule fixed before any excerpt was read selects every utterance tagged exactly b or exactly aa that follows a statement by the other speaker, is followed by the other speaker again, and is at most five words. Over the corpus's 1155 transcripts, 16,576 utterances qualify: a share of 0.941 are coded b and 983 aa.

The 223,606-utterance exact-tag census gives 36,180 b to 10,136 aa (3.57); the isolated-response filter raises it to 15.86, and three- or eight-word caps give 17.28 and 15.63. Among attested E4 forms it is 5,552 to 121, including 3,727 to 28 for isolated *yeah*. Thus sequential position and surface conditioning, not the five-word cap alone, drive the imbalance; Supplementary A.18 gives the full split.

Seeded sampling retains the earliest eligible candidate per conversation, then draws 120 four-turn excerpts, 60 per class. Three S1 deployments run each item three times. Rates condition on anchoring a proposition at the statement turn.

Table 12: The target-response strength boundary on human-annotated natural dialogue. NB carries the coders' b tag, NA their aa. The rate is the share of runs whose evidence record at the response turn is strong-positive, over runs that anchored a proposition at the statement turn; the sensitivity in brackets is the same rate over all answered runs. The d′ column uses those bracketed all-answered rates, not the adjacent conditional rates. Conversations are the resampling unit and each supplies one item.

Cell	Vendor	NB rate	NA rate	d′, all answered (NA vs NB)
M1N	OpenAI	**0.151** (18/119) [0.100]	0.679 (72/106) [0.402]	1.023 [0.591, 1.559]
M4aN	Anthropic	0.102 (18/177) [0.100]	0.678 (118/174) [0.656]	1.667 [1.209, 2.259]
M4bN	Google	0.019 (3/161) [0.017]	0.567 (85/150) [0.472]	1.998 [1.426, 2.928]

All intervals exclude zero in the predicted direction, triangulating the dialogue-act distinction but not our items or action-time policy. M1N and M4bN credit human-tagged back-channels on 18/119 = 0.151 and 3/161 = 0.019 runs. M1's natural d′ of 1.023 is below its constructed E1/E4 value of 3.251, and every deployment misses 0.32–0.43 of human-tagged agreements.

M1N's 0.151 is below constructed AAER-P 0.269, but unpaired populations preclude a tight bound. The corpus supplies neither a downstream commitment needed to validate `premature` nor judgments of our items (§8).

6.5 Stating the rule

The main sweep uses a prompt that defines `backchannel` inside a taxonomy and never says what a bare *yeah* implies, so "cannot distinguish" and "was never asked to" are not separated there. `perception_v2_instructed` inserts a single paragraph that operationalizes the manual's isolated-*yeah* rule and extends its boundary to listed similar tokens; it changes nothing else. The two arms are interleaved in a separate window at repeat indices 50–54, over the 24 E4 and 24 E1 items, with zero cache hits. Supplementary Table 34 gives both arms.

Stating the rule removes most of the target-response strength error, taking AAER-P to 9/120 = 0.075, and it does so without moving the criterion: E1 is credited on 120/120 runs in both arms and the paired E1 delta is exactly zero, so this is discrimination and not a suppressed label. d′(E1 vs E4) rises from 3.161 to 4.056.

The instruction also reduces masking from 87/120 to 0.425, paired -0.300 [-0.433, -0.175], but leaves 51 of 120 runs masked. Its exclusive post-commitment partition falls from 0.417 to 0.350: a deployment that has been told the rule still marks the target `policy_threshold_met` after grounding on evidence that arrives after the commitment, because the rule says nothing about when evidence may be read. The instruction repairs one route and reduces, but does not remove, the other.

Generated values and scopes are bound to `config/claim_registry.json`; full per-cell and sensitivity outputs are in `results/phase_d/analysis/`.

7 Discussion

7.1 A focal affected party is the person put on record

On M1/S1, 90/119 E4 runs are masked, rate 0.756 [0.567, 0.924], across 24 items and six scaffolds, leaving no `premature` warning. A later reader may see "Kai agreed" or an assigned action without the weak or late basis. The record cannot

expose that difference; harm is not measured. This links AI-mediated attribution and trust studied by Hancock, Naaman & Levy [18] and Hohenstein & Jung [19] to the person recorded and later readers, within peer-like pairs and without claims about authority, consequence or remedy.

The identical-output control measures **historical snapshot availability**: the event preserves 57 warnings erased by terminal state. It retains the frozen-policy snapshot only when perception identifies and timestamps an action; 20/119 are missed and two mistimed. This establishes bounded availability, not complete detection, unique minimality, policy-aware recomputation or successful contestation.

S1's retained, turn-indexed evidence table is itself one candidate basis for a historical query; the event is not uniquely required. Full-context filtering gives 0.283 masking while exact-prefix perception gives 0.408, so post-hoc filtering and preserved action-time inference are not interchangeable.

7.2 The display has costs and actions

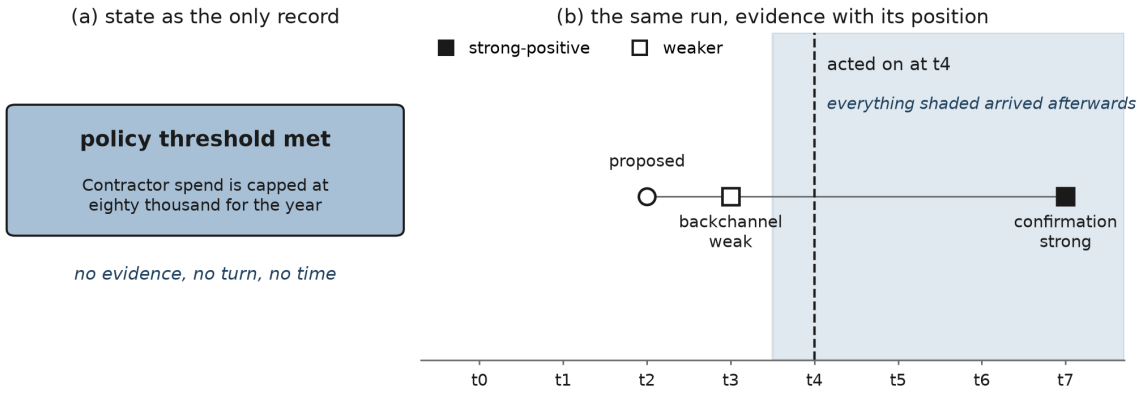

Fig. 7: Conceptual diagnostic view, not an evaluated interface. One masked M1/S1 E4 run appears under both representations. (a) A policy state stored without its basis. (b) The same run's evidence on the turn axis. The back-channel at t3 is weak; the strong confirmation arrives at t7, after reliance at t4, which is the distinction (a) cannot express. No review workflow is evaluated here.

Figure 7 suggests testable interface choices: restrict evidence links to claims that trigger decisions, show source context on demand, or route weak and late cases for confirmation. Their screen-space, review-time and behavioural effects are unknown. Ingestion time, model and prompt provenance, supersession, access and integrity controls are candidate audit fields, not requirements established by this experiment.

7.3 From provenance to contestation

Provenance is necessary but not sufficient for contestability. Notice, standing, correction, remedy and accountable ownership are separate governance variables that this study neither implements nor ranks. Retained time-indexed information is a precondition for recomputing or inspecting this system's action-time basis; it is a design for none of those governance choices.

The same record can protect and surveil. Retaining full source context may expose bystanders, clients, hesitation or disagreement to managers long after the exchange. Minimal event metadata, purpose limitation, retention tiers, access

logs, correction and deletion are alternatives for a future governance study. Nor is explicit confirmation free: it can shift review labor onto lower-power participants or turn coerced assent into stronger documentation.

7.4 What the computation establishes

The hybrid split establishes attribution, not a general accuracy advantage. M3 nearly removes the post-commitment-only partition while worsening the registered target-response strength component; reasoning contrasts remain unresolved. Family, training, size and default reasoning remain confounded.

Two alternatives close empirically. The instruction cuts the target-response strength error without changing E1 in the tested collection, but terminal masking remains because late evidence is a different problem (§6.5). Human annotations triangulate the b/aa dialogue-act distinction, not our action-time policy (§6.4). Restoring same-speaker evidence removes 7 of 57 historical warning-loss cases and leaves terminal masking unchanged.

A classifier repair addresses the label; retained history addresses timing. The magnitude of weak-setting masking is taxonomy- and policy-contingent, while the identical-output control establishes historical queryability rather than a taxonomy-independent rate or unique representation. The bounded pre-interface result does not establish workflow benefit. Interface effects remain outside the evidence.

8 Limitations

Construct and population. SWBD-DAMSL supplies dialogue acts and grounding theory supplies states and gaps; the contested E4 mapping remains theory-derived. §6.4 triangulates b/aa, but nobody outside this work judged our items, no pair-specific reliability is published, and `premature` has no external anchor. The oracle tests internal expressibility, not construct validity. The primary analysis has 24 E4 items across six semantic scaffolds. Constructed English cannot estimate prevalence or capture how sequence, prosody, culture and hierarchy change a back-channel's force; back-channels remain pragmatically ambiguous.

Two-party scope. Multi-party grounding raises different questions about whose acceptance is required and what silence means. Results apply to peer-like pairs, not group agreement. Items encode no deontic roles; policy sufficiency establishes neither permission nor authority to act.

Medium. The theory and annotation anchor are largely spoken; the items are written. §4.6 restricts the borrowing to what text carries, but does not test that transfer. Every item answers immediately, leaving delay, silence and revision before sending outside the design.

Models and time. Seven cells span five families but confound scale, training, effort, tokenizer and decoding. OpenAI Codex helped draft initial items, leaving provider relatedness to three cells unisolated; their high masking opposes a simple same-provider advantage. The main canary reused a key, and four interleaved collections lack independent canaries (§5.6). Without build IDs or pinned temperature, results support within-window contrasts, not stability.

Inference. Six scaffolds yield 64 sign assignments. The post-collection family-max result sits at the resolution floor and exact Holm does not support it. No prefix family was fixed: the +0.125 context contrast has exact $p = 0.125$; instruction $p = 0.03125$ becomes 0.0625 under the post-hoc Holm sensitivity. With six outer clusters these are descriptive. The contrasts support neither prevalence nor stability and are not premises of the analytic entailment or identical-output control.

Instrument boundaries. Adding `hedge` repairs E5 but raises E4 by +0.175, so rates remain relative to the Table 3 labels; one alternative taxonomy and one instruction do not sample either design space. Four frozen prompt wordings leave the coding-manual arm unmoved (Supplementary A), which rules out knife-edge dependence on one wording

but not prompt dependence in general; prior work finds sensitivity to order [26], formatting [40], and paraphrase [29]. Restoring inadmissible same-speaker evidence removes 7 historical-loss cases but defines no deployed policy. Commitment distance remains confounded, and the mixed comparison shows no general hybrid advantage. The taxonomy has no gap for reliance on an already `rejected` proposition, so E8 expects none by construction.

Deployment and governance. The offline event preserves a frozen-policy snapshot when perception identifies and timestamps an action, but misses 20 of 119 actions and mistimes two. It neither supports policy-aware recomputation nor establishes a uniquely minimal representation. Evidence logs and equivalent histories are not compared; Figure 7 is conceptual. Retention may aid auditing while increasing surveillance, privacy loss and review burden. No governance policy follows.

9 Conclusion

Grounding trackers can turn weak or late evidence into a settled state and erase that earlier reliance did not yet meet a stated policy. Exact-prefix perception shows what one tested model inferred without future dialogue; event-sourced replay shows that the historical warning can survive a later terminal-state change. A coding instruction descriptively reduces the target-response strength error, while the temporal problem requires preserving the frozen-policy answer at the action. The tested event does so only when perception identifies and timestamps that action; policy-aware recomputation would require evidence and a policy version. A system required to answer this historical query cannot let state replace the time-indexed basis from which that answer is obtained. Evidence fields are a proposed audit extension, not an evaluated governance design. The benchmark establishes conformance failures against a constructed two-party policy; natural dialogue triangulates its `b/aa` distinction, not the prevalence or organizational meaning of those failures.

References

[1] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Article 254. doi:10.1145/3290605.3300484

[2] Erin Beneteau, Olivia K. Richards, Mingrui Ray Zhang, Julie A. Kientz, Jason C. Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Article 243. doi:10.1145/3290605.3300473

[3] Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A Semantically-Based Standard for Dialogue Annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*. 430–437. doi:10.63317/2vxg5i9ctqt8

[4] Herbert H. Clark. 1996. *Using Language*. Cambridge University Press. doi:10.1017/CBO9780511620539

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

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

[7] Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a Collaborative Process. *Cognition* 22, 1 (1986), 1–39. doi:10.1016/0010-0277(86)90010-7

[8] Mark G. Core and James F. Allen. 1997. Coding Dialogs with the DAMSL Annotation Scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*. 28–35. https://www.eecis.udel.edu/~carberry/CIS-885/Papers/Core-Allen-Coding-Damsl.pdf

[9] Barbara Di Eugenio, Pamela W. Jordan, Richmond H. Thomason, and Johanna D. Moore. 2000. The Agreement Process: An Empirical Investigation of Human–Human Computer-Mediated Collaborative Dialogs. *International Journal of Human-Computer Studies* 53, 6 (2000), 1017–1076. doi:10.1006/ijhc.2000.0428

[10] Mark Dingemanse, Seán G. Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S. Gisladottir, Kobin H. Kendrick, Stephen C. Levinson, Elizabeth Manrique, Giovanni Rossi, and N. J. Enfield. 2015. Universal Principles in the Repair of Communication Problems. *PLOS ONE* 10, 9 (2015), e0136100. doi:10.1371/journal.pone.0136100

[11] Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall.

[12] Michelle Elizabeth, Gwénolé Lecorvé, Lina M. Rojas Barahona, and Magalie Ochs. 2026. Conversational Grounding in Large Language Models: Evaluation Methods, Challenges and Future Directions. In *Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 151–163. https://aclanthology.org/2026.sigdial-1.11/

[13] Martin Fowler. 2005. Event Sourcing. MartinFowler.com. https://martinfowler.com/eaaDev/EventSourcing.html

[14] 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. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1307–1323. doi:10.18653/v1/2020.findings-emnlp.117

[15] Rod Gardner. 2001. *When Listeners Talk: Response Tokens and Listener Stance*. John Benjamins, Amsterdam. doi:10.1075/pbns.92

[16] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 517–520. doi:10.1109/ICASSP.1992.225858

[17] David M. Green and John A. Swets. 1966. *Signal Detection Theory and Psychophysics*. Wiley.

[18] Jeffrey T. Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (2020), 89–100. doi:10.1093/jcmc/zmz022

[19] Jess Hohenstein and Malte Jung. 2020. AI as a Moral Crumple Zone: The Effects of AI-Mediated Communication on Attribution and Trust. *Computers in Human Behavior* 106 (2020), 106190. doi:10.1016/j.chb.2019.106190

[20] Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A Benchmark Dataset for Meeting Summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 16409–16423. doi:10.18653/v1/2023.acl-long.906

[21] Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. *Switchboard-DAMSL Labeling Project Coder's Manual*. Technical Report 97-02. University of Colorado Institute of Cognitive Science. https://web.stanford.edu/~jurafsky/ws97/manual.august1.html

[22] Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the Difference That Makes a Difference with Counterfactually-Augmented Data. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=Sklgs0NFvr

[23] 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. 2024. Common Ground Tracking in Multimodal Dialogue. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*. 3587–3602. https://aclanthology.org/2024.lrec-main.318/

[24] Daniel N. Kluttz, Nitin Kohli, and Deirdre K. Mulligan. 2022. Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions. In *Ethics of Data and Analytics*. Auerbach Publications, Boca Raton, 420–428. doi:10.1201/9781003278290-62

[25] Alex Lascarides and Nicholas Asher. 2009. Agreement, Disputes and Commitments in Dialogue. *Journal of Semantics* 26, 2 (2009), 109–158. doi:10.1093/jos/ffn013

[26] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 8086–8098. doi:10.18653/v1/2022.acl-long.556

[27] Hannah Mieczkowski, Jeffrey T. Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–14. doi:10.1145/3449091

[28] Evan Miller. 2024. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. arXiv:2411.00640. doi:10.48550/arXiv.2411.00640

[29] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of What Art? A Call for Multi-Prompt LLM Evaluation. *Transactions of the Association for Computational Linguistics* 12 (2024), 933–949. doi:10.1162/tacl_a_00681

[30] Biswesh Mohapatra, Manav Nitin Kapadnis, Laurent Romary, and Justine Cassell. 2024. Evaluating the Effectiveness of Large Language Models in Establishing Conversational Grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 9767–9781. doi:10.18653/v1/2024.emnlp-main.545

[31] Luc Moreau, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Beth Plale, Yogesh Simmhan, Eric Stephan, and Jan Van den Bussche. 2011. The Open Provenance Model Core Specification (v1.1). *Future Generation Computer Systems* 27, 6 (2011), 743–756. doi:10.1016/j.future.2010.07.005

[32] Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. 2007. Detecting and Summarizing Action Items in Multi-Party Dialogue. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*. 18–25. doi:10.18653/v1/2007.sigdial-1.4

[33] Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 3727–3733. doi:10.18653/v1/N19-1373

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

[35] Antonio Roque and David R. Traum. 2008. Degrees of Grounding Based on Evidence of Understanding. In *Proceedings of the SIGDIAL Workshop on Discourse and Dialogue*. 54–63. https://aclanthology.org/W08-0107/

[36] Antonio Roque and David R. Traum. 2009. Improving a Virtual Human Using a Model of Degrees of Grounding. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 1537–1542. https://www.ijcai.org/Proceedings/09/Papers/257.pdf

[37] Emanuel A. Schegloff. 1992. Repair After Next Turn: The Last Structurally Provided Defense of Intersubjectivity in Conversation. *Amer. J. Sociology* 97, 5 (1992), 1295–1345. doi:10.1086/229903

[38] Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language* 53, 2 (1977), 361–382. doi:10.2307/413107

[39] Michael F. Schober and Herbert H. Clark. 1989. Understanding by Addressees and Overhearers. *Cognitive Psychology* 21, 2 (1989), 211–232. doi:10.1016/0010-0285(89)90008-X

[40] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://iclr.cc/virtual/2024/poster/18650

[41] Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. 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)*. 6279–6296. doi:10.18653/v1/2024.naacl-long.348

[42] Robert Stalnaker. 2002. Common Ground. *Linguistics and Philosophy* 25, 5–6 (2002), 701–721. doi:10.1023/A:1020867916902

[43] Melisa Stevanovic and Anssi Peräkylä. 2012. Deontic Authority in Interaction: The Right to Announce, Propose, and Decide. *Research on Language and Social Interaction* 45, 3 (2012), 297–321. doi:10.1080/08351813.2012.699260

[44] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics* 26, 3 (2000), 339–373. doi:10.1162/089120100561737

[45] Thibaut Thonet, Laurent Besacier, and Jos Rozen. 2025. ELITR-Bench: A Meeting Assistant Benchmark for Long-Context Language Models. In *Proceedings of the International Conference on Computational Linguistics (COLING)*. 407–428. https://aclanthology.org/2025.coling-main.28/

[46] David R. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph. D. Dissertation. University of Rochester. https://hdl.handle.net/1802/809

[47] David R. Traum and Elizabeth A. Hinkelman. 1992. Conversation Acts in Task-Oriented Spoken Dialogue. *Computational Intelligence* 8, 3 (1992), 575–599. doi:10.1111/j.1467-8640.1992.tb00380.x

[48] 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. 2025. 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)*. 40–50. doi:10.18653/v1/2025.naacl-demo.5

[49] Victor H. Yngve. 1970. On Getting a Word in Edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*. 567–578.