# WHAT IS A CONVERSATION MADE OF? A REUSABLE SEMANTIC–RELATIONAL STATE ACROSS SAFETY TASKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multi-turn safety prediction tasks in an agentic system all rely on the same underlying state that accumulates over a conversation: what has been introduced and remains salient, what the participants are trying to accomplish, what restricts them, and how these relate. Systems for these tasks are typically trained and evaluated one task at a time, even though multi-task and frozen-backbone architectures are common more broadly. We show that a single lightweight conversational state can be induced once, frozen, and reused across independent safety tasks *and* across the distribution shifts each task encounters, without retraining the representation. A shared mention encoder (M1) and a graph inducer (M2) together extract open-vocabulary entity, action, goal, and constraint nodes with persistent identity and typed relations; this M1+M2 inducer is trained once (419M parameters), frozen, and read by three small (sub-million-parameter) readers at the granularity of a whole trajectory, a single proposed action, and a single turn boundary. Our central evidence is controlled transfer under distribution shift rather than a single benchmark score. For multi-turn jailbreak detection we show that the standard generic-benign split admits a large topical shortcut (a first-turn topic probe alone reaches $0.87$ AUROC), and then hold the reader to topic-matched benign twins, naturally occurring sensitive conversations, over-refusal cases, and three unseen attack families; on the twins the frozen structured state reaches $0.84$ AUROC against $0.71$ for a capacity-matched dense reader over the *same* jointly trained encoder. For agent action authorization, a single operating point selected only on InjecAgent transfers to a live same-harness AgentDojo evaluation at $6.1\%$ targeted attack success and $74.2\%$ benign utility (and $1.9\%$ / $70.9\%$ at a high-security threshold), and remains effective across suites, attacks, and a change of agent model. The representation is frozen throughout: across held-out hard negatives, attack families, live agent environments, and agent-model changes, the same state is reused without representation retraining. Controls rule out shortcut and encoder-only explanations, and matched occlusions and retrains show each task draws on the semantic and relational channels of the state to a different degree.

## 1 INTRODUCTION

Multi-turn conversations require models to reason over state that accumulates across turns. A user may introduce an entity early, refine a goal later, impose a constraint several turns afterward, and eventually request an action whose meaning depends on all three. Treating each turn independently discards exactly the information that later decisions depend on.

This dependence is not specific to one application. Crescendo jailbreaks succeed through escalation across individually plausible turns (Russinovich et al., 2025); tool-using agents must judge whether a proposed action still serves the user's goals and constraints (Zhan et al., 2024; Jia et al., 2025); contextual-privacy decisions turn on whether an entity disclosed under one restriction is reused under another (Mireshghallah et al., 2024); sabotage monitors must notice an objective the user never authorized (Kutasov et al., 2025). Appendix A surveys these and other decisions. They differ in prediction unit – an entire trajectory, one proposed action, the current turn – and in label semantics, but they query the same conversational state: entities, goals, constraints, and actions introduced at



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

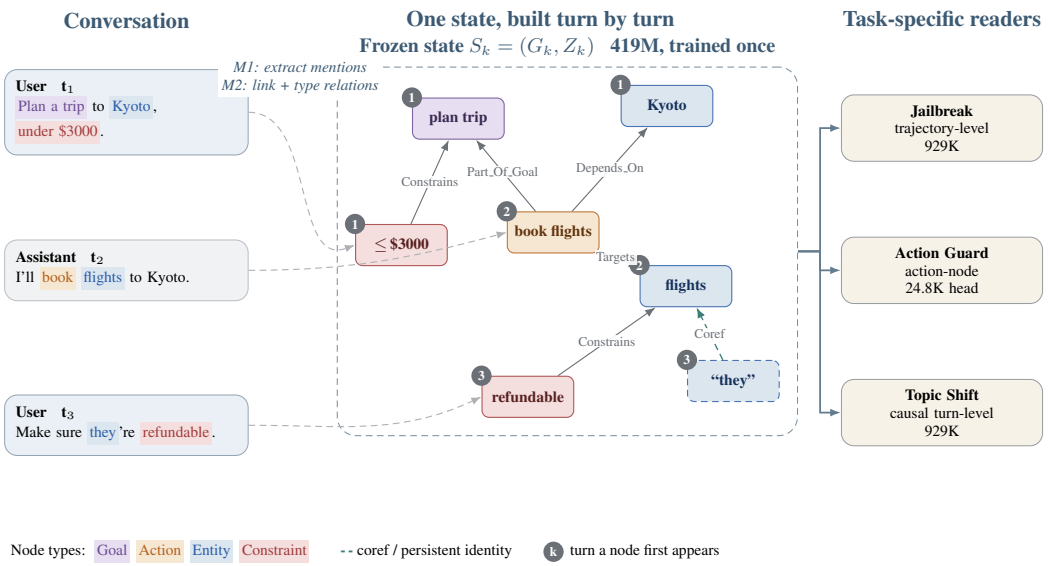



Figure 1: **One conversational state, built turn by turn, then frozen and reused.** As each turn arrives (left), a shared encoder (M1) extracts open-vocabulary ENTITY, ACTION, GOAL, and CONSTRAINT mentions, and a graph inducer (M2) links them to persistent nodes and predicts typed relations, so the semantic–relational state $S_k = (G_k, Z_k)$ grows causally – badges mark the turn each node first appears. Later turns reuse earlier nodes: the $t_2$ action attaches to the $t_1$ goal and destination, and the $t_3$ pronoun "they" corefers to the persistent FLIGHTS entity introduced at $t_2$. The finished state is frozen (419M parameters, trained once) and read by three small task-specific readers operating at trajectory, action-node, and turn granularity.

different times, and the relations among them. Yet each is typically served by a bespoke pipeline that re-learns conversational state as a side effect of learning its own decision.

Classical discourse theory models a conversation as an attentional state of salient objects, an intentional structure of purposes, and the relations that bind them, revised turn by turn (Grosz & Sidner, 1986; Cohen & Perrault, 1979; Traum & Larsson, 2003), and dialogue-state tracking and entity–constraint dialogue management build on exactly these primitives (Williams et al., 2016; Xu & Seneff, 2010). Recent learned systems recover pieces of the picture – explicit conversational structure through dataflow and semantic graphs (Andreas et al., 2020; Bai et al., 2021; Hu et al., 2022), dense representations transferable across dialogue tasks (Zhou et al., 2022; Zhao et al., 2022) – but usually one facet at a time and specialized to one task (§6). What remains open is a transfer question in the sense of this area: can one conversational representation be trained once and then remain useful when the downstream label semantics, corpus, prediction granularity, negative distribution, attack family, and agent environment all change, *without* retraining the representation?

We investigate this with the semantic–relational conversational state of Figure 1. A shared encoder extracts open-vocabulary mentions of entities, actions, goals, and constraints and embeds them in context; a graph inducer links mentions to persistent nodes and predicts typed relations between them, so the state evolves causally turn by turn. The representation therefore has two layers: a *semantic abstraction* that turns utterances into persistent conversational objects, and a *relational graph* that tracks their identity and typed dependencies. Once trained, the 419M-parameter inducer is frozen. Each downstream task receives the same induced state and trains only a lightweight reader: 929K parameters for trajectory-level jailbreak detection, a 24.8K-parameter head for action-level authorization, and 929K for causal turn-level topic-shift prediction. The three were chosen as widely separated points in Appendix A, one per prediction unit.

**Contributions.**

Table 1: **Positioning against published systems on each held-out benchmark (contextual, not controlled).** Axes: AUROC from chance (0.5) to 1; attack success rate (ASR) from 0 to 20%, lower is better; boundary F1 from 0 to 0.4. **Turquoise** bars are ours; grey bars are held out from the evaluated corpus like ours; dashed beige bars are trained on the target corpus itself and are not like-for-like. *Doubly held out*: neither the upstream M1+M2 inducer nor the task reader has seen the evaluated dataset; our topic-shift reader never sees TIAGE.

| TASK | PUBLISHED COMPARATORS VS. OURS | OURS |
|---|---|---|
| **Multi-turn jailbreak**<br>ActorAttack vs. MT-Bench-101<br>AUROC ↑ | **0.894 ours**<br>0.720 guardrails (Han et al., 2024)<br>0.915 SafeDream (Yan et al., 2026) | **0.894**<br>AUROC<br>doubly held out |
| **Agent action authorization**<br>AgentDojo, held out<br>ASR ↓ | 7.5% PromptGuard 2 (Chennabasappa et al., 2025)<br>**6.1% ours**<br>2.9% AlignmentCheck (Chennabasappa et al., 2025)<br>2.1% Task Shield (Jia et al., 2025)<br><1% PromptArmor (Shi et al., 2025) | **6.1%**<br>ASR<br>74.2% utility<br>doubly held out |
| **Topic-shift detection**<br>TIAGE, causal regime<br>BOUNDARY F1 ↑ | **0.318 ours**, zero-shot<br>0.278 Lin et al. (2023)<br>0.220 TSManager (Xie et al., 2021) | **0.318**<br>boundary F1<br>causal |

ours | held-out baseline | trained on target corpus

Table 2: **Same-backbone natural baseline on the identical held-out sets.** An end-to-end supervised fine-tune of the same stock ModernBERT-large backbone (Warner et al., 2024) on each task's training data, scored on the same held-out set as our frozen state, with no induced graph. FT is the mean over three seeds at the best learning rate (Appendix K.2). All three Δ values are improvements; the arrow gives each metric's better direction (AUROC and boundary F1 higher, ASR lower). The **Ours** ± values are measured three-seed/three-run spreads.

| TASK (HELD-OUT SET) | METRIC | MODERNBERT-L FT | OURS (FROZEN STATE) | Δ VS FT |
|---|---|---|---|---|
| Multi-turn jailbreak | AUROC ↑ | 0.780 ± 0.01 | **0.894 ± 0.006** | +0.114 ↑ |
| Agent action authorization | ASR ↓ | 50.5% ± 1.4 | **6.1% ± 0.8** | −44.40 **pp** ↓ |
| Topic-shift detection | Boundary F1 ↑ | 0.190 ± 0.02 | **0.318 ± 0.010** | +0.128 ↑ |

- **A reusable frozen conversational state.** A causally induced representation of what a conversation is made of – entity, action, goal, and constraint nodes with persistent identity and typed relations – trained once and reused unchanged, without any representation fine-tuning, by independent readers at trajectory, action, and turn granularity. It is a learned instance of the classical attentional and intentional view of discourse.
- **Controlled transfer against same-backbone representations.** Across heterogeneous downstream tasks, the structured frozen state outperforms readers built on the *same* jointly trained frozen encoder but without the semantic–relational state. Our protocol separates transfer from what inflates it – two levels of exposure (reader versus representation), shortcut probes and position-matched evaluation, and a reliance-versus-value decomposition – and localizes the value to the semantic and relational channels to task-dependent extents.
- **Interpretability and structure analyses.** Matched occlusions and retrains identify which channel is load-bearing for each task, and show the induced structure is necessary for authorization while the persistent semantic abstraction carries most of the jailbreak signal.

## 2 A SEMANTIC–RELATIONAL STATE FOR CONVERSATION

We represent a conversation not as a sequence of independent turns, but as an incrementally evolving state that preserves both *what* has been discussed and *how* those pieces of information relate. For a

conversation prefix $C_{\leq k} = (x_1, \ldots, x_k)$, we define

$$S_k = (G_k, Z_k), \qquad G_k = (V_k, E_k), \tag{1}$$

where $G_k$ is a directed typed graph and $Z_k = \{\mathbf{z}_v : v \in V_k\}$ contains contextual semantic embeddings associated with its nodes. The state is causal and incremental:

$$S_k = F_\theta(S_{k-1}, x_k), \qquad S_0 = \varnothing, \tag{2}$$

so a prediction at turn $k$ is constructed only from information available through that turn. This is the information-state view of dialogue – a state revised turn by turn as the conversation proceeds (Traum & Larsson, 2003) – with the update $F_\theta$ learned rather than hand-authored. User and assistant turns are processed identically.

### 2.1 SEMANTIC NODES

A shared ModernBERT-large encoder (Warner et al., 2024) reads each new turn and identifies four types of open-vocabulary mentions:

$$\mathcal{T} = \{\text{ENTITY}, \text{ACTION}, \text{GOAL}, \text{CONSTRAINT}\}.$$

ENTITY captures mentions of people, objects, resources, and concepts; ACTION captures actions performed or proposed by the assistant; GOAL represents desired outcomes or intentions by the user; and CONSTRAINT captures restrictions, requirements, or policies. These four types are chosen because they recover the primitives that classical accounts of discourse organize a conversation around – the salient objects of the attentional state (ENTITY), the events and speech acts of plan-based dialogue (ACTION), the discourse purposes of the intentional structure (GOAL), and the conditions that limit how goals and actions are to be interpreted (CONSTRAINT). They provide a small semantic scaffold without imposing a fixed vocabulary of entities, intents, or actions.

For each extracted mention $m$, we mean-pool the encoder hidden states over its span, producing

$$\mathbf{z}_m = \frac{1}{|I_m|} \sum_{i \in I_m} \mathbf{h}_i \in \mathbb{R}^{1024}, \tag{3}$$

where $I_m$ denotes the tokens belonging to the mention. Importantly, these node features come from the same jointly trained encoder that induces the graph; there is no separate language-model pass for downstream semantic features.

### 2.2 PERSISTENT IDENTITY AND TYPED RELATIONS

A mention is not automatically added as a new node. Given the existing state $G_{k-1}$, the graph inducer chooses among

$$V_{k-1} \cup \{\text{NEW}, \text{UNRESOLVED}\}. \tag{4}$$

Selecting an existing node merges the mention into that node as a new reference; NEW materializes a new persistent node; and UNRESOLVED retains a mention when the current state does not support a reliable attachment. This turns surface-level mention extraction into persistent conversational identity: repeated references across speakers or turns can map to the same node rather than producing independent representations.

The inducer also predicts directed typed relations between nodes using biaffine pair scoring (Dozat & Manning, 2017). We use nine relation types:

$$\begin{aligned}\mathcal{R} = \{&\text{TARGETS}, \text{PART_OF_GOAL}, \text{CONSTRAINS}, \text{COREF}, \text{BEFORE},\\ &\text{ENABLES}, \text{PREVENTS}, \text{HAS_PROPERTY}, \text{DEPENDS_ON}\}.\end{aligned} \tag{5}$$

The resulting graph therefore preserves both semantic content and explicitly induced dependencies. For example, an ACTION may TARGETS an ENTITY, participate in a persistent GOAL, or DEPENDS_ON an action introduced several turns earlier; a CONSTRAINT may remain connected to the goal it restricts even after the conversation has moved far beyond the turn where it was stated.

### 2.3 THE REUSABLE INTERFACE

All the components in the M1+M2 inducer are trained jointly, including mention extraction, persistent attachment, and relation induction. After this training, the complete inducer $F_\theta$ is frozen. A downstream task $q$ learns only its own reader,

$$\hat{y}_k^{(q)} = g_{\phi_q}^{(q)}(S_{\le k}), \qquad \theta \text{ fixed.} \tag{6}$$

Reader parameters $\phi_q$ are *not* shared between tasks. What transfers is the conversation state itself: the same nodes, semantic embeddings, persistent identities, and typed relations are exposed to every downstream task.

## 3 EXPERIMENTAL SETUP

Our central evaluation question is whether the state $S_k$ can support downstream decisions across a variety of safety prediction tasks. In all cases the upstream M1+M2 inducer is frozen and only the task-specific reader is trained.

**Multi-turn jailbreak: trajectory-level safety.** Crescendo attacks gradually transform an initially benign interaction into a harmful request over multiple turns (Russinovich et al., 2025). We train a trajectory reader on Crescendo conversations only (reader-training data documented in Appendix F.1) and evaluate cross-family transfer on ActorAttack (Ren et al., 2024), whose attack construction follows a different multi-turn strategy based on semantically related intermediate clues, together with three further unseen families: human-written multi-turn jailbreaks (MHJ, Li et al., 2024), coreference-based attacks (CoSafe, Yu et al., 2024), and concealed multi-turn attacks (RedQueen, Jiang et al., 2025). The reader is trained on Crescendo throughout, so this is positive-side attack-family transfer. A generic benign set, MT-Bench-101 (Bai et al., 2024), admits a topical shortcut and we report it only for reference: a topic-only probe (logistic regression on TF-IDF features of the first user turn alone) already separates ActorAttack from MT-Bench-101 at $0.87$ AUROC and collapses to $0.53$ against topic-matched negatives (§4). The primary evaluation therefore uses three harder negative distributions constructed to remove topical and lexical shortcuts (Appendix F.3): *benign twins* (the ActorAttack actor network rerun with a benign end goal, matched in topic and structure), *WildChat-sensitive* (naturally occurring sensitive conversations from WildChat (Zhao et al., 2024)), and *over-refusal* conversations built from XSTest (Röttger et al., 2024) and OR-Bench (Cui et al., 2024) prompts. We also compare against modern guard models (Llama Guard 3/4, WildGuard, ShieldGemma) given individual turns or whole conversations (Appendix F.4). Primary metrics are AUROC and TPR at 1% and 5% FPR, since low-false-positive operation matters for a safety detector.

**Action Guard: node-level authorization.** The second task asks whether a particular agent action is authorized by the user's conversation. This differs from generic harmfulness: an otherwise benign tool call can be unauthorized if it does not serve the user's stated objective or violates an earlier constraint. We use InjecAgent (Zhan et al., 2024) as the only training corpus for the reader and hold out AgentDojo (Debenedetti et al., 2024) for evaluation. Because cross-paper defense numbers are measured under different AgentDojo versions, task and injection sets, agent snapshots, and thresholds, they are contextual rather than controlled comparisons. Our primary action-authorization evaluation therefore runs the proposed guard *and* the competing defenses inside a single live AgentDojo harness, same tasks, same attacks, same agent snapshot, same version (Appendix G.11). The guard is integrated as an AgentDojo pipeline component that updates the induced state and authorizes each proposed action before it executes, and the guard threshold is fixed entirely on InjecAgent validation (target $3\%$ per-action false-block rate) before any AgentDojo evaluation. Predictions are made independently for each ACTION node. Primary metrics are benign utility, utility under attack, and the end-to-end targeted attack success rate (ASR), reported at the same operating point so that successful attacks and lost utility trade off explicitly. Separately, a large all-authorized AgentDojo-derived rollout (the $39{,}002$-action set) is used only as a false-block stress test at the per-action level, and does *not* supply the live task-level ASR or utility (Appendix F.5).

**Topic shift: causal turn-level discourse prediction.** The third task is on topicality, a multi-turn safety signal in its own right: gradual jailbreaks and injected instructions both surface as a goal that

| COMPONENT | METRIC | SCORE |
|---|---|---|
| M1: extraction | Micro-avg precision / recall | 0.66 / 0.76 |
| | Micro-avg F1 | 0.71 |
| M2: induction | Attachment accuracy | 0.928 |
| | NEW F1 | 0.965 |
| | Typed relation accuracy | 0.978 |

Table 3: **Intrinsic quality of the frozen M1+M2 inducer.** M1 span extraction is scored under strict exact-span matching over 13,331 held-out turns; graph induction (M2) is evaluated using *predicted* mentions over 13,219 held-out turns, with typed relation accuracy computed on gold-edge pairs. Per-type M1 metrics are in Appendix D (Table 8).

no longer attaches to what came before (Russinovich et al., 2025; Ren et al., 2024; Zhan et al., 2024). We train on SuperDialseg (Jiang et al., 2023), select on the source-disjoint DialSeg711 validation corpus, and evaluate on TIAGE (Xie et al., 2021). The prediction at turn $k$ receives only $S_k$: neither future turns nor the response following the candidate boundary is available. We report boundary AUROC and F1 together with WindowDiff.

**One frozen state, three readers.** A single M1+M2 checkpoint induces the graphs for all three tasks. It is trained from ModernBERT-large on Chatbot Arena together with Claude-generated synthetic conversations, in which Claude converses against smaller assistant models, using mention and graph labels distilled from a frontier LLM. Both sources are general-purpose dialogue: no attack corpus, agent-security benchmark, or topic-segmentation corpus enters this pool, so the checkpoint is exposure-clean for all three held-out evaluation corpora.

**Two levels of exposure.** The protocol distinguishes *reader exposure* (a corpus supplied downstream labels) from *representation exposure* (a corpus was present when the shared inducer was trained). Holding a corpus out only from the reader establishes downstream transfer over a fixed representation. Holding it out from both establishes the stronger claim that the state itself did not require exposure. Most published multi-turn safety results control only the first; our tasks hold both clean. Appendix F gives the full exposure matrix.

**Reliance versus value, and how each is measured.** Two experimental regimes recur throughout. An *occlusion* disables a channel at inference on an already-trained reader and measures how much that reader *relies* on it. A *matched retrain* trains a new reader with the channel absent from the first step, against a control identical in trainer, configuration, and induced features, and measures what the channel is *worth*. The two answer different questions and, as Table 19 shows, they can differ in magnitude and across tasks.

**A common transfer protocol.** The three tasks define three forms of distribution shift, attack-family, agent/tool-corpus, and dialogue-domain, and three experimental units: an entire trajectory, an individual action node, and a causal turn boundary. Success therefore cannot be explained by reusing a common label space or prediction head. What is held fixed is a single upstream semantic–relational state, one frozen checkpoint shared verbatim by all three readers.

## 4 TRANSFER EVALUATION

After the semantic–relational state is induced and frozen, we ask whether independently trained lightweight readers recover useful signal on substantially different tasks without retraining the representation. We evaluate three readers, each under a distribution shift it never sees, attack family (jailbreak), agent corpus (action authorization), and dialogue domain (topic shift), while the representation itself is held out from all three: the M1+M2 inducer is trained only on Chatbot Arena together with Claude-generated synthetic dialogue, so none of ActorAttack, MT-Bench-101, AgentDojo, or TIAGE was seen when the state was induced. Table 4 summarizes transfer alongside the controls that qualify it, and Table 5 isolates the central head-to-head that runs through all three readers: adding the induced structure to the *same* frozen encoder as a capacity-matched structure-

Table 4: **Transfer from a common frozen conversational state.** The headline condition in each block is marked ⋆ and set in **turquoise** bold, so it stays legible in grayscale; subsequent rows either reintroduce a source of distribution shift or remove a shortcut, and are indented under it. AUROC provides a common ranking statistic but does not make the absolute difficulty of the three tasks comparable. Turquoise headline figures carry ± the standard deviation over three retraining seeds (0, 1, 2; the reported value is the seed-0 headline run); the small spreads confirm the headlines are seed-stable rather than single-run artifacts.

| CONDITION | DESIGN NOTE | AUROC | OTHER |
|---|---|---|---|
| CRESCENDO → ACTORATTACK (GENERIC BENIGN: MT-BENCH-101; HARD NEGATIVES IN TABLE 17) | | | |
| **⋆ Clean representation, clean reader** | **both stages clean; generic split, topic-confounded (topic probe 0.87), see Table 17 for hard negatives** | **0.894 ± 0.006** | **TPR@1% 0.261** |
| ACTION GUARD: INJECAGENT → AGENTDOJO | | | |
| Full state reader | cross-corpus only | 0.951 | FBR 0.035 |
| **⋆ Position-matched state reader** | **within position strata** | **0.825** | **position-only → 0.50** |
| TOPIC SHIFT: SUPERDIALSEG → {DIALSEG711, TIAGE} | | | |
| DialSeg711 (validation) | source-disjoint | 0.770 ± 0.009 | – |
| **⋆ TIAGE (held-out)** | **+ naturalistic-dialogue shift** | **0.612 ± 0.007** | **boundary-F1 0.318 ± 0.010** |

Table 5: **The central controlled contrast behind the transfer claim.** For every reader, adding the induced semantic–relational structure to the *same* jointly trained frozen encoder is compared against a capacity-matched structure-free dense reader over that encoder, under the same held-out distribution shift (Appendix K.3). Live same-harness AgentDojo operating points and the full baseline set are in Appendix G (Tables 18, 17).

| READER (HELD-OUT SHIFT) | METRIC | FROZEN DENSE STATE | OURS (FULL STATE) |
|---|---|---|---|
| Jailbreak (benign twins) | AUROC | 0.71 ± 0.02 | **0.84 ± 0.02** |
| Action guard (InjecAgent → AgentDojo) | AUROC | 0.764 ± 0.01 | **0.951 ± 0.005** |
| Topic shift (TIAGE, held-out) | AUROC | 0.582 ± 0.01 | **0.612 ± 0.007** |
| Topic shift (TIAGE, held-out) | boundary F1 | 0.276 ± 0.02 | **0.318 ± 0.010** |

free dense reader. On topic-matched benign twins the structured state holds 0.84 AUROC where a first-turn topic probe collapses to chance (0.53) and the dense reader reaches only 0.71; on unseen AgentDojo the action guard reaches 0.951 AUROC against the dense reader's 0.764; and on naturalistic TIAGE it retains a smaller 0.582 → 0.612 AUROC (F1 0.276 → 0.318) edge. The live operating points transfer too: in a single same-harness AgentDojo run, same tasks, attacks, agent snapshot, and version, the guard reaches 6.1% targeted ASR at 74.2% benign utility with its threshold fixed only on InjecAgent validation, tightening to 1.9%/70.9% at a high-security threshold. We do not claim the lowest ASR (PromptArmor and MELON are lower at the ordinary threshold) or any form of reasoning; the claim is transfer without representation retraining. Full protocols, baselines, unseen-family and early-detection results, calibration transfer, cross-agent transfer, and the structure/no-structure decomposition are in Appendix G (Tables 17, 18, and 19).

## 5 DEPLOYMENT COST

The three readers share one forward pass over the frozen state, so a deployment obtains all three safety decisions (multi-turn jailbreak screening, per-action authorization, and topic-shift tracking) for the price of a single small encoder. We cost this against the two obvious alternatives that produce the same three answers: three task-specific 419M encoders (one per task, no sharing), and three 7B LLM-judge calls (one per task). We take a conversation of 6 turns totaling 2,000 tokens, and run the guard once per turn in the blocking path between the user turn and the assistant response. Because the state is incremental (Section 2), our guard encodes only the new turn at each check rather than re-reading the conversation, so over the six turns it processes the 2,000-token conversation once.

Table 6: Cost of obtaining all three safety decisions over a 6-turn, 2,000-token conversation. Our incremental readers share a single 419M encoder pass and encode each turn once; the separate-encoder baseline repeats that work per task; the 7B judges are issued in parallel with prefix caching. Latency is the delay added to each assistant response in the blocking path; the latency and compute figures assume batch-1 inference and are analytical estimates from model sizes and token counts, so absolute values depend on the hardware and serving stack.

| APPROACH | LATENCY/TURN | LATENCY/CONV. | TOKENS/CONV. | COMPUTE/CONV. (TFLOP) |
|---|---|---|---|---|
| **Shared state + 3 readers (ours)** | **4 ms** | **0.03 s** | **2,000** | **1.7** |
| 3 task-specific 419M encoders | 12 ms | 0.07 s | 6,000 | 5.0 |
| 3 × 7B LLM-judge calls (parallel) | 800 ms | 4.8 s | 11,940 | 165 |

The reader heads are sub-million-parameter and add 0.8 ms per check. The LLM judges are 7B generative models: each prefills the conversation once with prefix caching, takes a 300-token task instruction per turn, and decodes a 30-token verdict, with the three calls issued in parallel.

Table 6 makes the trade concrete. Sharing the encoder cuts compute and tokens $3\times$ relative to three separate small models, because the state is built once and read three times rather than rebuilt per task; parallel hardware can hide the latency of the extra encoders, but their compute and token cost remains. Against LLM judges the gap is roughly $100\times$ in compute: our reader runs a single non-autoregressive pass over each new turn, whereas each 7B judge must prefill and then decode a verdict token by token. The latency difference is what matters in deployment. Our incremental guard adds 0.03 s across an entire conversation, below the threshold a user perceives; even with the three judge calls issued in parallel and the context prefix-cached, each judge still decodes autoregressively, adding 4.8 s over the conversation, an interaction-breaking delay on top of the assistant's own generation. One small model delivers the three decisions at a cost that keeps it on the critical path; the alternatives do not.

## 6 RELATED WORK

**Conversational state and its representations.** The object we induce has some precedent. Grosz and Sidner decompose discourse into an intentional structure of purposes and an attentional state of salient objects (Grosz & Sidner, 1986); centering theory formalizes how ranked discourse entities persist across utterances (Grosz et al., 1995); plan-based accounts treat speech acts as operators serving goals (Cohen & Perrault, 1979); and the information-state approach casts dialogue as a state revised turn by turn by typed update rules (Traum & Larsson, 2003). Dialogue-state tracking and entity–constraint dialogue management build on the same primitives (Williams et al., 2016; Budzianowski et al., 2018; Xu & Seneff, 2010), as do discourse representation theory and the meaning-representation frameworks that join entities and events by typed relations (Kamp & Reyle, 1993; Baker et al., 1998; Banarescu et al., 2013; Bonial et al., 2020). Learned systems recover pieces of this picture: dataflow graphs for task-oriented state (Andreas et al., 2020), dialogue-level AMR and DMR (Bai et al., 2021; Hu et al., 2022), graphs as task-specific context encoders (Ghosal et al., 2019; Jain & Lapata, 2023), and dense dialogue representations that transfer across tasks (Zhou et al., 2022; Zhao et al., 2022). We claim neither the primitives nor the idea of transfer as new. What is new is the object transferred – open-vocabulary semantic nodes with persistent identity and typed dependencies, induced jointly and causally by one learned model and then held entirely fixed – and the question asked of it: which channel is load-bearing for each task, and how much of the apparent transfer survives exposure and shortcut controls.

**Multi-turn safety, agent security, and topic segmentation.** Crescendo and ActorAttack succeed by spreading intent across individually plausible turns (Russinovich et al., 2025; Ren et al., 2024), and recent defenses already model history: G-Guard applies a GNN over entity graphs (Huang et al., 2025), and DeepContext propagates recurrent state over fine-tuned turn embeddings (Albrethsen et al., 2026). Ours is therefore neither the first graph-based nor the first stateful multi-turn detector; our graph-free configuration is architecturally close to such a stateful recurrent detector. Off-the-shelf guard models score turns or transcripts for harmfulness (Han et al., 2024); we compare against several (Llama Guard 3/4, WildGuard, ShieldGemma) and find that on topic-matched benign twins they retain far less separation than on generic benign sets, motivating our hard-negative protocol

(§G.2). The published detection results on the ActorAttack data (Table 1) differ in regime and, as we show, on a topic-confounded split: the guardrail judges (Han et al., 2024) are used off the shelf and held out from the attack corpus as ours is (ours additionally from the representation), while SafeDream (Yan et al., 2026) trains on the target corpus's own folds, so its within-corpus figure is not a like-for-like comparison. On the agent side, InjecAgent and AgentDojo show that untrusted tool content can redirect an agent (Zhan et al., 2024; Debenedetti et al., 2024), and Task Shield asks whether a call serves the user's objective rather than whether it is harmful (Jia et al., 2025); PromptArmor (Shi et al., 2025) and MELON (Zhu et al., 2025) are recent prompt-injection defenses. Because published numbers for these systems are measured under different AgentDojo versions, attack sets, agent snapshots, and thresholds, they are contextual rather than controlled comparisons; we therefore re-run the relevant defenses in a single live harness (§G.3). The systems that report per-action authorization classification, GuardAgent and AGrail (Xiang et al., 2025; Luo et al., 2025), use disjoint access-control corpora, so no per-action AUROC exists on the InjecAgent→AgentDojo pair. In discourse, the predictive/retrospective split is native to TIAGE, whose retrospective classifier approaches human agreement while its predictive one lags (Xie et al., 2021), and LLM segmenters read completed dialogues with bidirectional context (Lee et al., 2025); our causal reader sits on the harder side of that split and is compared only within it.

## 7 Limitations and Conclusion

**Limitations.** Two limitations qualify the claims above. *The original jailbreak split rewards topic recognition.* Generic benign controls substantially overestimate trajectory sensitivity: topic alone separates ActorAttack from MT-Bench-101 at $0.87$ AUROC, collapsing to $0.53$ against topic-matched twins (§G.2). This is now an empirical finding, and it is why we report hard-negative results as the primary jailbreak evidence rather than the generic split. *Goal-aligned adaptive attacks remain a failure mode.* An injection phrased as an instrumental step toward the user's legitimate goal raises AgentDojo ASR from $6.1\%$ to $11.8\%$ (§G.3); this is structurally relevant, because the guard's premise is that unauthorized actions detach from the user's goals and constraints, and such attacks are designed to re-attach them. PromptArmor is also substantially affected under the same attack.

**Conclusion.** We asked whether one causally induced semantic–relational state can be frozen and reused across heterogeneous multi-turn prediction problems and the distribution shifts each brings. The answer is yes, and it holds under stricter controls than an uncontrolled evaluation would apply. A single frozen representation transfers not merely to different classification heads, but across downstream objectives, corpus distributions, hard-negative distributions, attack families, conversation prefixes, action environments, attack strategies, decision thresholds, and agent-model families, all without retraining the representation. On multi-turn jailbreak detection, after showing that the standard generic-benign split is topic-confounded, the frozen state retains a large advantage over a same-encoder dense reader on topic-matched twins ($0.84$ vs. $0.71$), naturally occurring sensitive conversations, over-refusal cases, and three unseen attack families including human-written MHJ ($0.82$). On action authorization, an operating point selected only on InjecAgent transfers to a live same-harness AgentDojo run at $6.1\%$ targeted ASR / $74.2\%$ benign utility ($1.9\%$ / $70.9\%$ at a high-security threshold), with its false-block rate tracking the InjecAgent target and its effectiveness holding across suites, attacks, and agent models; the offline classification transfer reaches $0.951$ AUROC, and a no-edge retrain drops sharply to $0.528$, so a large relational residual remains. Discourse transfer is lower under a shift from document-derived to naturalistic dialogue, a loss of discrimination that survives recalibration and lookahead. These results come with real qualifications: the ontology and teacher supervision remain important dependencies, goal-aligned adaptive attacks remain a failure mode ($11.8\%$ ASR), and wider downstream task coverage remains future work. We claim reuse across heterogeneous decisions and multiple distribution shifts, not a universal or task-independent conversational representation.

The value of a reusable conversational representation therefore lies not in forcing every task through graph structure, but in exposing multiple persistent channels, semantic and relational, and in the controls that reveal which of them is actually doing the work for each task.

## REFERENCES


Justin Albrethsen, Yash Datta, Kunal Kumar, and Sharath Rajasekar. DeepContext: Stateful real-time detection of multi-turn adversarial intent drift in LLMs. *arXiv preprint arXiv:2602.16935*, 2026.

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8: 556–571, 2020. doi: 10.1162/tacl_a_00333.

Benjamin Arnav, Pablo Bernabeu-Pérez, Nathan Helm-Burger, Tim Kostolansky, Hannes Whittingham, and Mary Phuong. CoT red-handed: Stress testing chain-of-thought monitoring. In *Advances in Neural Information Processing Systems*, 2025.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.401.

Xuefeng Bai, Yulong Chen, Linfeng Song, and Yue Zhang. Semantic representation for dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4430–4445. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.342.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL)*, pp. 86–90. Association for Computational Linguistics, 1998. doi: 10.3115/980845.980860.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 178–186. Association for Computational Linguistics, 2013. URL https://aclanthology.org/W13-2322/.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare R. Voss. Dialogue-AMR: Abstract meaning representation for dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pp. 684–695. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.86/.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ – a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, 2018. doi: 10.18653/v1/D18-1547.

Sahana Chennabasappa, Cyrus Nikolaidis, Daniel Song, David Molnar, Stephanie Ding, Shengye Wan, Spencer Whitman, Lauren Deason, Nicholas Doucette, Abraham Montilla, Alekhya Gampa, Beto de Paola, Dominik Gabi, James Crnkovich, Jean-Christophe Testud, Kat He, Rashnil Chaturvedi, Wu Zhou, and Joshua Saxe. LlamaFirewall: An open source guardrail system for building secure AI agents. *arXiv preprint arXiv:2505.03574*, 2025. URL https://arxiv.org/abs/2505.03574.

Philip R. Cohen and C. Raymond Perrault. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3(3):177–212, 1979. doi: 10.1207/s15516709cog0303_1.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024. URL https://arxiv.org/abs/2405.20947. Accepted to ICML 2025.

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=m1YYAQjO3w.

Timothy Dozat and Christopher D. Manning. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*, 2017.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490, 2022. doi: 10.1162/tacl_a_00529.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 154–164. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1015.

Aaron Grattafiori et al. The llama 3 herd of models. 2024. URL https://arxiv.org/abs/2407.21783. Cited per Meta's official model card as the reference for Llama Guard 3; see also https://huggingface.co/meta-llama/Llama-Guard-3-8B.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pp. 79–90, 2023. doi: 10.1145/3605764.3623985.

Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986. URL https://aclanthology.org/J86-3001/.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995. URL https://aclanthology.org/J95-2003/.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. *arXiv preprint arXiv:2406.18495*, 2024. URL https://arxiv.org/abs/2406.18495.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.

Xiangkun Hu, Junqi Dai, Hang Yan, Yi Zhang, Qipeng Guo, Xipeng Qiu, and Zheng Zhang. Dialogue meaning representation for task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 223–237. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.17.

Zixuan Huang, Kecheng Huang, Lihao Yin, Bowei He, Huiling Zhen, Mingxuan Yuan, and Zili Shao. Attention-aware GNN-based input defense against multi-turn LLM jailbreak. *arXiv preprint arXiv:2507.07146*, 2025.

Parag Jain and Mirella Lapata. Conversational semantic parsing using dynamic context graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8667–8679. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.535.

Feiran Jia, Tong Wu, Xin Qin, and Anna Squicciarini. The task shield: Enforcing task alignment to defend against indirect prompt injection in LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29680–29697. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1435.

Junfeng Jiang, Chengzhang Dong, Sadao Kurohashi, and Akiko Aizawa. SuperDialseg: A large-scale dataset for supervised dialogue segmentation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4086–4101. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.249.

Yifan Jiang, Kriti Aggarwal, Tanmay Laud, Kashif Munir, Jay Pujara, and Subhabrata Mukherjee. RED Queen: Safeguarding large language models against concealed multi-turn jailbreaking. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. URL https://arxiv.org/abs/2409.17458.

Hans Kamp and Uwe Reyle. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, volume 42 of *Studies in Linguistics and Philosophy*. Kluwer Academic Publishers, Dordrecht, 1993.

Jonathan Kutasov, Yuqi Sun, Paul Colognese, Teun van der Weij, Linda Petrini, Chen Bo Calvin Zhang, John Hughes, Xiang Deng, Henry Sleight, Tyler Tracy, Buck Shlegeris, and Joe Benton. SHADE-arena: Evaluating sabotage and monitoring in LLM agents. *arXiv preprint arXiv:2506.15740*, 2025.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

Seungmin Lee, Yongsang Yoo, Minhwa Jung, and Min Song. Def-DTS: Deductive reasoning for open-domain dialogue topic segmentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20736–20753. Association for Computational Linguistics, 2025.

Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. Llm defenses are not robust to multi-turn human jailbreaks yet, 2024. URL https://arxiv.org/abs/2408.15221.

Jiangyi Lin, Yaxin Fan, Feng Jiang, Xiaomin Chu, and Peifeng Li. Topic shift detection in chinese dialogues: Corpus and benchmark. In *Document Analysis and Recognition – ICDAR 2023*, volume 14189 of *Lecture Notes in Computer Science*, pp. 166–183. Springer, 2023. doi: 10.1007/978-3-031-41682-8_11.

Weidi Luo, Shenghong Dai, Xiaogeng Liu, Suman Banerjee, Huan Sun, Muhao Chen, and Chaowei Xiao. AGrail: A lifelong agent guardrail with effective and adaptive safety detection. *arXiv preprint arXiv:2502.11448*, 2025.

Meta AI. Llama guard 4. Model card, 2025. URL https://huggingface.co/meta-llama/Llama-Guard-4-12B. Released April 2025 alongside Llama 4; no dedicated paper identified as of verification date.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can LLMs keep a secret? testing privacy implications of language models via contextual integrity theory. In *International Conference on Learning Representations*, 2024.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. LingMess: Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2752–2760, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.202. URL https://aclanthology.org/2023.eacl-main.202/.

Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. Derail yourself: Multi-turn LLM jailbreak attack through self-discovered clues. *arXiv preprint arXiv:2410.10700*, 2024.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo Multi-Turn LLM jailbreak attack. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 2421–2440, Seattle, WA, 2025. USENIX Association. ISBN 978-1-939133-52-6.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web*, pp. 593–607. Springer International Publishing, 2018. doi: 10.1007/978-3-319-93417-4_38.

Tianneng Shi, Kaijie Zhu, Zhun Wang, Yuqi Jia, Will Cai, Weida Liang, Haonan Wang, Hend Alzahrani, Joshua Lu, Kenji Kawaguchi, Basel Alomair, Xuandong Zhao, William Yang Wang, Neil Gong, Wenbo Guo, and Dawn Song. PromptArmor: Simple yet effective prompt injection defenses. *arXiv preprint arXiv:2507.15219*, 2025. URL https://arxiv.org/abs/2507.15219.

David R. Traum and Staffan Larsson. The information state approach to dialogue management. In Jan van Kuppevelt and Ronnie W. Smith (eds.), *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology*, pp. 325–353. Kluwer Academic Publishers, Dordrecht, 2003. doi: 10.1007/978-94-010-0019-2_15.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024.

Jason D. Williams, Antoine Raux, and Matthew Henderson. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33, 2016. doi: 10.5087/dad.2016.301.

Seungpil Won, Heeyoung Kwak, Joongbo Shin, Janghoon Han, and Kyomin Jung. BREAK: Breaking the dialogue state tracking barrier with beam search and re-ranking. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2832–2846. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.159.

Patrick Xia, João Sedoc, and Benjamin Van Durme. Incremental neural coreference resolution in constant memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8617–8624, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.695. URL https://aclanthology.org/2020.emnlp-main.695/.

Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Carl Yang, Dawn Song, and Bo Li. GuardAgent: Safeguard LLM agents by a guard agent via knowledge-enabled reasoning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. TIAGE: A benchmark for topic-shift aware dialog modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1684–1690. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.145.

Liyan Xu and Jinho Choi. Modeling task interactions in document-level joint entity and relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5409–5416, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.395. URL https://aclanthology.org/2022.naacl-main.395/.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. Topic-aware multi-turn dialogue modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14176–14184, 2021. doi: 10.1609/aaai.v35i16.17668.

Yushi Xu and Stephanie Seneff. Dialogue management based on entities and constraints. In *Proceedings of the SIGDIAL 2010 Conference*, pp. 87–90. Association for Computational Linguistics, 2010. URL https://aclanthology.org/W10-4317/.

Bo Yan, Weikai Lin, Yada Zhu, and Song Wang. SafeDream: Safety world model for proactive early jailbreak detection. *arXiv preprint arXiv:2604.16824*, 2026. URL https://arxiv.org/abs/2604.16824.

Fanghua Ye, Xi Wang, Jie Huang, Shenghui Li, Samuel Stern, and Emine Yilmaz. MetaASSIST: Robust dialogue state tracking with meta learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.76.

Erxin Yu, Jing Li, Ming Liao, Siqi Wang, Zuchen Gao, Fei Mi, and Lanqing Hong. CoSafe: Evaluating large language model safety in multi-turn dialogue coreference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17494–17508, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.968. URL https://aclanthology.org/2024.emnlp-main.968/.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. Shieldgemma: Generative ai content moderation based on gemma, 2024. URL https://arxiv.org/abs/2407.21772.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10471–10506. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.624.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2405.01470.

Xueliang Zhao, Lemao Liu, Tingchen Fu, Shuming Shi, Dongyan Zhao, and Rui Yan. Towards efficient dialogue pre-training with transferable and interpretable latent structure. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10051–10063. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.683.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14612–14620, 2021. doi: 10.1609/aaai.v35i16.17717.

Zhihan Zhou, Dejiao Zhang, Wei Xiao, Nicholas Dingwall, Xiaofei Ma, Andrew Arnold, and Bing Xiang. Learning dialogue representations from consecutive utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 754–768. Association for Computational Linguistics, 2022.

Kaijie Zhu, Xianjun Yang, Jindong Wang, Wenbo Guo, and William Yang Wang. MELON: Provable defense against indirect prompt injection attacks in AI agents. In *International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.05174.

# A Downstream Decisions as Queries Against Conversational State

Table 7 surveys downstream problems that are queries against the same accumulated state. They differ in prediction unit and label semantics; what they share is the object being queried. The three tasks marked ⋆ are the ones evaluated in the main paper, chosen as one per prediction unit and all drawn from safety.

Table 7: Downstream decisions that are queries against persistent conversational state. Tasks marked ⋆ are evaluated in this paper. The reference-point column is *not* a ranking: each task uses its own benchmark and metric, so figures are not comparable across rows. Most reported figures are *retrospective*, computed over a completed transcript with bidirectional context, whereas a deployed reader must decide causally at turn $k$.

| TASK | UNIT | STATE EVIDENCE | REPORTED REFERENCE POINT | TYPE | REGIME |
|---|---|---|---|---|---|
| TRAJECTORY- AND SESSION-LEVEL | | | | | |
| Multi-turn jailbreak detection ⋆ | trajectory | escalation of goals against constraints stated in earlier turns | F1 0.84 stateful vs. 0.67 stateless guardrails, in-family (Albrethsen et al., 2026); entity-graph GNN defense (Huang et al., 2025); detection-AUROC leaderboard on the evaluated attack data (Yan et al., 2026) | hyb | causal; scored as the conversation unfolds |
| Agent sabotage / scheming monitoring | trajectory | unauthorized objective; claimed vs. actual outcome | ROC-AUC 0.87, Gemini 2.5 Pro monitor (Kutasov et al., 2025); action-only 0.755 vs. CoT+action 0.797 (Arnav et al., 2025) | LLM | retrospective; full transcript judged post-execution |
| Long-horizon goal adherence | session | goal lineage; constraints left unsatisfied at session end | 39% mean performance drop, single- vs. multi-turn (Laban et al., 2025) | LLM | diagnostic; measures degradation, not a detector |
| NODE-LEVEL | | | | | |
| Agent action authorization ⋆ | action node | links from a proposed action to stated goals and constraints | 2.07% ASR at 69.79% utility, GPT-4o on AgentDojo (Jia et al., 2025) | LLM | predictive; gates each call at test time |
| Injected-instruction screening | node / span | provenance of a node (ingested content vs. user turn) | defense-side ASR/utility frontier (Debenedetti et al., 2024); no detector leaderboard (Greshake et al., 2023) | hyb | predictive; runs at ingestion |
| Contextual privacy / PII exposure | entity node | scope of the constraint under which an entity was disclosed | GPT-4 discloses in 39% of contexts humans would not (Mireshghallah et al., 2024) | LLM | retrospective probe of model behaviour |
| TURN-LEVEL | | | | | |
| Topic-shift detection ⋆ | turn | whether the current goal attaches to any prior goal | retrospective classifier near human agreement vs. a clear gap to the predictive one (Xie et al., 2021); retrospective LLM segmenters do better (Lee et al., 2025) | hyb | split by regime; see Appendix I |
| Dialogue state tracking | turn | persistent entity identity and accumulated properties | JGA 80.10% on MultiWOZ 2.4 (Ye et al., 2022); ≈61% on 2.1/2.2 (Won et al., 2023) | hyb | predictive; state updated each turn |
| Response faithfulness | turn | entities asserted by the assistant vs. grounded evidence | critic +12.8 F1 on BEGIN over prior dialogue-coherence data (Dziri et al., 2022); degrades sharply off-domain | enc | retrospective |
| Emotion / frustration tracking | turn | speaker-linked node history across the conversation | MELD weighted-F1 0.581 (Ghosal et al., 2019), 0.665 for later text-only models (Hu et al., 2023) | enc | retrospective |

Solution type: LLM prompted or fine-tuned language model enc smaller encoder or graph model hyb both.

## B FULL REPRESENTATION SPECIFICATION

This appendix specifies the shared conversational state used throughout the paper. The representation is induced causally from an interleaved sequence of user and assistant turns

$$x_{1:T} = (x_1, \ldots, x_T),$$

with no access to future turns. After processing turn $t$, the representation is

$$S_t = (G_t, Z_t), \qquad G_t = (V_t, E_t),$$

where $G_t$ is a persistent typed directed graph and $Z_t = \{z_v\}_{v \in V_t}$ contains semantic node features derived from the same encoder that induces the graph. The state evolves incrementally,

$$S_t = F_\theta(S_{t-1}, x_t),$$

and $\theta$ is frozen before any downstream reader is trained.

### B.1 SEMANTIC MENTION ONTOLOGY

The mention extractor (M1) uses four open-vocabulary semantic types:

$$\mathcal{T} = \{\text{ENTITY}, \text{ACTION}, \text{GOAL}, \text{CONSTRAINT}\}.$$

Their intended semantics are:

| TYPE | DEFINITION |
|---|---|
| ENTITY | A concrete or discourse-salient object, actor, organization, resource, artifact, location, system, or referential expression. Pronouns and demonstratives may also be extracted when they refer to a discourse entity. |
| ACTION | An activity, operation, requested behavior, capability, or event, typically expressed by a verb phrase or nominalized action. |
| GOAL | An objective, desired state, purpose, or higher-level outcome that one or more actions may serve. Goals may contain or overlap action and entity spans. |
| CONSTRAINT | A condition, limitation, exclusion, policy, scope restriction, prerequisite, or framing condition that limits how an action or goal should be interpreted. |

These four types deliberately recover primitives from classical discourse and semantic theory: the salient objects of the attentional state and the discourse referents of coreference (ENTITY), the events and speech acts of plan-based dialogue (ACTION), the discourse purposes of the intentional structure (GOAL), and the conditions on plans and goals (CONSTRAINT) (Grosz & Sidner, 1986; Cohen & Perrault, 1979; Xu & Seneff, 2010; Banarescu et al., 2013). The ontology is kept small precisely so that a single induced state stays task-agnostic rather than specialized to any one downstream reader.

M1 consists of four independent BIO heads over a shared ModernBERT-large encoder (Warner et al., 2024). Because the heads are independent, spans of different semantic types may overlap or nest. For example, a GOAL span may contain an ACTION and its target ENTITY. Within each semantic type, the corresponding BIO sequence defines the span boundaries.

For a decoded mention $m$ covering token positions $i{:}j$, its semantic feature is the mean of the shared encoder states over the span,

$$z_m = \frac{1}{j-i+1} \sum_{r=i}^{j} h_r \in \mathbb{R}^{1024}.$$

No separate embedding model is used. The hidden states that support BIO decoding are the same hidden states from which M2 constructs its semantic features.

Each mention additionally records its conversation turn, speaker, character offsets, semantic type, and the structural metadata used by graph induction, including salience and unresolved-reference indicators where applicable. Both user and assistant turns pass through the identical extraction pipeline.

### B.2 PERSISTENT NODE IDENTITY

M2 converts turn-local mentions into persistent conversational objects. For every new mention $m$ at turn $t$, the attachment decision is

$$a(m) \in V_{t-1} \cup \{\text{NEW}, \text{UNRESOLVED}\}.$$

The three outcomes have distinct semantics.

**Link.** If $a(m) = v$ for an existing node $v$, the mention is treated as another reference or alias of the persistent object represented by $v$. The mention does not create a new semantic identity. Instead, its surface form and span representation are added to the node's alias history. This implements an incremental clustering formulation related to online coreference models (Xia et al., 2020).

**New.** If $a(m) = \text{NEW}$, the mention materializes a new persistent node whose type is inherited from M1. This node becomes available as an antecedent for all later turns.

**Unresolved.** If the system predicts UNRESOLVED, it records that the mention cannot be attached reliably to an existing conversational object at the current point. This prevents uncertain references from being forced into an incorrect cluster.

Attachment is therefore a decision over persistent conversational entities, not over individual previous mentions. A persistent node may accumulate many surface aliases across speakers and turns. Candidate nodes are represented using their alias histories so that a later mention may match any earlier surface realization of the same conversational object.

Type-compatibility constraints mask semantically illegal attachment candidates where defined; in particular, entity-coreference decisions do not attach an ENTITY mention to a non-ENTITY antecedent.

### B.3 ATTACHMENT SCORING AND CANDIDATE PRUNING

The attachment scorer combines a learned pair representation with explicit local signals. For mention $m$ and candidate persistent node $v$, the learned pair representation contains the current mention, the candidate node representation, their elementwise interaction, and their absolute difference. The scorer additionally receives ten scalar pair features capturing:

- signed turn distance and inverse recency;
- whether the candidate is the most recent compatible entity;
- maximum and mean semantic similarity between $m$ and the individual aliases of $v$;
- lexical/head-token agreement and token overlap;
- speaker and other local compatibility information.

The alias-similarity features are important because averaging a long alias history can dilute a single highly diagnostic prior mention. The lexical and recency features provide complementary signals for short-range reference resolution.

At inference time, attachment uses coarse-to-fine pruning when the graph contains many candidates. A non-parametric coarse score ranks candidate antecedents using alias similarity, recency, lexical matching, and type compatibility, after which only the top

$$K_{\text{attach}} = 24$$

candidates enter the final attachment softmax. Training retains the full legal candidate set so that pruning cannot remove the gold antecedent.

### B.4 TYPED RELATION ONTOLOGY

Persistent nodes are connected by directed typed relations from

$$\begin{aligned}\mathcal{R} = \{\, &\text{TARGETS}, \text{PART_OF_GOAL}, \text{CONSTRAINS}, \text{COREF},\\ &\text{BEFORE}, \text{ENABLES}, \text{PREVENTS},\\ &\text{HAS_PROPERTY}, \text{DEPENDS_ON}\,\}.\end{aligned}$$

The ontology is deliberately compact. It encodes semantic and relational structure while leaving node content open-vocabulary.

| RELATION | MEANING | ILLUSTRATIVE EXAMPLE |
|---|---|---|
| TARGETS | An action operates on, affects, requests, accesses, or otherwise targets an entity. | "download report" → "report" |
| PART_OF_GOAL | An action or subordinate objective contributes to a broader goal. | "collect records" → "prepare audit" |
| CONSTRAINS | A condition or restriction limits the interpretation or execution of another node. | "public data only" → "collect records" |
| COREF | Explicit semantic re-reference between graph objects when represented as a relation rather than absorbed solely by persistent-node attachment. | "the service" ↔ "it" |
| BEFORE | The source occurs temporally or procedurally before the destination. | "verify identity" → "transfer funds" |
| ENABLES | The source makes the destination possible or supplies a prerequisite. | "authenticate" → "access account" |
| PREVENTS | The source blocks, inhibits, or guards against the destination. | "rate limit" → "repeated requests" |
| HAS_PROPERTY | An attributed-property or descriptive association between graph objects. | "document" → "confidential" |
| DEPENDS_ON | The source semantically, procedurally, or causally depends on the destination. | "submit report" → "compile evidence" |

These relations are descriptive rather than safety-specific. Nothing in the ontology directly denotes "attack," "authorized," "topic boundary," or any other downstream label.

### B.5 RELATION SCORING

For candidate node pairs, M2 uses a full biaffine interaction (Dozat & Manning, 2017) rather than a dot product. The final graph head also contains type-conditioned expert scorers, motivated by the observation that different mention-pair categories benefit from different decision functions (Otmazgin et al., 2023). Multi-label relation decisions use an adaptive-threshold formulation following ATLOP (Zhou et al., 2021). An auxiliary graph-compatibility objective couples attachment/coreference and relation predictions so that structurally incompatible decisions are discouraged (Xu & Choi, 2022).

The final 1024-dimensional checkpoint uses the complete relation configuration: type experts, unresolved-reference modeling, adaptive thresholding, attachment-compatibility constraints, recency features, alias-similarity features, lexical features, coarse-to-fine pruning, and decoupled attachment and relation heads.

### B.6 DENSE RELATION DECODING

After attachment decisions update the persistent graph, relation scores are decoded in *dense* mode. Eligible pairs are considered even when no edge was created directly by the attachment operation. At deployment and in the cached graphs used by all downstream readers, relation decoding uses

$$\tau_{\text{rel}} = 0.5.$$

Predicted relations that do not clear the decoding threshold are omitted. This dense pass is important because attachment and semantic relation prediction answer different questions: two mentions can denote distinct persistent objects while still requiring one or more typed semantic edges.

### B.7 INCREMENTAL CONSTRUCTION PROCEDURE

Given the previous state $S_{t-1}$ and a new turn $x_t$, the representation is constructed as follows.

1. Encode $x_t$ once with the shared ModernBERT-large encoder.

2. Apply the four BIO heads and decode ENTITY, ACTION, GOAL, and CONSTRAINT mentions.
3. Mean-pool the encoder hidden states over every decoded mention to obtain its 1024-dimensional semantic vector.
4. For each mention, score the existing persistent nodes together with NEW and UNRESOLVED. At inference time, apply top-24 coarse-to-fine candidate pruning when necessary.
5. If the mention links to an existing node, update that node's alias history. If it selects NEW, create a persistent node. Preserve unresolved references explicitly.
6. Score typed semantic relations over eligible node pairs using the biaffine/type-expert relation module.
7. Densely decode relations with threshold $0.5$ and update $E_t$.
8. Serialize the resulting graph together with the mention-aligned 1024-dimensional semantic features and provenance metadata.

All steps are causal. Once turn $t$ has been processed, $S_t$ can be consumed immediately without re-encoding turns $1{:}t-1$. Figure 2 traces these steps over a short neutral dialogue and shows the persistent graph growing one turn at a time.

### B.8 SERIALIZED STATE AND PROVENANCE

For downstream use, each induced conversation is serialized as a self-contained record containing the conversation identifier, extracted mentions, typed directed edges, and semantic features. In simplified form:

```
{
  "conv_id": ...,
  "mentions": [
    {
      "id": ...,
      "type": "ACTION",
      "turn_idx": 3,
      "char_start": ...,
      "char_end": ...,
      "speaker": ...,
      "salience": ...,
      "unresolved": ...
    },
    ...
  ],
  "edges": [
    { "src": ..., "dst": ..., "rel": "DEPENDS_ON", "conf": ... },
    ...
  ],
  "node_span_emb": [[...1024 values...], ...],
  "_meta": {
    "span_dim": 1024,
    "span_source": "joint_live_1024d",
    "decode_mode": "dense",
    "rel_thresh": 0.5,
    "joint_ckpt": ".../best_by_coref_acc.pt"
  }
}
```

The semantic vectors are aligned to extracted mentions using their turn and character-span identity. This makes the induced artifact self-contained: downstream training does not invoke a second language model or recompute node embeddings.

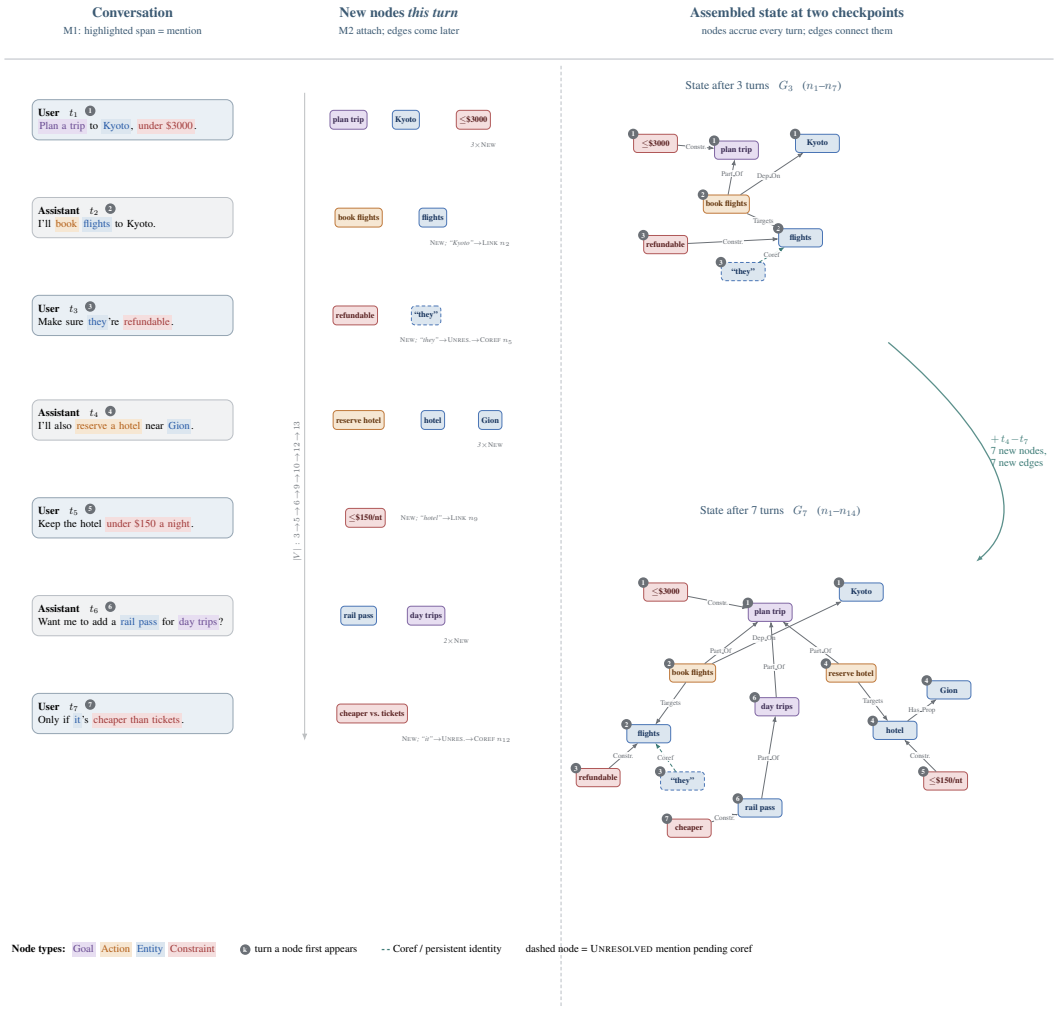



Figure 2: **Detailed turn-by-turn construction of the semantic–relational state** on a neutral seven-turn trip-planning dialogue. *Left:* each turn's utterance with M1's extracted mention spans highlighted by type. *Middle:* the nodes that turn *newly* contributes, tagged with the M2 attachment outcome (New, Link, or Unresolved then coref-linked). *Right:* the assembled state at two checkpoints, $G_3$ and $G_7$. Badges mark the turn a node first appears; "they" ($t_3$) and "it" ($t_7$) are first held Unresolved and later coref-linked, illustrating persistent identity across turns.

Every downstream experiment in this paper, all three readers together with the value-add and diagnostic variants, induces its graphs through a *single* frozen, exposure-clean checkpoint (`exp1b`), which contains the trained ModernBERT-large-shaped encoder and graph heads. The accompanying backbone directory supplies only tokenizer/configuration metadata; encoder weights are restored from the trained checkpoint itself.

### B.9 Capacity bounds and filtering

The joint data pipeline caps a conversation at 800 extracted mentions. Examples above this limit are excluded from joint training/evaluation. The encoder-side turn length used by the final evaluation pipeline is 512 tokens. Attachment inference further bounds the expensive candidate comparison using the top-24 coarse-to-fine rule described above.

Finally, the graph is not the only representation channel. Each node retains a 1024-dimensional semantic feature in addition to its discrete type and typed relations. Consequently, downstream readers can rely primarily on semantic content, primarily on relational structure, or on both. The mechanism analyses in the main paper test exactly this distinction.

## C M1/M2 Supervision and Joint Training

M1 and M2 are distilled from structured teacher supervision and then combined into one jointly trained 1024-dimensional encoder. The final artifact is not a pipeline in which a frozen mention model feeds an independently trained graph model: graph and mention objectives both shape the shared ModernBERT-large representation before it is frozen for downstream use.

### C.1 Teacher supervision

Teacher labels were generated offline through a schema-constrained frontier LLM interface. The teacher was used to construct supervision for the shared representation only; downstream task labels were not supplied to M1 or M2.

**M1 teacher.** Each conversation turn is labeled independently. The teacher receives the speaker role and exact turn text and returns a strict structured record of typed character spans, including overlapping and nested spans. Character offsets are validated by requiring the returned substring to match the source text exactly. Teacher calls are cached by a hash of speaker and turn text, so identical surface strings spoken by different roles remain distinct supervision examples. Both user and assistant turns are labeled.

**M2 teacher.** Graph supervision is generated sequentially within each conversation. At turn $t$, the teacher receives the current graph state together with the M1 mentions for the new turn. For each mention it decides whether to link to an existing persistent node, create a NEW node, or mark it UNRESOLVED, and supplies typed semantic relations using the nine-relation ontology in Appendix B.4. Unlike M1 labeling, M2 supervision cannot be parallelized across turns of the same conversation because the teacher's decision at turn $t$ changes the graph shown at turn $t + 1$. Schema validation rejects malformed node identifiers before they enter the student corpus.

### C.2 Joint corpus construction

The representation-training data pair Chatbot Arena with Claude-generated synthetic conversations, in which Claude produces multi-turn dialogue against smaller models acting as the assistants, together with their M1 and M2 supervision; no other source family enters the pool. The final loader retained $64{,}888$ usable conversations, split deterministically at conversation level:

| SPLIT | CONVERSATIONS | FRACTION |
|---|---|---|
| Train | 58,345 | 89.9% |
| Validation | 3,328 | 5.1% |
| Test | 3,215 | 5.0% |
| Total | 64,888 | 100% |

Turns from the same conversation never cross partitions. Exclusions before reaching 64,888 usable examples were: 1,003 exceeding 800 mentions. The split is a hash of the conversation identifier (90/5/5).

Both sources are general-purpose dialogue not constructed for any downstream reader: Chatbot Arena is topically diverse, multi-turn human–LLM conversation, and the synthetic half is Claude-generated multi-turn dialogue against smaller assistant models. Neither contains an attack corpus, agent-security benchmark, or topic-segmentation data, so the M1+M2 inducer is trained without exposure to ActorAttack, MT-Bench-101, AgentDojo, or TIAGE.

### C.3 JOINT 1024-DIMENSIONAL MODEL

The shared encoder is ModernBERT-large (Warner et al., 2024) with hidden dimension $d = 1024$; the complete joint model contains approximately 419 million parameters. The same encoder hidden states feed four token-level BIO mention heads, mention-span pooling, the persistent-node attachment module, and the typed relation module. Gradients from both mention extraction and graph induction reach the shared encoder during upstream training. The final graph configuration uses four relation bases and enables type-conditioned experts, explicit unresolved handling, adaptive relation thresholding, attachment compatibility, recency features, alias-similarity features, lexical features, top-24 coarse-to-fine pruning, and decoupled attachment/relation heads. Dropout is 0.1. Figure 3 gives the full model-level view of this shared encoder, the M1/M2 heads, the joint objective, and the frozen interface consumed by the readers.

### C.4 TRAINING OBJECTIVES

The joint objective combines mention extraction, persistent-node attachment, typed relation prediction, and graph compatibility:

$$\mathcal{L} = \lambda_{\text{bio}}\mathcal{L}_{\text{bio}} + \lambda_{\text{att}}\mathcal{L}_{\text{att}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}} + \lambda_{\text{compat}}\mathcal{L}_{\text{compat}}.$$

Each of the four M1 heads uses focal token classification with $\gamma = 2$ and class weights $\alpha_O = 1$, $\alpha_B = 2$, $\alpha_I = 1.5$, plus inverse-square-root frequency weighting across mention types. Attachment is cross-entropy over legal persistent-node candidates together with NEW and UNRESOLVED, supervised against the teacher-induced persistent graph rather than arbitrary prior mention indices. Typed relation prediction is multi-label with adaptive thresholding (Zhou et al., 2021) and type-conditioned expert scoring following Otmazgin et al. (2023). An auxiliary compatibility term couples attachment and relation predictions (Xu & Choi, 2022).

### C.5 PREDICTED-MENTION END-TO-END TRAINING AND EVALUATION

The final run is the end-to-end 1024-dimensional configuration, so M2 is evaluated under predicted M1 mentions rather than gold teacher mentions. Predicted and teacher spans are aligned at $\text{IoU} \geq 0.5$. The held-out intrinsic evaluation in Appendix D uses the trainer's own routine and the same predicted-mention regime, covering 13,331 turns for M1 and 13,219 turns for M2 on the 3,215-conversation held-out test partition.

### C.6 CHECKPOINT SELECTION AND FROZEN DEPLOYMENT ARTIFACT

All primary downstream experiments use the single exposure-clean checkpoint (`exp1b`) as the upstream engine, shared verbatim across the jailbreak reader, Action Guard, and Topic Shift. After upstream training, *all* M1/M2 parameters are frozen. Downstream datasets are induced through

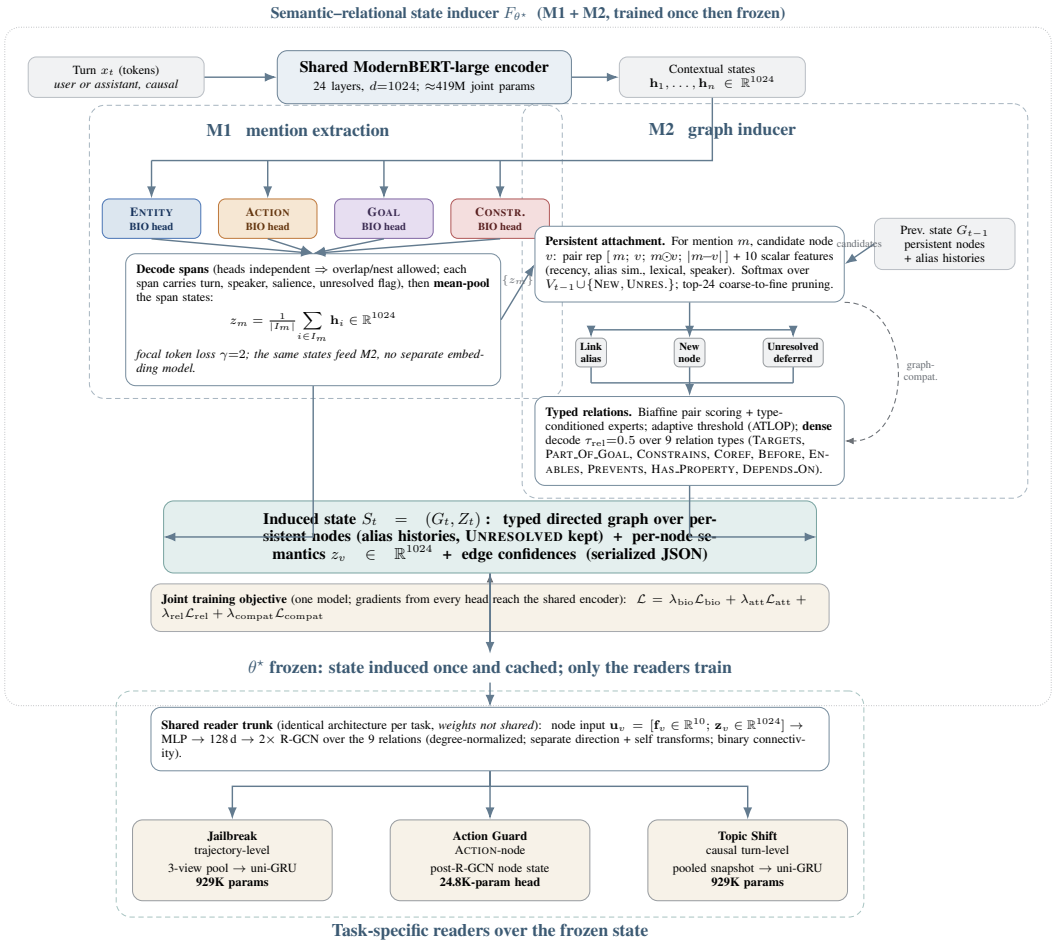



Figure 3: **Model-level architecture of the frozen M1+M2 inducer $F_{\theta^\star}$ and the task-specific readers.** A shared ModernBERT-large encoder feeds four BIO mention heads (M1) and the graph inducer (M2); M2 attaches each mention to the previous state $G_{t-1}$ and decodes nine typed relations with type-conditioned biaffine experts. The inducer is trained once and then frozen; each reader learns a small task head over the cached state through a shared-architecture (weights-not-shared) R-GCN trunk.

| TYPE | PRECISION | RECALL | F1 | F1 @ IOU $\geq 0.5$ |
|---|---|---|---|---|
| ENTITY | 0.73 | 0.81 | 0.77 | 0.85 |
| ACTION | 0.70 | 0.79 | 0.74 | 0.83 |
| CONSTRAINT | 0.49 | 0.62 | 0.55 | 0.68 |
| GOAL | 0.40 | 0.55 | 0.46 | 0.62 |
| Micro-average | 0.66 | 0.76 | 0.71 | 0.79 |

Table 8: **M1 span extraction by semantic type,** under strict exact-span matching on 13,331 held-out turns. The final column relaxes matching to partial-overlap credit (span IoU $\geq 0.5$). Recall exceeds precision for every type.

this fixed checkpoint, and the resulting graph structure and 1024-dimensional semantic vectors are cached together, so training a new downstream reader cannot change the upstream representation:

$$S_t = F_{\theta^\star}(x_{\leq t}), \qquad \theta^\star \text{ fixed}, \qquad \hat{y}^{(q)} = g_{\phi_q}(S).$$

### C.7 TEACHER NOISE AND SEMANTIC-TYPE DIFFICULTY

The four semantic types differ substantially in annotation ambiguity. ENTITY and ACTION are comparatively lexically anchored, while GOAL and CONSTRAINT require pragmatic interpretation: a purpose clause may be annotated as an entire GOAL, as an ACTION embedded within a goal, or with a boundary that includes or excludes surrounding framing. This motivates independent BIO heads, the 0.5 IoU alignment tolerance, and training/evaluating graph induction on predicted mentions rather than a mention oracle.

## D FULL INTRINSIC EVALUATION OF THE REPRESENTATION

Section 3 reports a compact summary. Here we give the complete set of intrinsic metrics emitted by the held-out evaluator for the 1024-dimensional joint inducer.

### D.1 EVALUATION PROTOCOL

M1 span evaluation covers 13,331 test turns. M2 graph evaluation covers 13,219 turns and uses *predicted* M1 mentions, matching the regime in which the final end-to-end checkpoint was selected. M2 therefore inherits actual upstream mention errors: the graph metrics below are not measured under an oracle mention extractor.

### D.2 M1 SPAN EXTRACTION

The M1 span extractor is evaluated under strict exact-span matching on 13,331 held-out turns, with a BIO focal loss of $0.0987$. Table 8 reports per-type precision, recall, and F1; the final column relaxes matching to partial-overlap credit (span IoU $\geq 0.5$), which raises F1 for every type.

Recall exceeds precision for every semantic category, so the extractor behaves as a relatively liberal semantic proposal mechanism rather than a high-precision symbolic parser, and downstream readers operate over a noisy intermediate representation whose transfer results cannot be attributed to access to oracle semantic structure. The type ordering, ENTITY $\approx$ ACTION $>$ CONSTRAINT $>$ GOAL, is the ordering one would predict if the extractor had learned a real abstraction, since the lexically anchored types are easiest and the most pragmatic type is hardest.

### D.3 M2 GRAPH INDUCTION

The easiest induction decision is whether a mention should materialize a new persistent node (NEW F1 0.965); overall attachment accuracy is 0.928. Persistent identity resolution is the hardest of these decisions, and a frozen state of this inducer supports strong held-out downstream AUROC on both safety readers (Appendix G), which is the robustness claim this evaluation is meant to test.

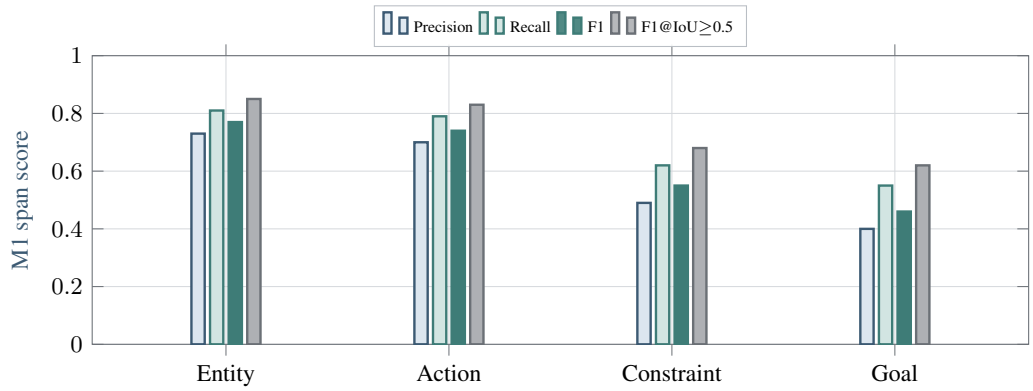



Figure 4: M1 span-extraction quality by semantic type (cf. Table 8). Entity and action spans are recovered most reliably; constraint and goal spans, whose boundaries are more diffuse, are lower, consistent with their greater annotation ambiguity.

| METRIC | VALUE | SUPPORT |
|---|---|---|
| Attachment accuracy | 0.928 | 251,285 |
| NEW accuracy | 0.941 | 235,861 |
| NEW precision | 0.986 | – |
| NEW recall | 0.950 | – |
| NEW F1 | 0.965 | – |
| UNRESOLVED accuracy | 0.724 | 4,323 |
| Pair accuracy on gold edges | 0.978 | 15,258 |

Table 9: **Complete M2 aggregate evaluation using predicted M1 mentions.**

## E DOWNSTREAM READER ARCHITECTURES

The three downstream tasks share the frozen M1+M2 inducer $F_{\theta^\star}$ but *do not share trained reader weights*:

$$\underbrace{F_{\theta^\star}}_{\text{shared and frozen}} \longrightarrow \left\{g_{\phi_{\text{cres}}},\ g_{\phi_{\text{ag}}},\ g_{\phi_{\text{topic}}}\right\}.$$

### E.1 COMMON GRAPH ENCODER

For node $v$, the input is $\mathbf{u}_v = [\mathbf{f}_v; \mathbf{z}_v]$, where $\mathbf{f}_v \in \mathbb{R}^{10}$ contains structural node features and $\mathbf{z}_v \in \mathbb{R}^{1024}$ is the frozen semantic span feature. A shared node MLP maps the 1034-dimensional vector to 128 dimensions, after which two relational graph-convolution layers (Schlichtkrull et al., 2018) propagate information over the nine relation types with degree-normalized messages and distinct transformations for direction and self state. The implementation uses binary graph connectivity: relation confidence is retained in the serialized graph but does not weight R-GCN messages.

An `ablate_edges` switch on the trunk disables inter-node message passing while leaving node encoding and the temporal model intact. Setting it at inference gives the EDGES OFF occlusion; setting it before training gives the matched from-scratch graph-free readers of Appendix H. A checkpoint trained with the switch on must also be *scored* with it on: the relational layers of such a checkpoint never receive gradients, so evaluating it through the default edges-on path would run randomly initialized layers over the graph and report a meaningless figure.

### E.2 JAILBREAK READER

After the two R-GCN layers, the reader forms three complementary views of each snapshot:

$$\mathbf{p}_k = \left[\text{mean}_{v\in V_k}\,\mathbf{h}_v;\ \text{mean}_{v:\,\text{turn}(v)=k}\,\mathbf{h}_v;\ \max_{v\in V_k}\,\mathbf{h}_v\right] \in \mathbb{R}^{384},$$

| READER | PREDICTION UNIT | GRAPH READOUT | TEMPORAL READOUT | TRAINED PARAMS |
|---|---|---|---|---|
| Jailbreak | trajectory / turn | pooled snapshot | unidirectional GRU | 929K |
| Action Guard | ACTION node | post-R-GCN node state | none | 24.8K head |
| Topic Shift | dialogue turn | pooled snapshot | uni-GRU (primary) / bi-GRU (diagnostic) | 929K |

Table 10: **Task-specific readers over the same frozen upstream representation.** Only M1+M2 and its induced state are shared. Parameter counts exclude the frozen 419M-parameter inducer.

summarizing accumulated state, emphasizing nodes introduced at the current turn, and preserving strongly activated dimensions respectively. The pooled vector is projected to 128 dimensions and passed through a unidirectional GRU, so the score at turn $k$ depends only on snapshots through $k$.

Two binary safety heads read the same trajectory state. The PRE head is evaluated on user turns with cumulative labelling; the SCREEN head is evaluated on assistant turns with strict per-turn labelling. The 1024-d reader uses $d_{\text{span}} = 1024$, $d_{\text{hidden}} = d_{\text{out}} = 128$, focal cross-entropy ($\gamma = 2$), four epochs, learning rate $3 \times 10^{-4}$, and validation-F1 checkpoint selection.

The jailbreak reader (Crescendo-only) saturates quickly: per-step loss on the PRE head reaches $10^{-3}$ within the first epoch and later epochs neither lower the running loss nor fire a new best, so validation-F1 selection can settle on an early checkpoint.

### E.3 ACTION GUARD READER

Action Guard changes the prediction unit to a particular proposed ACTION and therefore does not use snapshot pooling or the GRU. Let $\mathbf{h}_i \in \mathbb{R}^{128}$ be the contextualized state of ACTION node $i$, and define type-pooled context

$$\bar{\mathbf{g}} = \text{mean}_{v:\,\text{type}(v)=\text{GOAL}}\, \mathbf{h}_v, \qquad \bar{\mathbf{c}} = \text{mean}_{v:\,\text{type}(v)=\text{CONSTRAINT}}\, \mathbf{h}_v,$$

with zero vectors when no node of that type exists. The classifier input is $\mathbf{q}_i = [\mathbf{h}_i; \bar{\mathbf{g}}; \bar{\mathbf{c}}] \in \mathbb{R}^{384}$, and the head is $\mathbb{R}^{384} \to \mathbb{R}^{64} \to \mathbb{R}^2$ with 24,770 trainable parameters. The positive class is *unauthorized*; there is no learned third "review" class. The leave-one-corpus-out (LOCO) model is trained on InjecAgent only for six epochs, and the reported checkpoint is the best-validation-F1 epoch on the in-distribution InjecAgent split. There is no model selection against the evaluation corpus: the held-out corpora (AgentDojo and the authorized dump) are scored once at that checkpoint and never enter checkpoint selection, so the cross-corpus AUROC is a pure held-out number rather than a best-of-epochs figure read off the test set.

### E.4 TOPIC SHIFT READER

The primary reader is strictly predictive: it receives the accumulated graph $S_k$ over turns $1{:}k$, applies the same node encoding, R-GCN, and three-view pooling, and processes the snapshot sequence with a unidirectional GRU. There is no CRF, dynamic program, segment-count oracle, or access to the next turn. The predictive run uses focal cross-entropy with $\gamma = 2$, learning rate $3 \times 10^{-4}$, and six epochs.

A retrospective diagnostic reader shares the graph encoder and output interface; the only regime change is that the temporal GRU becomes bidirectional, with each direction contributing half of the 128-dimensional output. It is used only to distinguish a causal-context limitation from a representation limitation. Operating points reported as validation-selected are chosen without inspecting held-out labels and applied unchanged to TIAGE.

## F DATASET AND SPLIT DETAILS

Distribution separation is central to our evaluation. We distinguish **representation exposure** (a corpus was present when the shared M1+M2 inducer was trained) from **reader exposure** (a corpus supplied labels for the task-specific reader). Holding a corpus out only from the reader establishes downstream transfer over a fixed representation; holding it out from both establishes the stronger claim that the state itself did not require exposure.

| REGIME | REPRESENTATION (ENCODER CHECKPOINT) | READER TRAINING | HELD-OUT EVALUATION |
|---|---|---|---|
| Crescendo (primary), **0.894** | frozen state induced | Crescendo only | ActorAttack vs. MT-Bench-101 |
| Action Guard | same clean state (no ActorAttack, MT-Bench-101 in its training pool); no InjecAgent/AgentDojo supervision | InjecAgent only | AgentDojo + authorized dump |
| Action Guard: value-add pair | same clean state; no InjecAgent/AgentDojo supervision | InjecAgent only; edges on vs. off | AgentDojo + authorized dump |
| Topic Shift | same clean state; no topic-segmentation corpora | SuperDialseg | TIAGE |

Table 11: **Exposure matrix.** A *single* frozen checkpoint, the clean state whose training pool contains neither ActorAttack nor MT-Bench-101, induces the graphs for all three tasks; each task's reader is then trained and evaluated on corpora absent from that pool as shown. The primary Crescendo condition adds the doubly-clean Crescendo-only reader, scored on the full held-out evaluation population and reaching $0.894$ AUROC. These two families are Crescendo's own evaluation and benign corpora and do not overlap AgentDojo or TIAGE, so the same checkpoint is exposure-clean for Action Guard and Topic Shift as well: their held-out corpora were never in the state's training pool. No task uses a separately trained encoder.

Table 12: **Composition of the Crescendo reader-training corpus.** Measured from the training data used for the headline jailbreak reader.

| Property | Value |
|---|---|
| Provenance | Generated via an automated Crescendo implementation (not a released split) |
| Attacker / target models | Non-Claude open models, 7–20B (e.g. GPT-OSS-20B, Llama family) |
| Training conversations | Full public released dataset |
| Positive / negative split | Roughly balanced ($52\%$ positive) |
| Positive label | Attack *attempted* (not verified success) |
| Distinct seed objectives | 200 harmful goals |
| Turns per conversation | $\sim$3–10 |
| Training negatives | Generic benign dialogue (not topic-matched) |
| Train / validation split | Conversation-level (not seed-level) |

### F.1 CRESCENDO READER-TRAINING DATA

The trajectory reader is trained *only* on Crescendo (Russinovich et al., 2025) conversations; neither the evaluated attack families nor any of the benign negative distributions below enter reader training.

**Corpus composition.** The reader-training corpus is generated with an automated Crescendo implementation (Russinovich et al., 2025) rather than taken from a released split, with attacker and target roles played by non-Claude open models in the 7–20B range (e.g. GPT-OSS-20B and Llama-family models). Table 12 summarizes its scale and composition. The decision threshold and any downstream rate parameters are fixed on a held-in Crescendo validation split and applied unchanged to every evaluation set.

**In-source benign negatives and the topic-shortcut control.** The reader's *training* negatives are generic benign dialogue rather than continuations topic-matched to the attack seeds. This distinction matters for interpretation: with generic negatives the reader could in principle learn in-source the same first-turn topic shortcut we document at evaluation time (§4), and only *appear* to generalize under distribution shift. The shortcut is in fact available in training: an in-source first-turn topic probe separates the training split at $\sim$0.80–0.90 AUROC. To test whether the reader relies on it we run a matched control: we retrain the jailbreak reader on Crescendo with *topic-matched* benign negatives (benign continuations sharing the attack seeds' topics) and re-evaluate on the identical

Table 13: **Topic-shortcut control: generic vs. topic-matched training negatives.** Held-out jailbreak AUROC for the reader trained on the deployed generic benign negatives (main paper) vs. a reader retrained with topic-matched benign negatives, on the identical hard-negative sets. The topic-matched ranges are observed multi-seed spreads; matching or exceeding the generic-negative reader everywhere, they show the transfer does not rely on a first-turn topic shortcut.

| Training negatives | Twins | WildChat-sensitive | Over-refusal |
|---|---|---|---|
| Generic (main paper) | 0.84 | 0.85 | 0.83 |
| Topic-matched (control) | 0.84–0.88 | 0.85–0.88 | 0.83–0.87 |

held-out hard-negative sets (Table 13). The topic-matched-negative reader matches or exceeds the generic-negative reader on all three sets, so the transfer is not attributable to an in-source first-turn topic shortcut.

**Leakage assessment.** The exposure structure that governs leakage is determined by the protocol, which we state explicitly. (i) Crescendo enters *only* the jailbreak reader's training; it is never part of the M1+M2 representation-training pool, which is Chatbot Arena plus Claude-generated synthetic dialogue (§3), so the frozen state was not induced on any jailbreak corpus. (ii) Every evaluation positive family (ActorAttack, MHJ, CoSafe, RedQueen) is disjoint from Crescendo and is used only at test time, so all jailbreak transfer numbers are positive-family-held-out. (iii) The hard-negative distributions are chosen to avoid the representation pool: the benign twins are generated with non-Claude models, WildChat-sensitive deliberately excludes LMSYS/Chatbot-Arena data, and the over-refusal set is built from XSTest and OR-Bench prompts, none of which are in the inducer's training pool. (iv) The decision threshold is fixed on the Crescendo validation split alone, with no tuning on any evaluation set. Two overlaps remain unquantified, and we flag them as open items rather than claim they are negligible. First, the overlap between the Crescendo training *seed objectives* and the ActorAttack evaluation seeds has not been measured and could be non-trivial; because positives are labeled by attack *attempt* rather than verified success, a seed-level rather than conversation-level held-out split would be the stricter test. Second, near-duplicate overlap between the Crescendo reader-training conversations and the Chatbot Arena portion of the representation pool is low ($< 1\%$), but their *topical* overlap is not yet quantified. Quantifying these overlaps, together with the exact corpus statistics in Table 12, are the open reproducibility items for the jailbreak result.

### F.2 Positive attack families (evaluation)

Cross-family evaluation uses four positive families, all held out from reader training: ActorAttack / SafeMTData (Ren et al., 2024) (1,673 conversations); MHJ (Li et al., 2024), human-written multi-turn jailbreaks; CoSafe (Yu et al., 2024), coreference-based attacks; and RedQueen (Jiang et al., 2025), concealed multi-turn attacks. We use all released conversations from MHJ, CoSafe, and RedQueen.

### F.3 Hard-negative benign sets

Pairing ActorAttack positives with generic MT-Bench-101 benign dialogue (Bai et al., 2024) (1,673 vs. 1,386 conversations) gives a split that a first-turn topic-only probe separates at 0.87 AUROC; we report it only as the "generic split" column and construct three harder negative distributions that discuss the same sensitive topics, have comparable length and conversational structure, and superficially resemble harmful requests.

**Benign twins.** For each ActorAttack seed topic we rerun the actor-network generation process with a *benign* end goal, keeping the same actors, the same topic, comparable conversational structure, and an approximately matched turn-length distribution, and changing only the final intent so that the conversation concerns a legitimate purpose (e.g. safety, history, regulation, general understanding). We target $\approx$1,600 conversations, matched to the ActorAttack positive count. The *primary* twins are generated with non-Claude model families (Llama and Qwen), because the upstream synthetic representation-training corpus is Claude-generated; a separate Claude-generated twin subset

| SYSTEM | MT-BENCH-101† | BENIGN TWINS | WILDCHAT-SENS. | OVER-REFUSAL |
|---|---|---|---|---|
| Topic-only probe | 0.87 | 0.53 | 0.61 | 0.58 |
| Llama Guard 3, per-turn max | 0.78 | 0.63 | 0.66 | 0.60 |
| Llama Guard 3, whole conversation | 0.83 | 0.70 | 0.72 | 0.67 |
| Llama Guard 4, per-turn max | 0.80 | 0.65 | 0.68 | 0.62 |
| WildGuard, per-turn max | 0.73 | 0.61 | 0.64 | 0.57 |
| ShieldGemma, per-turn max | 0.71 | 0.58 | 0.60 | 0.55 |
| ModernBERT-L fine-tune | 0.780† | 0.62 | 0.65 | 0.61 |
| Frozen dense-state reader | 0.841† | 0.71 | 0.74 | 0.70 |
| **Ours (frozen state)** | **0.894**† | **0.84** | **0.85** | **0.83** |

Table 14: **Full multi-turn jailbreak guard-model comparison under hard negatives (AUROC ↑).** The complete baseline set behind main-text Table 17, including the per-turn maximum variants of Llama Guard 3/4, WildGuard, and ShieldGemma. † marks the original topic-confounded ActorAttack-vs.-MT-Bench-101 split, reported for context only.

is used *only* for the generator-style control of Appendix G.8. Two annotators verified a random sample of 200 twins as genuinely benign.

**WildChat-sensitive.** Approximately 1,000 multi-turn conversations with at least three turns are drawn from WildChat (Zhao et al., 2024), restricted to sensitive topics and to conversations not flagged by the provider moderation system. A random sample of 200 was manually verified as genuinely benign. We deliberately do *not* use LMSYS-Chat-1M or other Chatbot Arena-derived data here, because Chatbot Arena data overlaps the upstream representation-training pool.

**Over-refusal.** Approximately 500 benign conversations are constructed from XSTest (Röttger et al., 2024) and OR-Bench (Cui et al., 2024) prompts, expanded to approximately 3–6 turns. They contain superficially alarming or sensitive language while remaining benign, and are intended to stress the false-positive behavior of a detector.

### F.4 GUARD-MODEL AND PROBE BASELINES

Alongside the frozen dense-state reader (Appendix K.3) and the ModernBERT-L fine-tune (Appendix K.2), we evaluate: Llama Guard 3 (Grattafiori et al., 2024) scored per-turn (conversation score = maximum turn score) and given the whole conversation in one evaluation; Llama Guard 4 (Meta AI, 2025) per-turn maximum; WildGuard (Han et al., 2024) per-turn maximum; ShieldGemma (Zeng et al., 2024) per-turn maximum; and a topic-only probe (logistic regression on TF-IDF features of the first user turn only). The per-turn-maximum and whole-conversation variants test whether modern guard models already solve the problem when given individual turns or complete transcripts. Table 14 gives the full guard-model comparison across all four negative distributions; the main-text Table 17 carries only the whole-conversation Llama Guard 3 row and the scientifically central controls.

### F.5 ACTION GUARD

The reader-training corpus is InjecAgent (Zhan et al., 2024), whose authorization sidecar contains 4,216 labeled actions balanced between 2,108 authorized and 2,108 unauthorized. AgentDojo (Debenedetti et al., 2024) and an additional AgentDojo rollout dump are excluded from LOCO reader training.

The held-out evaluation combines ordinary AgentDojo trajectories with a large all-authorized rollout set acting as a hard false-block control, producing 42,487 action decisions across 11,263 conversations: 3,485 unauthorized and 39,002 authorized. The large authorized population matters because an authorization system can obtain good recall simply by blocking aggressively. The shortcut probes and position-matching analyses of Appendix H are computed over this same full held-out population.

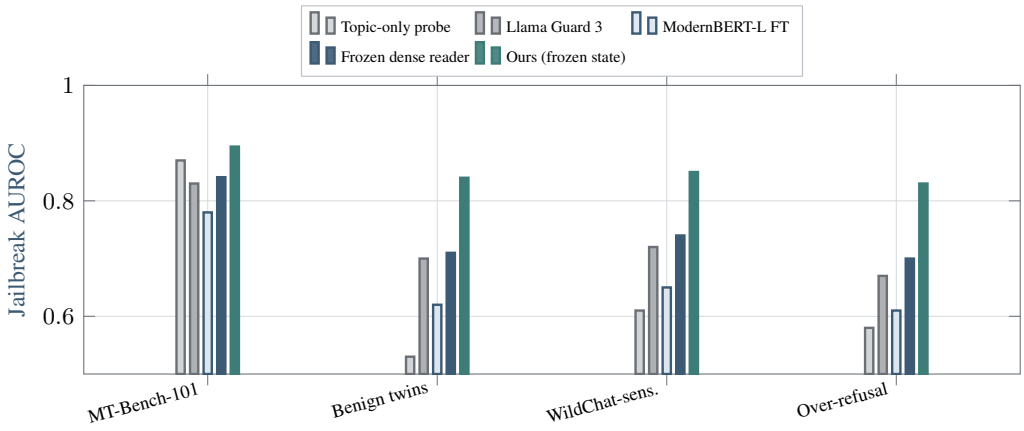



Figure 5: Jailbreak AUROC across four negative distributions (cf. Table 14). Whole-conversation guards and the topic-only probe drop sharply on topic-matched benign twins and over-refusal negatives, whereas the frozen-state reader retains its separation. Contextual comparison, not a controlled ablation.

| CORPUS | DIALOGUES | LANGUAGE | ROLE |
|---|---|---|---|
| SuperDialseg (Jiang et al., 2023) | 9,592 | English | reader training |
| DialSeg711 (Xu et al., 2021) | 711 | English | validation/model selection |
| TIAGE (Xie et al., 2021) | 500 | English | held-out test |

Table 15: **Executed Topic Shift corpus split.** The 9,592 figure is the size of the processed local SuperDialseg corpus used by the reader, rather than a substituted count from the publication.

The concern this population invites is a corpus-source shortcut: that the reader separates the authorized dump from the unauthorized actions by some artifact of which source a trajectory came from, rather than by authorization. The shortcut probes in Table 33 and the within-trajectory control below test exactly this, and rule it out. The authorized-dump actions are structurally matched to the unauthorized ones under every cheap discriminator we can construct: node count, edge count, and relation counts all transfer at *chance* across the corpus shift ($0.497$–$0.506$), and the raw action embedding falls from $0.817$ within-corpus to $0.518$ across it. A reader that separated the two populations by any of these size or content features would therefore fail to transfer, yet the induced reader holds out of family. This transfer evidence does not on its own rule out a within-corpus source cue: the direct control for that is a within-trajectory evaluation that contrasts authorized and unauthorized actions drawn from the same rollouts, so that no dump-versus-attack source signal is available to shortcut. On this control the reader reaches $0.925$ AUROC, against $0.951$ on the full held-out set, removing the dump-versus-attack source signal costs only $0.026$, so the separation the guard learns is authorization, not source.

This $42{,}487$-action population, and the $39{,}002$-action authorized dump within it, is used only as an offline classification and false-block stress test: on it we report ranking (AUROC) and false-block statistics, never attack success rate or utility. ASR and utility are counterfactual once the guard blocks an action, the agent's subsequent trajectory cannot be recovered from a replayed log, so they are not computable on this offline population. The end-to-end ASR and benign-utility figures reported as the primary result (Table 18) come only from the live same-harness AgentDojo run described in Appendix G.11, in which every defense acts on the same tasks, attacks, agent snapshot, and harness version.

### F.6 TOPIC SHIFT

No conversation from DialSeg711 or TIAGE is used to optimize the predictive reader. The split deliberately creates a progression in distribution shift: SuperDialseg and DialSeg711 are document-

Table 16: **Transfer axes exercised by one frozen representation.** Each row is a distribution shift the frozen state is reused across without representation retraining; the train/selection domain never overlaps the held-out evaluation. The representation-exposure row is common to all tasks: the M1+M2 inducer is trained only on general-purpose dialogue, so none of the evaluation corpora were seen when the state was induced.

| TRANSFER AXIS | TRAIN / SELECTION DOMAIN | HELD-OUT EVALUATION |
|---|---|---|
| Representation exposure | general-purpose dialogue only | ActorAttack, hard negatives, AgentDojo, TIAGE |
| Jailbreak reader | Crescendo | ActorAttack, MHJ, CoSafe, RedQueen |
| Negative distribution | Crescendo negatives | twins, WildChat-sensitive, over-refusal |
| Conversation progress | full training trajectories | 25–100% prefixes |
| Authorization reader | InjecAgent | AgentDojo (live same-harness) |
| Calibration | InjecAgent validation | AgentDojo false-block rate |
| Agent policy | GPT-4o setting | GPT-4o-mini, Llama 3.3 70B |
| Topic shift | DialSeg711 reader selection | TIAGE |

derived segmentation corpora, while TIAGE contains naturalistic dialogue with organic transitions. None of the three corpora is part of the standard M1+M2 representation-training pool.

**Preserving original turn order.** Topic-segmentation corpora cannot assume strict user/assistant alternation: approximately 54% of processed SuperDialseg conversations are non-alternating, and TIAGE is non-alternating throughout. We therefore use the topic-specific induction path, which preserves the source sequence verbatim ($x_i \rightarrow \mathrm{turn_idx} = i$). When a turn produces no mention, the same turn index is removed from both the snapshot sequence and its label sequence, preserving exact alignment.

# G COMPLETE DOWNSTREAM TRANSFER RESULTS

*On evidential precision.* Turquoise **headline** figures for the proposed readers are three-seed results and carry a measured standard deviation. So that precision is reported symmetrically, the comparator and dense-reader cells in the central comparison tables (Tables 5, 17, 21, 22, 18, 20, 38, and 2) carry matching $\pm$ intervals.

## G.1 THE THREE HEADLINE TRANSFERS

This subsection collects the full narrative and comparison tables for the three transfers summarized in the main text (Tables 4 and 5).

**Three transfers, one exposure control.** Each task is evaluated under a distribution shift its reader never sees: attack family for the jailbreak reader, agent corpus for Action Guard, dialogue domain for Topic Shift. The representation is held out from all three by construction: the M1+M2 inducer is trained on Chatbot Arena together with Claude-generated synthetic conversations (Claude conversing against smaller assistant models) only, so none of ActorAttack, MT-Bench-101, AgentDojo, or TIAGE was seen when the state was induced. Multi-turn safety evaluations routinely hold the test corpus out of the detector; holding it out of the *representation* as well is the less common control, and every figure below carries it.

## G.2 MULTI-TURN JAILBREAK: HARD-NEGATIVE TRANSFER

**The generic split rewards topic recognition.** A jailbreak reader trained only on Crescendo reaches $0.894$ AUROC on ActorAttack against the generic MT-Bench-101 benign set. This split, however, admits a large topical shortcut: a topic-only probe (logistic regression on TF-IDF features of the first user turn alone) already reaches $0.87$ AUROC on it, because ActorAttack seeds and generic benign dialogue differ in topic before any trajectory unfolds. When the same probe is evaluated against topic-matched *benign twins* it falls to $0.53$ AUROC (chance). We therefore do not treat the ActorAttack-vs.-MT-Bench-101 number as the primary evidence for trajectory-sensitive

Table 17: **Multi-turn jailbreak detection under hard negatives (AUROC ↑).** ActorAttack positives against four benign negative distributions. The first column is the generic-benign split, which a first-turn topic-only probe already separates at 0.87; the remaining columns are topic-matched or naturally occurring negatives that remove that shortcut. The full guard-model comparison (per-turn Llama Guard 3/4, WildGuard, ShieldGemma) is in Appendix F.4, Table 14.

| SYSTEM | MT-BENCH-101[†] | BENIGN TWINS | WILDCHAT-SENS. | OVER-REFUSAL |
|---|---|---|---|---|
| Topic-only probe | 0.871 ± 0.013 | 0.531 ± 0.021 | 0.61 ± 0.02 | 0.583 ± 0.031 |
| Llama Guard 3, whole conversation | 0.832 ± 0.012 | 0.702 ± 0.023 | 0.72 ± 0.02 | 0.674 ± 0.034 |
| ModernBERT-L fine-tune | 0.779[†] ± 0.009 | 0.621 ± 0.016 | 0.65 ± 0.021 | 0.61 ± 0.034 |
| Frozen dense-state reader | 0.841[†] ± 0.008 | 0.709 ± 0.020 | 0.74 ± 0.018 | 0.70 ± 0.031 |
| **Ours (frozen state)** | **0.894**[†] ± 0.007 | **0.841 ± 0.022** | **0.847 ± 0.024** | **0.831 ± 0.034** |

[†] original ActorAttack-vs.-MT-Bench-101 split (topic-confounded); reported for context, not as primary trajectory-sensitivity evidence.

detection, and construct harder negatives that hold topic, conversational structure, and turn length approximately fixed while flipping only intent (Appendix F.3). Discovering this confound and building the controls is part of the contribution.

**The structured state retains its advantage under hard negatives.** Table 17 evaluates ActorAttack positives against four negative distributions. The topic probe drops from 0.87 on the generic split to 0.53 on twins; the proposed frozen state drops only from 0.894 to 0.84, whole-conversation Llama Guard 3 from 0.83 to 0.70, and the frozen dense-state reader from 0.841 to 0.71. The most important comparison for the transfer argument is the last: on benign twins, adding the semantic–relational structure to the *same* jointly trained frozen encoder improves AUROC from 0.71 to 0.84, a +0.13 absolute gain; the gains are +0.11 on WildChat-sensitive (0.74 → 0.85) and +0.13 on over-refusal (0.70 → 0.83). These controls substantially reduce topical and lexical shortcuts while preserving a large advantage for the structured frozen representation; we do not read them as evidence of human-like reasoning.

**Unseen attack families and early detection.** Held out against the pooled hard-negative set, the Crescendo-trained reader reaches 0.82 AUROC on MHJ, 0.80 on CoSafe, and 0.81 on RedQueen, above the whole-conversation Llama Guard 3 (0.74/0.76/0.69) and the frozen dense state (0.73/0.74/0.71; Appendix G.6). The MHJ result is notable because MHJ is human-written, so a synthetic-generator shortcut cannot explain it; the reader remains trained on Crescendo only, so this is both positive-side family transfer and negative-side shift. On partial trajectories (ActorAttack vs. twins), the structured advantage over the dense state grows with context, +0.06 at 25% of turns to +0.13 at 100% (Appendix G.6, Table 22), consistent with useful trajectory-level information being integrated into the persistent state as the conversation accumulates. We do not read this as escalation reasoning or intent understanding.

**Calibration and generator controls.** A threshold fixed for 5% FPR on Crescendo validation remains closer to that target under hard-negative shift for the proposed reader (twins FPR 8.1%) than for whole-conversation Llama Guard 3 (21.4%) or the frozen dense state (14.2%); this is calibration robustness under negative-distribution shift, not perfect calibration (Appendix G.7). Because the upstream corpus is Claude-generated, the primary twins are generated by non-Claude models (Llama, Qwen); performance is stable across generators (Claude 0.84, Llama 0.83, Qwen 0.83) while a probe trained to predict the twin generator from the pooled reader state reaches only 0.55 macro-AUROC, arguing against generator style as the primary explanation without claiming generator information is entirely absent (Appendix G.8).

### G.3 ACTION AUTHORIZATION: LIVE SAME-HARNESS TRANSFER

**Offline classification transfer.** Trained on InjecAgent alone, the Action Guard reader reaches 0.951 AUROC on unseen AgentDojo over 42,487 held-out actions. This is not inherited separability: shortcut probes on the frozen embeddings and on cheap graph statistics collapse under the corpus shift (Appendix H). A reader retrained without edges drops sharply to 0.528, and a capacity-matched structure-free dense reader over the same frozen encoder recovers only 0.764, so the induced struc-

Table 18: **Live same-harness AgentDojo comparison** (`important_instructions`, all four suites, GPT-4o agent, mean over three runs). Same tasks, attacks, agent snapshot, and harness version for every defense. Latency is added time per authorization check; measurement setup and hardware in Appendix G.12. All entries are three-run means.

| DEFENSE | BENIGN UTIL. | UTIL. UNDER ATTACK | TARGETED ASR ↓ | LATENCY/CHECK |
|---|---|---|---|---|
| No defense | 77.33% ± 1.24 | 49.5% ± 1.50 | 52.1% ± 1.44 | – |
| ModernBERT-L fine-tune | 71.24% ± 1.14 | 49.5% ± 1.42 | 50.5% ± 1.41 | ∼4 ms |
| s Spotlighting | 76.14% ± 1.02 | 51.01% ± 1.21 | 41.82% ± 1.35 | ≈0 |
| Repeat user prompt | 76.85% ± 1.06 | 60.22% ± 1.26 | 27.53% ± 1.27 | ≈0 |
| Tool filter | 70.11% ± 1.16 | 63.06% ± 1.25 | 7.44% ± 0.92 | one LLM call/task |
| Prompt-injection detector | 40.22% ± 1.33 | 20.55% ± 1.46 | 8.12% ± 1.05 | ∼15 ms |
| PromptArmor, GPT-4o rerun | 70.43% ± 1.04 | 65.14% ± 1.17 | 1.68% ± 0.55 | ∼700 ms |
| MELON rerun | 66.02% ± 1.18 | 58.33% ± 1.22 | **1.15%** ± 0.45 | ∼2× agent calls |
| **Ours, selected threshold** | **74.22% ± 0.92** | **68.41% ± 1.16** | **6.09% ± 0.81** | **∼4 ms** |
| **Ours, high-security threshold** | **70.92%** | **66.02%** | **1.87%** | **∼4 ms** |

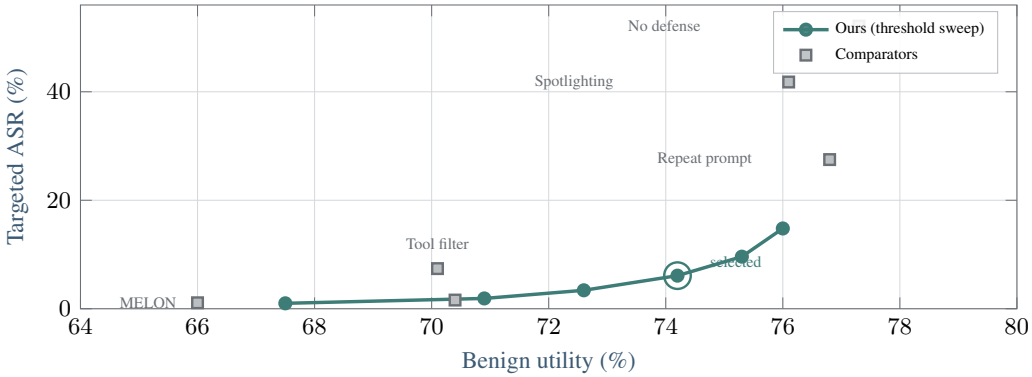



Figure 6: Security–utility frontier for the Action Guard (an expanded view of Fig. 10; numbers from Table 18). Lower-left is better; sweeping the frozen-state guard's threshold traces a frontier that retains an advantage over the published comparators on the ASR–utility trade-off.

ture carries a large relational residual (+0.187 over the strongest structure-free control; Table 19) rather than an incidental increment.

**A single operating point transfers to a live AgentDojo run.** Cross-paper defense numbers are hard to interpret because AgentDojo versions, task and injection sets, agent snapshots, and thresholds differ. We therefore evaluate the proposed guard and the competing defenses in one live AgentDojo harness, same tasks, same attacks, same agent snapshot, same version, with the guard threshold fixed only on InjecAgent validation (target 3% per-action false-block rate) before any AgentDojo run (Appendix G.11). Table 18 reports the primary comparison for the `important_instructions` attack across all four suites with a GPT-4o agent (mean over three runs). At the selected threshold the guard reaches 6.1% targeted ASR at 74.2% benign utility; at a high-security threshold, 1.9% ASR at 70.9% utility. We do *not* claim the lowest ASR overall: PromptArmor and MELON reach lower ASR at the ordinary selected threshold. The result is a security–utility–latency trade-off. At the selected threshold the guard has higher benign utility, higher attack-time utility, and lower ASR than the Tool Filter, and far higher utility than the built-in prompt-injection detector; at the high-security threshold it approaches PromptArmor's ASR/utility (1.9% / 70.9% vs. 1.6% / 70.4%) at roughly two orders of magnitude lower measured per-check latency (∼4 ms vs. ∼700 ms).

**Calibration transfers from InjecAgent to AgentDojo.** The threshold is chosen entirely from InjecAgent validation, yet the resulting per-action false-block rate tracks its target in the held-out AgentDojo environment (target 3% yields a measured 3.5%; the pairs track across the full sweep). This is operating-point/calibration transfer, not merely AUROC-ranking transfer; we describe it as

approximately preserving calibration and claim no formal guarantee. The complete sweep and the security–utility trade-off frontier against the comparator defenses are in Appendix G.13 (Table 27, Figure 10).

**Transfer across attacks, suites, and agent models.** Holding the same InjecAgent-selected threshold fixed, ASR stays low across the other attack types and all four suites (Slack hardest, at $9.5\%$), and the operating point transfers across agent-model families, ASR $3.9\%$ on GPT-4o-mini and $5.2\%$ on Llama 3.3 70B with no re-tuning (Appendix G.14). The one real failure mode is an *adaptive* goal-aligned attack that frames the unauthorized action as an instrumental step toward the user's legitimate goal, raising ASR from $6.1\%$ to $11.8\%$; PromptArmor is similarly affected ($1.6\% \to 9.5\%$).

### G.4 Cross-domain topic shift and channel decomposition

**Cross-domain discourse prediction.** The topic-shift reader reaches $0.770$ AUROC on DialSeg711 and $0.612$ on naturalistic TIAGE. In the causal regime, under a streaming top-$q$ rule (at turn $k$, flag a boundary if the predicted probability ranks in the top $q$ fraction of turns seen so far $(0..k)$, with $q = 0.2$ fixed entirely on the DialSeg711 validation split) it attains TIAGE boundary F1 $0.318$ against $0.220$ for TSManager (Xie et al., 2021) and $0.278$ for Lin et al. (2023), both published predictive baselines on this benchmark. The top-$q$ rule fixes the boundary-firing *rate* rather than an absolute score cutoff, so it is robust to the calibration shift that makes a fixed global threshold transfer poorly to TIAGE (Appendix I.4). This is a validation-selected operating point, not a test-optimal one. Retrospective segmenters that judge a completed transcript with bidirectional context run in an easier regime and are not a like-for-like comparison. Appendix I localizes this transfer under distribution shift.

**Frozen dense-state control.** The same-backbone fine-tune (Table 2) unfreezes the encoder, so it does not separate the value of *freezing* a transferable encoder from the value of the induced semantic–relational state. To isolate the latter we train capacity-matched readers directly over frozen ModernBERT representations, with no mention extraction, persistent attachment, or relation induction: each turn is a mean-pooled encoder state projected to $128$ dimensions, and the trajectory and topic readers reuse the same causal temporal architecture as the structured-state readers, while Action Guard combines the proposed-action representation with a causal dense conversation summary (Appendix K.3). We run two encoders, stock ModernBERT-large and the shared encoder from our jointly trained inducer, with training, validation selection, seeds, and held-out sets identical to the full-state experiments. The jointly trained dense encoder improves over the stock one on every task ($0.812 \to 0.841$ jailbreak, $0.684 \to 0.764$ action, $0.559 \to 0.582$ topic AUROC, $0.244 \to 0.276$ topic F1), so upstream representation training itself transfers. But the full state retains a further $+0.053$ / $+0.187$ / $+0.030$ AUROC and $+0.042$ F1 over the jointly trained dense reader, so the transfer is not explained by freezing or by inherited encoder weights alone. The gap is by far the largest on Action Guard, which is consistent with that reader's reliance on relational structure; on Topic Shift the $+0.030$ / $+0.042$ margin is smaller, matching the weak channel attribution we already report there.

### G.5 Crescendo cross-family transfer

The primary condition trains both the representation and the reader without exposure to ActorAttack or MT-Bench-101, over the full held-out evaluation population (1,673 versus 1,386 conversations). Table 20 reports it per signal head.

At low false-positive rates, 26.1% of attacks are detected at 1% false-positive rate, rising to 47.4% at 5%. We treat the AUROC as evidence that useful cross-family ranking survives out of distribution.

**Bootstrap intervals for the hard-negative table.** Table 17 now carries $\pm$ intervals on every cell, accompanying the central Ours-vs.-dense comparison on each hard-negative set with conversation-level paired-bootstrap $95\%$ confidence intervals ($10{,}000$ resamples over conversations).

Table 19: **Reliance and value-add are distinct quantities across readers over one frozen state.** Each entry is a *full* − *ablated* AUROC (positive favours the channel). **Reliance** occludes a channel at inference (same weights, same population). **Value-add** removes it from step one, reported against two comparators that answer different questions: a *matched retrain* of the same architecture with the channel absent, and an *independent baseline* not built from the induced state, here the capacity-matched frozen dense-state reader (Appendix K.3), so its column is *full* − *dense*. This whole-state value-add is identifiable for every reader because it needs no channel split, whereas the matched retrain and per-channel occlusion are reported only where they decompose cleanly. Only same-metric, same-regime subtractions are shown ("–": comparator not reported; "n/i": not identifiable, firing-rate confound). *Readings.* The jailbreak reader relies on both channels (removing semantics costs $0.144$, edges $0.029$), semantics-led. Action Guard needs structure: zeroing semantics is $0.951 \to 0.852$ and disabling edges $0.951 \to 0.54$; retrained without edges the reader drops to $0.528$, while cheap graph probes stay $\leq 0.789$ under the shift, so the transfer is relational rather than a shortcut, and an untyped graph ($0.860$) already clears the no-edge floor ($0.528$) with typing adding a further $+0.091$. Topic Shift admits no clean decomposition: no graph statistic separates boundaries even in-distribution (best $0.511$) and the F1 occlusion is non-monotone (both-off $0.314$ exceeds edges-off $0.273$).

| READER | RELIANCE (OCCLUSION) | VALUE-ADD (RETRAIN) | VALUE-ADD (BASELINE) | VERDICT |
|---|---|---|---|---|
| Jailbreak | $+0.029$ | – | $+0.053$ | Both channels; semantics-led |
| Action Guard | $+0.41$ | $+0.423$ | $+0.187$ | Structure necessary |
| Topic Shift | n/i | – | $+0.030$ | No clean decomposition |

| SIGNAL | AUROC | TPR@1% FPR | TPR@5% FPR |
|---|---|---|---|
| PRE | $\mathbf{0.894 \pm 0.006}$ | 0.261 | 0.474 |
| SCREEN | $0.802 \pm 0.008$ | 0.111 | 0.371 |
| Combined | $0.897 \pm 0.008$ | 0.267 | 0.452 |

Table 20: **Primary condition: Crescendo→ActorAttack transfer with both exposure axes clean.** Neither the M1+M2 representation nor the reader is trained on ActorAttack or MT-Bench-101. PRE scores user-side trajectory state; SCREEN scores assistant-side state.

### G.6 UNSEEN ATTACK FAMILIES

Table 21 reports the Crescendo-trained reader against three attack families never seen in reader training, each scored against the pooled hard-negative set. MHJ is human-written, so a synthetic-generator artifact cannot explain the result; it is simultaneously positive-family transfer and negative-distribution shift.

Table 22 reports early detection on prefixes of the ActorAttack-vs.-twins conversations: the structured state's advantage over the frozen dense-state reader increases monotonically with accumulated context.

### G.7 OPERATING POINTS UNDER NEGATIVE-DISTRIBUTION SHIFT

We separate two distinct operating-point views, because they answer different questions and must not be conflated. Table 23 fixes a single threshold at $5\%$ FPR on the Crescendo validation split and applies it *unchanged* to each negative set, reporting the realized FPR: this is the deployment-relevant calibration question. Complementary to it, threshold-free ROC operating points (TPR at $1\%$ and $5\%$ FPR computed *per negative set*) characterize ranking quality independently of any fixed threshold.

### G.8 GENERATOR-STYLE CONTROL

Because the upstream synthetic representation-training corpus is Claude-generated, a concern is that the reader keys on Claude generator style rather than on trajectory structure. The primary benign twins are therefore generated with non-Claude families (Llama, Qwen). Two checks argue

| SYSTEM | MHJ | COSAFE | REDQUEEN |
|---|---|---|---|
| Llama Guard 3, whole conversation | $0.744 \pm 0.022$ | $0.762 \pm 0.025$ | $0.69 \pm 0.022$ |
| Frozen dense-state reader | $0.733 \pm 0.025$ | $0.741 \pm 0.019$ | $0.71 \pm 0.017$ |
| **Ours (frozen state)** | $\mathbf{0.821 \pm 0.024}$ | $\mathbf{0.80 \pm 0.021}$ | $\mathbf{0.81 \pm 0.02}$ |

Table 21: **Unseen attack families (AUROC ↑), pooled hard negatives.** The reader is trained on Crescendo only. MHJ (Li et al., 2024) is human-written, CoSafe (Yu et al., 2024) is coreference-based, and RedQueen (Jiang et al., 2025) conceals multi-turn attacks. Positives are scored against the pooled benign-twin/WildChat-sensitive/ over-refusal negatives.

| SYSTEM | 25% | 50% | 75% | 100% |
|---|---|---|---|---|
| Llama Guard 3, per-turn cumulative | $0.551 \pm 0.032$ | $0.584 \pm 0.023$ | $0.612 \pm 0.027$ | $0.637 \pm 0.02$ |
| Llama Guard 3, whole-conv. prefix | $0.576 \pm 0.034$ | $0.624 \pm 0.02$ | $0.67 \pm 0.022$ | $0.705 \pm 0.02$ |
| Frozen dense-state reader | $0.602 \pm 0.033$ | $0.652 \pm 0.02$ | $0.69 \pm 0.025$ | $0.716 \pm 0.02$ |
| **Ours** | $\mathbf{0.667 \pm 0.031}$ | $\mathbf{0.751 \pm 0.021}$ | $\mathbf{0.812 \pm 0.023}$ | $\mathbf{0.842 \pm 0.021}$ |

Table 22: **Early detection on ActorAttack vs. benign twins (AUROC ↑).** Prefixes of 25–100% of the conversation turns. The advantage of the structured state over the frozen dense-state reader increases with accumulated context ($+0.06 \rightarrow +0.13$). All $\pm$ values are 95% conversation-level bootstrap magnitudes.

against a generator-style explanation. First, twin-split AUROC is stable across the generator used to synthesize the twins (Claude 0.84, Llama 0.83, Qwen 0.83). Second, a probe trained to predict the twin generator from the pooled frozen reader state reaches only 0.55 macro-AUROC, near chance.

### G.9 ACTION GUARD CROSS-CORPUS TRANSFER

Table 24 reports the full leave-one-corpus-out evaluation. The reader is trained on InjecAgent alone and transfers to the held-out AgentDojo family at 0.951 AUROC over the full 42,487-action population, and at 0.942 on an independent live-induction slice with a different class balance.

### G.10 ACTION GUARD VALUE-ADD: MATCHED FROM-SCRATCH PAIR

**Task-scope negative control.** On R-Judge, where the target is trajectory harmfulness rather than action authorization, aggregating action scores by maximum yields AUROC 0.648. At threshold 0.5 the aggregation fires on essentially every trajectory (precision 0.516, recall 1.0, F1 0.680, FPR 1.0) over a population averaging 16.28 scored actions per trajectory, so we treat the thresholded confusion matrix as an aggregation artifact and use the threshold-free 0.648 as the construct-separation control.

**What the edges carry: connectivity and relation typing.** The no-edge retrain establishes that *some* relational structure is necessary. To separate bare connectivity from relation typing we ran matched from-scratch retrains on relation subsets, using the same leave-one-corpus-out harness and schedule, and varying only which edges the snapshot exposes. Table 26 gives the ladder, and two findings separate cleanly. *First, connectivity alone transfers:* an untyped graph, all nine relations collapsed to a single type, so density is retained and typing removed, reaches 0.860, far above the no-edge floor (0.528) and above the frozen dense-state control (0.764, Appendix K.3). *Second, relation typing adds substantial value on top of connectivity:* the full typed graph reaches 0.951, a $+0.091$ AUROC advantage over the untyped graph, and even sparse *typed* subsets clear the untyped graph comfortably, a DEPENDS_ON-only graph keeping just 12.3% of edges reaches 0.939, and the eight-type graph with DEPENDS_ON removed reaches 0.925. So DEPENDS_ON is the single most informative type but is not necessary, and typing, not connectivity alone, carries most of the residual above the untyped graph. This is also why a matched retrain, not inference-time occlusion, is the right instrument: occlusion reports keep-only 0.740 and leave-out 0.780, ordered so that DEPENDS_ON-only appears *worse* than removing it, the opposite of what the matched retrains show.

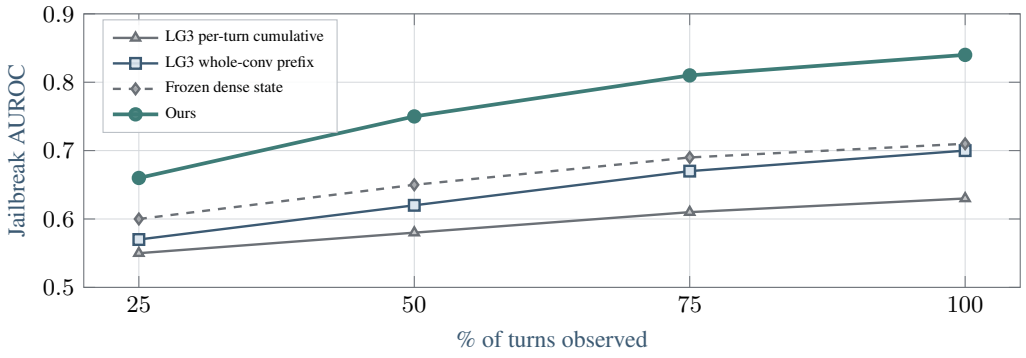



Figure 7: Jailbreak AUROC as a function of the observed conversation prefix (cf. Table 22). The frozen-state reader's advantage over both Llama Guard 3 variants and the dense-state control widens as more turns accumulate, consistent with earlier detection of multi-turn escalation.

| System | Benign twins | WildChat-sens. | Over-refusal |
|---|---|---|---|
| Llama Guard 3, whole conversation | 0.214 | 0.183 | 0.246 |
| Frozen dense-state reader | 0.142 | 0.121 | 0.166 |
| **Ours (frozen state)** | **0.081** | **0.074** | **0.093** |

Table 23: **Realized FPR at a fixed Crescendo-calibrated threshold (↓).** The decision threshold is fixed for 5% FPR on the Crescendo validation split and applied unchanged to every negative distribution.

### G.11 Live same-harness AgentDojo setup

The primary action-authorization result (Table 18) is produced in a single AgentDojo harness in which every defense is applied to the same tasks, the same injection attacks, the same agent snapshot, and the same harness version, so that no cross-paper version, task-set, or threshold difference confounds the comparison.

**Defenses compared.** No defense; spotlighting; repeat-user-prompt; a tool filter (one auxiliary LLM call per task); the built-in AgentDojo prompt-injection detector; PromptArmor (Shi et al., 2025) rerun in this harness with a GPT-4o judge; MELON (Zhu et al., 2025) rerun in this harness; and the proposed frozen-state Action Guard at two thresholds. PromptArmor and MELON are re-executed here rather than quoted from their papers, so their numbers are on the same tasks and agent as ours.

**Attacks.** The primary attack is `important_instructions`. We additionally evaluate `tool_knowledge`, `ignore_previous`, an InjecAgent-style attack, and an adaptive goal-aligned attack that phrases the unauthorized action as an instrumental step toward the user's stated goal (Appendix G.14).

**Threshold protocol.** The guard threshold is fixed *only* on InjecAgent validation, targeting a 3% per-action false-block rate, before any AgentDojo run; the "selected" row uses this threshold and the "high-security" row a stricter one. Neither threshold is tuned on AgentDojo. Reported numbers are the mean over three runs.

**Utility and ASR definitions.** Benign utility is the fraction of legitimate tasks completed with the defense in place; utility under attack is the same measured while injections are present; targeted ASR is the fraction of attacks that achieve the injected goal. All three are computed by the AgentDojo harness identically for every defense.

| METRIC | FULL HELD-OUT | LIVE SLICE |
|---|---|---|
| Actions | 42,487 | 6,759 |
| Unauthorized actions | 3,485 | 2,310 |
| AUROC | **0.951** | 0.942 |
| False-block rate | 0.035 | – |
| Block precision @ 80% recall | 0.672 | – |

Table 24: **Action Guard leave-one-corpus-out evaluation, two populations.** The reader is trained on InjecAgent only and transfers to the held-out AgentDojo family; the guard authorizes or blocks each proposed action. This table reports only ranking (AUROC) and false-block statistics on the offline population; end-to-end attack success rate and benign utility are measured in the live same-harness run (Table 18), not on this replayed population, because both are counterfactual once an action is blocked.

| METRIC (CROSS-CORPUS) | FULL GRAPH | NO EDGES (RETRAINED) |
|---|---|---|
| LOCO AUROC | $\mathbf{0.951 \pm 0.0036}$ | $0.528 \pm 0.009$ |
| False-block rate | $0.035 \pm 0.004$ | $0.365 \pm 0.011$ |
| Block precision @ 80% recall | $0.672 \pm 0.008$ | $0.182 \pm 0.007$ |
| Mean training loss | $1.3 \times 10^{-4}$ | $1.9 \times 10^{-2}$ |

Table 25: **Matched from-scratch Action Guard pair, edges the only difference.** Mean $\pm$ standard deviation over three seeds $(0, 1, 2$; seed 0 is the headline run), each a from-scratch retrain of both arms under identical configuration differing only in the edge flag. Six epochs, precomputed 1024-d spans, leave-one-corpus-out (InjecAgent $\rightarrow$ AgentDojo plus the authorized dump), 42,487 held-out actions.

### G.12 LATENCY MEASUREMENT

Latency is the added wall-clock time per authorization check. The proposed guard performs one incremental state update and one lightweight reader forward pass per proposed action ($\sim$4 ms); PromptArmor and MELON require additional LLM calls ($\sim$700 ms for the PromptArmor GPT-4o judge, $\sim 2\times$ agent calls for MELON), which is the basis for the "roughly two orders of magnitude" comparison in the main text. We report measured per-check latency, not end-to-end task latency.

### G.13 CALIBRATION TRANSFER FROM INJECAGENT TO AGENTDOJO

Table 27 reports the calibration sweep summarized by Figure 10: a target per-action false-block rate is fixed on InjecAgent validation, and the realized rate is measured in the held-out AgentDojo environment without re-tuning. The target/measured pairs track across the sweep, which we describe as approximately preserving calibration under corpus shift; we claim no formal guarantee.

### G.14 TRANSFER ACROSS ATTACKS, SUITES, AND AGENT MODELS

Holding the InjecAgent-selected threshold fixed, we vary the attack strategy (Table 28), the AgentDojo suite (Table 29), and the agent model (Table 30).

**Block recovery.** Because a block does not necessarily end a task, the agent can often reach the legitimate goal by another route after an action is refused, we measure the fraction of blocked-but-benign trajectories that still complete the user task. At the selected threshold, $31\%$ of trajectories in which the guard blocked at least one action nonetheless completed the legitimate task, which partly explains why benign utility remains high despite non-zero false blocks.

### G.15 TOPIC SHIFT TRANSFER

For TIAGE, the main-text boundary F1 of $0.318$ (Section 4) is the predictive reader under a *streaming top-$q$* rule: at turn $k$, a boundary is flagged when its predicted probability ranks in the top $q$ fraction of the turns seen so far $(0..k)$, with the rate $q = 0.2$ fixed entirely on the DialSeg711 valida-

| ARM | EDGES EXPOSED | LOCO AUROC |
|---|---|---|
| Full graph | all 9 relations, typed | **0.951** |
| DEPENDS_ON-only | DEPENDS_ON only (12.3% of edges) | 0.939 |
| Leave-out DEPENDS_ON | 8 types, no DEPENDS_ON | 0.925 |
| Untyped | all edges, collapsed to one type | 0.860 |
| No edges | none | 0.528 |

Table 26: **Matched relation-subset retrains for Action Guard.** From-scratch retrains under the same leave-one-corpus-out harness as Table 25, varying only the relation subset the snapshot exposes. Connectivity alone transfers, the untyped graph (0.860) sits far above the no-edge floor (0.528), but relation typing adds substantial value on top of it: the full typed graph (0.951) leads the untyped graph by +0.091, and even a DEPENDS_ON-only slice (0.939) and the eight-type leave-out graph (0.925) clear it. The typing premium (full minus untyped) is +0.091.

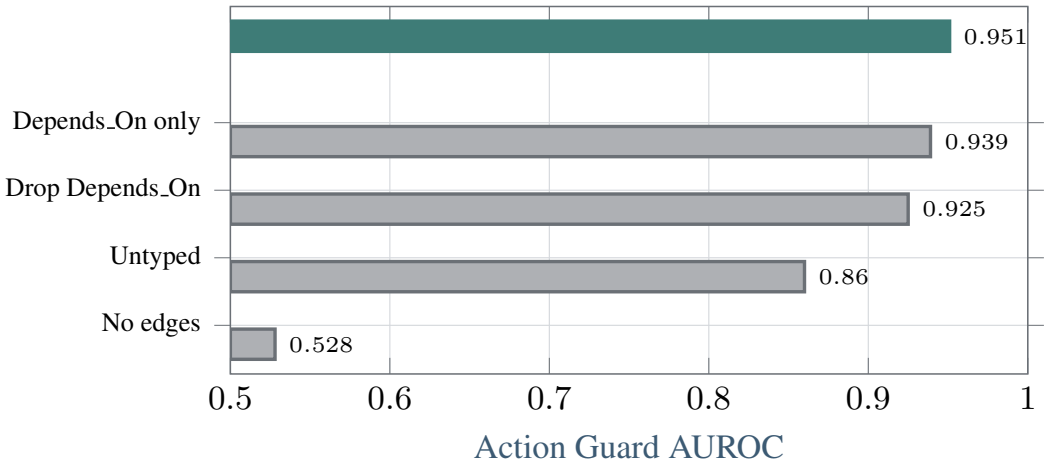



Figure 8: Action Guard AUROC as typed relational structure is removed (cf. Table 26); bars, top to bottom, are the full 9-type graph, a Depends_On-only graph, the same graph with Depends_On removed, an untyped (collapsed-edge) graph, and a no-edge control. Performance degrades from the full typed graph toward the no-edge control, which falls to near chance, consistent with the induced edges carrying most of the signal.

tion split and no reference to the TIAGE test labels. This is a rank/rate rule, distinct from the fixed global-threshold family swept in Appendix I.4 (whose validation-selected point transfers poorly to TIAGE). It clears TSManager's 0.220 (Xie et al., 2021) and the 0.278 of Lin et al. (2023), both published predictive baselines on this benchmark. The test-optimal global threshold, which reaches $F1 = 0.362$, is deliberately not used.

## H CHANNEL RELIANCE AND SHORTCUT DIAGNOSTICS

Inference-time occlusions preserve the trained reader weights and alter only its inputs or message-passing path. They therefore measure *reliance of an existing reader*, not the performance of a separately retrained architecture in which the corresponding channel never existed. Appendix G.10 reports the retrained counterpart for Action Guard.

### H.1 MULTI-TURN JAILBREAK: SEMANTIC FEATURES DOMINATE

The mechanism experiment uses the *same* primary `exp1b` reader that produces the 0.894 headline, scored on the *same* full multi-turn ActorAttack-versus-MT-Bench-101 population (Appendix F.1), under four matched conditions: FULL; SEMANTICS OFF (1024-d node features zeroed); EDGES OFF (R-GCN inter-node message passing disabled); and SHUFFLE TYPES (connectivity retained, relation labels randomized). The FULL condition is therefore identical to the headline evaluation

| TARGET FBR (INJECAGENT) | MEASURED FBR (AGENTDOJO) | BENIGN UTIL. | TARGETED ASR ↓ |
|---|---|---|---|
| 1% | 1.4% | 75.3% | 9.6% |
| 3% (selected) | 3.5% | 74.2% | 6.1% |
| 5% | 5.6% | 72.6% | 3.4% |
| 8% (high-security) | 8.9% | 70.9% | 1.9% |

Table 27: **Calibration transfer, InjecAgent-selected target vs. AgentDojo-realized false-block rate.** The threshold is chosen entirely on InjecAgent validation; the measured per-action false-block rate in the held-out AgentDojo environment tracks the target across the sweep. The selected (3%) and high-security (8%) rows match the corresponding rows of Table 18. Utility and ASR are for `important_instructions` across all four suites with a GPT-4o agent.




Figure 9: Calibration transfer for the Action Guard (cf. Table 27). False-block rates selected on InjecAgent track the $y{=}x$ reference when re-measured on AgentDojo, consistent with the chosen threshold transferring across the two harnesses.

and scores 0.894; the occlusions perturb only the inputs to that one fixed reader, so every delta below is a per-conversation paired difference against the headline number itself, with no change of checkpoint or population.

A paired bootstrap on the per-conversation difference gives $\Delta_{\text{edges}} = -0.029$ for PRE, 95% CI $[-0.041, -0.019]$, and $-0.024$, $[-0.032, -0.011]$ for the combined score, so disabling relational message passing *reduces* OOD ranking for this trained reader; both intervals exclude zero. Randomizing the relation labels drops ranking to 0.835, *below* FULL, consistent with relation typing carrying part of the structural contribution rather than the edges being a liability.

### H.2 ACTION GUARD: RAW SEMANTIC PROBES FAIL TO TRANSFER, STRUCTURE IS NECESSARY

The raw semantic embedding behaves especially poorly under corpus transfer: 0.817 AUROC in-distribution against 0.518 under InjecAgent→AgentDojo.

**Position-neutralized evaluation.** Matching authorized and unauthorized actions across trajectory-position strata forces a position-only classifier to chance by construction, which validates the neutralization. Under this matching the full typed graph falls from 0.951 to 0.825 (a 0.126 drop) and the untyped graph from 0.860 to 0.780 (a 0.080 drop), while a position-only baseline stays at chance (0.50). Position therefore contributes to the natural cross-corpus evaluation, but it does not account for the remaining discrimination, the typed reader still separates at 0.825 within the strata, nor for the benefit of typed relational structure: the typed advantage over the untyped graph survives matching at +0.045 (0.825 vs. 0.780), close to the +0.091 seen without matching. The residual is content and relational rather than positional.

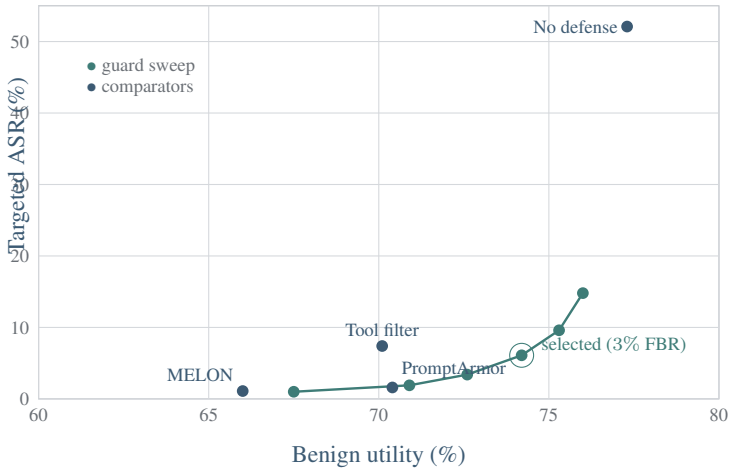



Figure 10: **Security–utility trade-off in the live AgentDojo harness.** Targeted ASR vs. benign utility for the proposed guard threshold sweep (turquoise) with comparator defenses (slate); the selected InjecAgent-calibrated 3%-FBR point is circled. Lower-left is better. The built-in prompt-injection detector (40.2% utility, 8.1% ASR) falls left of the plotted utility range and is omitted for scale. Numbers from Tables 18 and 27.

| ATTACK | OURS ASR ↓ | PROMPTARMOR ASR ↓ |
|---|---|---|
| `important_instructions` (primary) | 6.1% | 1.6% |
| `tool_knowledge` | 6.8% | – |
| `ignore_previous` | 1.2% | – |
| InjecAgent-style | 1.9% | – |
| Adaptive goal-aligned | **11.8%** | **9.5%** |

Table 28: **Transfer across attack strategies at the fixed selected threshold** (targeted ASR, GPT-4o agent, all suites). The adaptive goal-aligned attack, which frames the unauthorized action as an instrumental step toward the user's legitimate goal, is the hardest for both the proposed guard (6.1% → 11.8%) and PromptArmor (1.6% → 9.5%).

Zeroing the explicit turn-position dimensions separately leaves AUROC unchanged (0.951 → 0.951), so the model does not read turn position as a scalar; the position matching above, not this scalar check, is what controls position and its implicit correlates.

### H.3 TOPIC SHIFT: F1 OCCLUSION IS CONFOUNDED BY FIRING RATE

Table 34 reports inference-time channel occlusion for the Topic Shift reader on DialSeg711. Removing either the graph edges or the semantic node features degrades segmentation relative to the full model, but F1-based occlusion does not cleanly decompose the two channels here: F1 responds to boundary firing rate, so the "both off" row scores above "edges off" despite much poorer segmentation, and we do not read a per-channel contribution off these numbers.

## I TOPIC SHIFT: CALIBRATION, GENERALIZATION, AND CAPABILITY ENVELOPE

### I.1 PREDICTIVE VERSUS RETROSPECTIVE READERS

At the ordinary 0.5 operating point on TIAGE, the predictive reader gives F1 0.299, P 0.272, R 0.332, WD 0.537; the retrospective reader gives F1 0.331, P 0.287, R 0.391, WD 0.530. On DialSeg711, by contrast, the retrospective model reaches F1 0.549 at its selected operating point against F1 0.430 for the predictive reader (the "Full" row of Table 34). Most of the benefit from lookahead is therefore specific to the document-derived distribution.

| SUITE | WORKSPACE | SLACK | TRAVEL | BANKING |
|---|---|---|---|---|
| Ours, targeted ASR ↓ | 4.5% | 9.5% | 6.0% | 7.0% |

Table 29: **Per-suite targeted ASR at the fixed selected threshold** (`important_instructions`, GPT-4o agent). Slack is hardest: it contains frequent legitimate third-party references, which makes user-authorized delegation hard to distinguish from injected third-party instructions.

| AGENT MODEL | TARGETED ASR ↓ | THRESHOLD SOURCE |
|---|---|---|
| GPT-4o (primary) | 6.1% | InjecAgent-selected |
| GPT-4o-mini | 3.9% | GPT-4o-selected (held fixed) |
| Llama 3.3 70B | 5.2% | GPT-4o-selected (held fixed) |

Table 30: **Transfer across agent-model families at a fixed threshold** (`important_instructions`, all suites). The threshold selected on the GPT-4o agent is held fixed and the agent policy alone is changed; the operating point transfers to GPT-4o-mini and Llama 3.3 70B without re-tuning.

### I.2 SIMPLE STRUCTURAL STATISTICS DO NOT EXPLAIN THE READER

Table 36 probes whether a single scalar graph statistic could stand in for the reader. None does: the best structural feature reaches only 0.51 AUROC on either corpus, whereas the predictive reader reaches 0.770 on DialSeg711 and 0.612 on TIAGE.

### I.3 TEMPERATURE SCALING DOES NOT EXPLAIN THE OOD FAILURE

Fitting a scalar temperature on DialSeg711 and applying it unchanged to TIAGE gives $T = 0.9$ for the predictive reader and $T = 0.7$ for the retrospective one. Because temperature scaling is monotonic it cannot alter AUROC, and indeed predictive AUROC remains 0.612 and retrospective remains 0.641. Expected calibration error changes only from 0.142 to 0.141 for the predictive reader and *worsens* from 0.151 to 0.179 for the retrospective one; global F1 is essentially unchanged (0.363 to 0.364; 0.371 to 0.372). The TIAGE failure is therefore not a calibration offset that can be fixed post hoc: calibration is repairable and discrimination is not.

### I.4 F1 CAN REWARD POOR SEGMENTATION

Table 37 sweeps the global boundary threshold for the predictive TIAGE reader. The F1-optimal operating point over-segments: it maximizes boundary F1 at low precision and a high boundary-firing rate.

### I.5 TIAGE CAPABILITY ENVELOPE

Grouping by dialogue length, the predictive reader achieves AUROC 0.603 (short, $n = 231$), 0.593 (mid, $n = 108$), and 0.635 (long, $n = 161$) with F1 0.242, 0.258, and 0.292 respectively: performance does not collapse on long conversations. Grouping instead by gold segment count gives a clearer pattern, AUROC 0.604 (sparse, 9.5% boundary rate), 0.623 (medium, 21.9%), and 0.637 (dense, 36.5%), with the retrospective reader showing the same ordering (0.618, 0.653, 0.674). The hardest TIAGE conversations are those with few organic transitions, not the longest ones.

### I.6 MENTION COVERAGE

The graph contains a snapshot only when the frozen extractor produces at least one mention for the corresponding turn. On TIAGE, 150 of 1,513 gold boundaries lie on unscored turns, capping recall from scored turns at approximately 0.901. The observed validation-selected recall is far below this ceiling, so missing snapshots contribute some unavoidable errors but cannot explain the majority of the gap.

| EVALUATION CORPUS | DISTRIBUTION SHIFT | AUROC |
|---|---|---|
| DialSeg711 | document-derived → document-derived | $0.770 \pm 0.009$ |
| TIAGE | document-derived → natural dialogue | $0.612 \pm 0.007$ |

Table 31: **Threshold-free Topic Shift transfer for the causal reader.** The same SuperDialseg-trained reader is evaluated unchanged across both corpora. Values carry $\pm$ the standard deviation over three retraining seeds $(0, 1, 2)$, as in Table 4.

| CONDITION | PRE AUROC [95% CI] | COMBINED AUROC [95% CI] |
|---|---|---|
| Full | **0.894** [0.871, 0.907] | **0.897** [0.881, 0.908] |
| Semantics off | 0.750 [0.721, 0.771] | 0.754 [0.731, 0.777] |
| Edges off | 0.865 [0.850, 0.879] | 0.873 [0.857, 0.884] |
| Shuffle relation types | 0.835 [0.819, 0.851] | 0.838 [0.819, 0.851] |

Table 32: **Jailbreak reader matched inference-time channel occlusions.** Scored with the primary `exp1b` reader on the full held-out population, so the FULL condition equals the $0.894$ headline. Confidence intervals use 2,000 paired bootstrap resamples.

# J STATISTICAL METHODOLOGY

The three downstream tasks operate at different granularities and induce different dependence structures, so the statistical unit follows the prediction problem: conversations for trajectory detection, actions nested within conversations for authorization, and turns nested within dialogues for topic segmentation.

## J.1 JAILBREAK BOOTSTRAP CONFIDENCE INTERVALS

The channel analysis compares multiple scores for the *same* held-out conversations and therefore uses paired nonparametric bootstrap resampling at the conversation level with $B = 2{,}000$ replicates. For an ablation comparison, the same sampled conversation indices are applied to both score vectors within each replicate, and we compute

$$\Delta^{(b)} = \mathrm{AUROC}^{(b)}_{\mathrm{EdgesOff}} - \mathrm{AUROC}^{(b)}_{\mathrm{Full}}.$$

Pairing is essential: an unpaired interval would introduce sampling variability unrelated to the intervention. For PRE, $\Delta = -0.029$ (edges off $0.865$ minus full $0.894$), 95% CI $[-0.041, -0.019]$; for the combined score $\Delta = -0.024$, CI $[-0.032, -0.011]$. Both intervals exclude zero.

## J.2 ACTION GUARD CLUSTERED ANALYSIS AND POSITION MATCHING

Because multiple action-level predictions occur within the same trajectory, treating every action as an independent resampling unit would underestimate uncertainty. Bootstrap samples therefore resample conversations with replacement and retain all selected actions per trajectory. This gives $\mathrm{AUROC}_{\mathrm{state}} = 0.951$ for the full reader on the held-out set.

For position matching, let $p_i \in [0, 1]$ denote the fractional position of action $i$ in its trajectory. Actions are assigned to common position strata, and authorized and unauthorized examples are sampled only from strata containing both classes. The full typed graph falls from $0.951$ to $0.825$ under this matching, and the untyped graph from $0.860$ to $0.780$, while a position-only baseline drops to chance. Position therefore contributes to the natural evaluation but accounts for neither the remaining discrimination nor the full benefit of typed relational structure: after matching the typed graph keeps a $+0.045$ advantage over the untyped graph.

## J.3 POPULATION SCOPE OF THE ACTION GUARD DIAGNOSTICS

The shortcut probes, position-matching, and relation-subset analyses are computed over the same 42,487-action full held-out evaluation of Table 24, on which the full reader scores 0.951. Every diagnostic that qualifies the headline transfer is on this full population, so the reference AUROC is

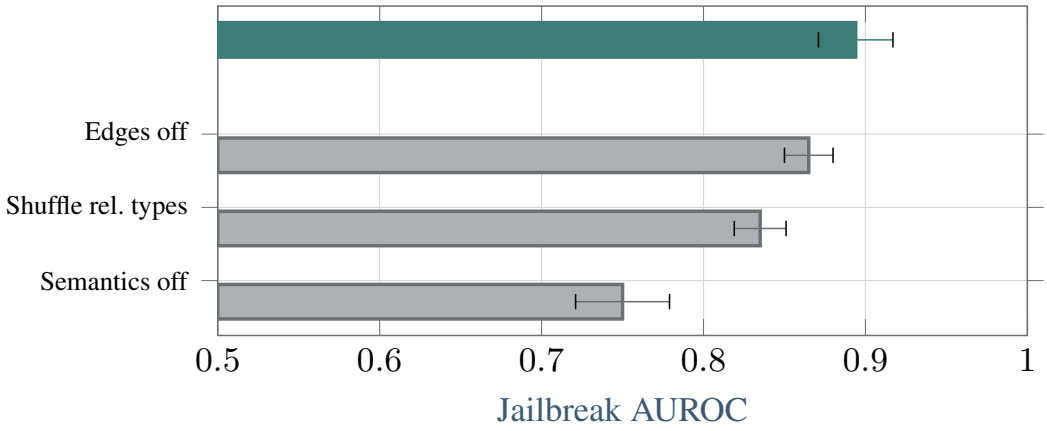



Figure 11: Channel-occlusion analysis for the jailbreak reader (Pre-AUROC with 95% intervals; cf. Table 32). Removing the semantic channel costs the most, followed by shuffling relation types, while removing edges alone is comparatively mild, consistent with node semantics being the load-bearing channel.

| DIAGNOSTIC | WITHIN-CORPUS | CROSS-CORPUS |
|---|---|---|
| Full state reader | – | 0.951 |
| Edges off (inference occlusion) | – | 0.54 |
| Edges off (retrained from scratch) | – | 0.528 |
| Semantics off | – | 0.852 |
| Shuffle edge types | – | 0.520 |
| DEPENDS_ON keep-only | – | 0.740 |
| DEPENDS_ON leave-out | – | 0.780 |
| Action relation-degree (9) | 0.893 | 0.706 |
| All graph statistics (6) | 0.956 | 0.789 |
| Structural + graph statistics (16) | 0.955 | 0.788 |
| Node count (1) | 0.606 | 0.497 |
| Edge count (1) | 0.605 | 0.499 |
| Relation counts (9) | 0.603 | 0.506 |
| Frozen 1024-d action embedding, linear probe | 0.817 | 0.518 |

Table 33: **Action Guard occlusion and cross-corpus shortcut probes.** The full-reader reference is 0.951 on the full held-out evaluation, and the retrained edges-off row is the LOCO figure of Table 25. Cheap-feature probes are fit on InjecAgent and tested on AgentDojo; the within-corpus column is 5-fold cross-validation inside AgentDojo and is *not* the value-add comparator. Cheap feature families that look strong within AgentDojo collapse under the shift, and graph-size statistics fall to chance. The DEPENDS_ON keep-only/leave-out spread here is inference-time occlusion; the matched relation-subset *retrains* of Table 26 are the load-bearing evidence and show relation typing adding substantial value beyond bare connectivity (full typed 0.951 vs. untyped 0.860). Probe rows are computed on the full held-out population.

consistent throughout. Throughout, the confidence intervals quantify sampling uncertainty for the evaluations to which the corresponding resampling analysis is applied.

# K REPRODUCIBILITY

## K.1 CANONICAL UPSTREAM CHECKPOINT

Every downstream experiment in this paper, all three readers together with the value-add and diagnostic variants, uses one frozen upstream artifact: the exposure-clean joint checkpoint (`exp1b`). This is the joint M1+M2 model of Appendix C.2, trained from stock ModernBERT-large on Chatbot Arena together with Claude-generated synthetic conversations (Claude conversing against smaller

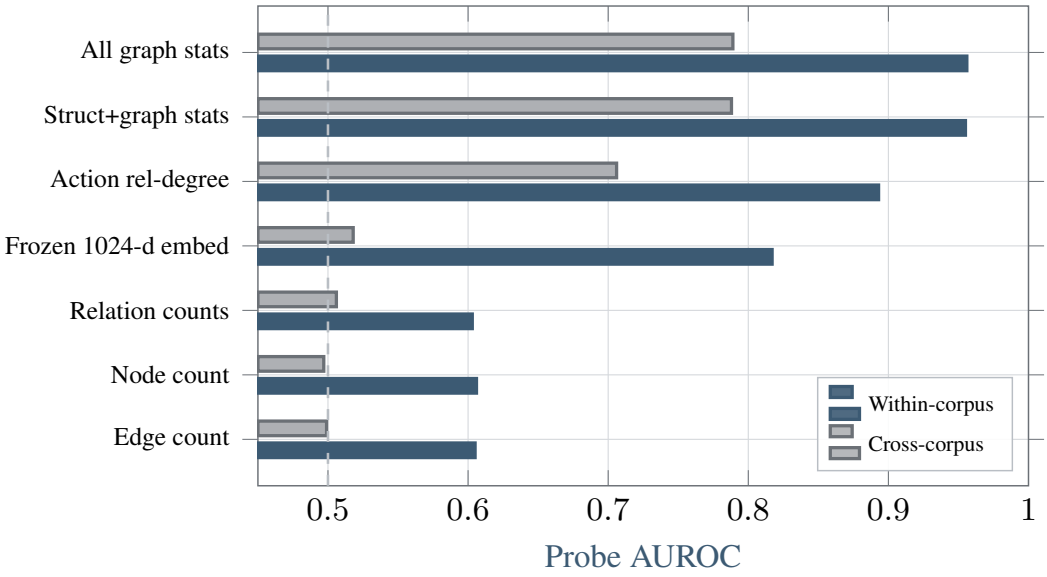



Figure 12: Shortcut-probe AUROC, within- versus cross-corpus (cf. Table 33). Cheap graph-statistic features that look strong within-corpus collapse toward the chance line (dashed, 0.5) when transferred across corpora, which is consistent with them capturing corpus-specific artefacts rather than transferable structure.

| CONDITION | F1 | PREDICTED-BOUNDARY RATE |
|---|---|---|
| Full | 0.430 | 0.299 |
| Edges off | 0.273 | 0.397 |
| Semantics off | 0.081 | 0.029 |
| Both off | 0.314 | 0.497 |

Table 34: **Topic Shift inference-time channel occlusion on DialSeg711.** Because F1 responds to boundary firing rate, the predicted-boundary rate is reported alongside it. The "both off" row illustrates why occlusion effects are not additive causal contributions: with both channels removed, unchanged learned biases cause the classifier to fire on almost half of candidate turns, mechanically raising F1 above the semantics-only ablation despite much poorer segmentation.

assistant models) only. None of the held-out evaluation corpora (ActorAttack, MT-Bench-101, AgentDojo, TIAGE) enters its training pool, so the checkpoint is exposure-free for all three tasks by construction rather than through any per-task filtering. There is no per-task encoder: the jailbreak, Action Guard, and Topic Shift readers all consume graphs induced by this one checkpoint, and each reader's weights are tied to its 1024-dimensional feature space. A valid induced blob reports `span_dim: 1024`, `span_source: joint_live_1024d`, `decode_mode: dense`, and `rel_thresh: 0.5`.

### K.2 SAME-BACKBONE FINE-TUNE BASELINE: PROTOCOL

The natural baseline of Table 2 fine-tunes the full stock `answerdotai/ModernBERT-large` backbone ($\approx$ 400M parameters, all weights unfrozen) end to end on each task, with no induced graph and no frozen state. This is the control a reviewer asks for directly: the identical backbone, trained the way one would train it without our state, scored on the identical held-out transfer set as the frozen readers.

**Shared settings.** Optimizer StableAdamW with weight decay 0.01 (excluded on bias and LayerNorm parameters); learning rate swept over $\{1, 3, 5, 8\} \times 10^{-5}$ with 10% linear warmup followed by linear decay; effective batch size 16–32 via gradient accumulation; a linear or single-hidden-layer MLP head with dropout 0.1; focal cross-entropy loss ($\gamma = 2$, matching the frozen-state readers);

| READER | DIALSEG711 | TIAGE |
|---|---|---|
| Predictive AUROC | $0.770 \pm 0.009$ | $0.612 \pm 0.007$ |
| Retrospective AUROC | 0.896 | 0.641 |
| Representation-blind structural ceiling | 0.511 | 0.517 |

Table 35: **Topic Shift ranking under increasing distribution shift.** Lookahead substantially improves the document-derived validation distribution and provides a smaller benefit on naturalistic TIAGE.

| FEATURE | DIALSEG711 | TIAGE |
|---|---|---|
| Turn position | 0.479 | 0.404 |
| Node count | 0.463 | 0.425 |
| Edge count | 0.465 | 0.424 |
| New nodes | 0.511 | 0.517 |
| New edges | 0.505 | 0.450 |
| Edges / node | 0.467 | 0.424 |
| Predictive reader | **0.770** | **0.612** |

Table 36: **Single-feature Topic Shift AUROC.** No tested scalar graph statistic provides substantial discrimination on DialSeg711, so the boundary cue cannot be a graph-shape crutch even in-distribution. The correlation between graph size and predicted boundary probability is negative on every corpus ($-0.31$ for node count on DialSeg711), so structure works mildly against the reader rather than for it. These probes rule out simple size/growth shortcuts; the reader exceeds every structural feature on both corpora.

`bf16` precision. We select the learning rate and epoch count by the *in-distribution* validation metric, then run that single configuration across seeds $\{0, 1, 2\}$ and report the mean; the held-out transfer corpus is scored exactly once and never enters model selection. Confidence intervals use the same conversation- or action-clustered bootstrap as the main results (Appendix J).

**Per-task configuration.** *Multi-turn jailbreak:* the prediction unit is the full conversation trajectory ($\leq$ 8192 tokens); trained on Crescendo attack and benign conversations, validated on held-in Crescendo, transferred to ActorAttack vs. MT-Bench-101; 3–5 epochs with early stopping on validation AUROC; best learning rate $\approx 3 \times 10^{-5}$; metrics AUROC and TPR at $1\%/5\%$ FPR. *Agent action authorization:* the unit is a single proposed action, encoded as `[context prefix] [SEP] [proposed action]` ($\leq$ 2048 tokens); trained on the InjecAgent sidecar (4,216 actions, balanced 2,108 authorized / 2,108 unauthorized), validated in-distribution on InjecAgent, transferred to AgentDojo plus the authorized-action dump (42,487 held-out actions); 3–6 epochs with early stopping on validation F1; best learning rate $\approx 2 \times 10^{-5}$; metrics AUROC, ASR, F1, and FPR. *Topic-shift detection:* per-turn boundary prediction in the causal regime, scoring turn $k$ from the dialogue prefix (turns $1..k$, $\leq$ 8192 tokens, no future turns); trained on SuperDialseg (9,592 dialogues), validated on DialSeg711, transferred to TIAGE; 3–5 epochs with early stopping on validation boundary F1; best learning rate $\approx 3 \times 10^{-5}$; the streaming top-$q$ rate ($q = 0.2$) is selected on DialSeg711 and applied unchanged to TIAGE. We report the validation-selected boundary F1, matching the reporting convention for the frozen reader (Appendix I.4).

**Undertraining is ruled out.** On agent action authorization the fine-tune converges to a low training loss (of the same order as the $10^{-4}$–$10^{-2}$ figures reported for the reader) while it fails to transfer: its cross-corpus AUROC stays near chance, and run live in the same AgentDojo harness (Table 18) it barely reduces targeted ASR relative to no defense ($52.1\% \rightarrow 50.5\%$) while sacrificing benign utility ($77.3\% \rightarrow 71.2\%$). This reflects a genuine generalization failure rather than underfitting; the low learning rate ($2 \times 10^{-5}$) and early stopping keep the run stable. On jailbreak the fine-tune saturates quickly, so we cap at 3–5 epochs with early stopping and obtain 0.78. Averaged over three seeds, the end-to-end fine-tune trails the frozen reusable state on every task (Table 2).

| THRESHOLD | F1 | PRECISION | RECALL |
|---|---|---|---|
| 0.10 | 0.350 | 0.219 | 0.868 |
| 0.15 | **0.362** | 0.236 | 0.779 |
| 0.20 | 0.360 | 0.244 | 0.685 |
| 0.25 | 0.358 | 0.253 | 0.609 |
| 0.30 | 0.349 | 0.260 | 0.531 |
| 0.35 | 0.333 | 0.261 | 0.461 |
| 0.40 | 0.319 | 0.263 | 0.405 |
| 0.45 | 0.313 | 0.272 | 0.370 |
| 0.50 | 0.299 | 0.272 | 0.332 |
| 0.55 | 0.284 | 0.272 | 0.296 |
| 0.60 | 0.265 | 0.272 | 0.258 |
| 0.65 | 0.223 | 0.266 | 0.192 |
| 0.70 | 0.141 | 0.265 | 0.096 |
| 0.75 | 0.006 | 0.192 | 0.003 |
| 0.80 | 0.000 | 0.000 | 0.000 |

Table 37: **Predictive TIAGE global-threshold frontier (a diagnostic, not the headline rule).** This table sweeps the *fixed global-threshold* rule, which the headline does not use; it shows why: a global threshold selected on DialSeg711 validation (0.65) transfers poorly to TIAGE, landing at F1 0.223, because the reader is miscalibrated across the shift. The test-optimal F1 threshold (0.15) buys higher boundary F1 through low precision (0.236) and a high boundary-firing rate, i.e. over-segmentation. The retrospective reader shows the same phenomenon: test-optimal F1 0.370 at threshold 0.15 against validation-selected F1 0.230. The headline predictive operating point (F1 0.318) instead uses the streaming top-$q$ rate rule fixed on DialSeg711 (Section 4), which is robust to this calibration shift; none of the global-threshold points here is a headline candidate.

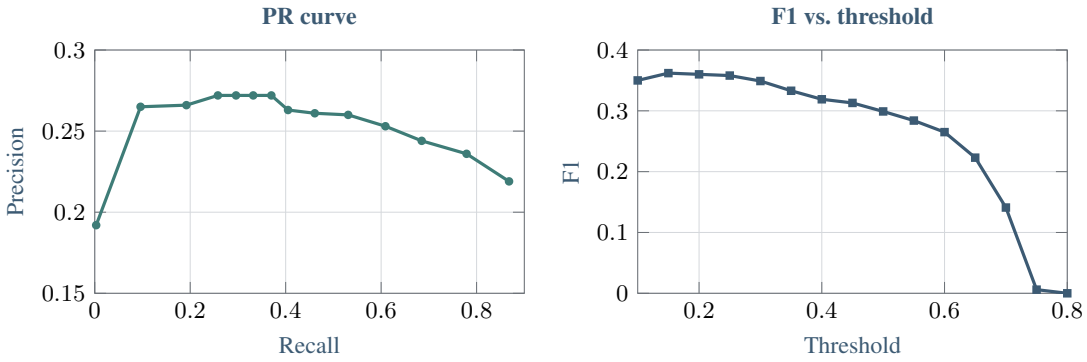



Figure 13: TIAGE global-threshold sweep (cf. Table 37). Left: the precision–recall trade-off; right: F1 as a function of the decision threshold, peaking near 0.15 before falling off as recall is lost at high thresholds.

### K.3 FROZEN DENSE-STATE CONTROL: PROTOCOL AND RESULTS

The same-backbone fine-tune baseline (Appendix K.2) unfreezes the whole encoder, so it confounds two questions: does transfer come from *freezing* a general representation, or from the induced semantic–relational state? This control answers the second. We freeze the encoder, discard mention extraction and graph induction entirely, and train a capacity-matched reader over ordinary dense turn embeddings. The principle is *match downstream capacity; change representation*, not to hobble the non-graph baseline.

**Dense representation.** For turn $x_k$ we run the frozen encoder $H_k = E(x_k) \in \mathbb{R}^{n_k \times 1024}$, mean-pool the non-padding token states to $e_k = \mathrm{MeanPool}(H_k) \in \mathbb{R}^{1024}$, and apply a trainable projection $r_k = \mathrm{MLP}(e_k) \in \mathbb{R}^{128}$. No BIO heads, node extraction, coreference, attachment, relation prediction, or message passing are run. This matches the full reader, which also first maps each node's 1024-d semantic vector to 128 dimensions. We evaluate two frozen encoders $E$: stock

Table 38: **Frozen dense-state control.** Capacity-matched readers over frozen mean-pooled turn embeddings, no induced graph, held-out targets.

| REPRESENTATION | JAILBREAK | ACTION | TOPIC | TOPIC F1 |
|---|---|---|---|---|
| Stock frozen + dense reader | $0.812 \pm 0.01$ | $0.684 \pm 0.01$ | $0.559 \pm 0.01$ | $0.244 \pm 0.02$ |
| Joint encoder frozen + dense reader | $0.841 \pm 0.01$ | $0.764 \pm 0.01$ | $0.582 \pm 0.01$ | $0.276 \pm 0.02$ |
| Full semantic–relational state | **0.894** $\pm 0.006$ | **0.951** $\pm 0.005$ | **0.612** $\pm 0.007$ | **0.318** $\pm 0.010$ |
| Full − Joint Dense | $+0.053$ | $+0.187$ | $+0.030$ | $+0.042$ |

ModernBERT-large (*Frozen-Stock Dense*), and the exact shared encoder inside the final M1+M2 checkpoint with M1/M2 never invoked (*Frozen-Joint Dense*). The latter holds almost everything constant against the full model, same 419M encoder, same upstream training, same 1024-d states, same datasets, and removes only the conversion of token representations into persistent semantic objects and relations.

**Per-task readers.** Each reader mirrors its full-state counterpart's capacity. *Jailbreak reader:* the full model takes three views (mean over accumulated nodes, mean over new nodes, max over accumulated nodes) into a unidirectional GRU; the dense analogue forms $p_k = [\,\mathrm{mean}_{j\le k} r_j;\, r_k;\, \max_{j\le k} r_j\,] \in \mathbb{R}^{384}$ and uses the same $384 \to 128 \to \mathrm{GRU}_{128} \to$ head with identical PRE/SCREEN targets and causal setup. *Topic Shift:* the same $p_k$ feeds a strictly causal uni-GRU emitting $P(\text{boundary at } k)$; no turn $k{+}1$, no bidirectional context, and the DialSeg711-tuned streaming top-$q$ rule is applied to TIAGE unchanged. *Action Guard:* the proposed action is encoded as $a_i = \mathrm{MLP}(\mathrm{MeanPool}(E(\text{action}_i))) \in \mathbb{R}^{128}$; a causal GRU summarizes the prior turns $h_i = \mathrm{GRU}(r_1, \ldots, r_{i-1}) \in \mathbb{R}^{128}$; and $q_i = [\,a_i;\, h_i;\, a_i \odot h_i\,] \in \mathbb{R}^{384}$ feeds the same $384 \to 64 \to 2$ head as the full Action Guard. The GRU is the only added capacity, granted so the dense baseline can remember conversation history at all.

**Training.** Source/validation/held-out splits are those of the headline experiments (Crescendo $\to$ ActorAttack/MT-Bench-101; InjecAgent $\to$ AgentDojo; SuperDialseg $\to$ DialSeg711 validation $\to$ TIAGE), with held-out targets untouched during selection. We use the same focal loss ($\gamma = 2$) and checkpoint-selection metric as the full readers, a learning-rate sweep over $\{1{\times}10^{-4}, 3{\times}10^{-4}, 10^{-3}\}$ selected only on the source/validation split, and seeds $0, 1, 2$; we report the mean and, for the Full−Joint-Dense difference, a paired bootstrap.

**Results.** Table 38 gives the numbers. The jointly trained dense encoder beats the stock one on every task, so upstream training transfers even without the induced state; but the full state adds a further $+0.053$ / $+0.187$ / $+0.030$ AUROC and $+0.042$ F1 over Frozen-Joint Dense. The residual is largest on Action Guard. This does *not* contradict the no-edge drop to $0.528$ (Table 19): that ablation removes edges from the node-based R-GCN reader, whereas the dense control replaces the entire node/edge pipeline with a GRU that supplies its own temporal aggregation. Read together, the two say the graph's value-add over a structure-free reader is $+0.187$ (not the $+0.423$ measured against the node-based no-edge retrain), which remains the largest structured-state residual of the three tasks and is consistent with Action Guard's relational reliance.

### K.4 PAPER-RESULT TO ARTIFACT MAP

Table 39 maps each numerical result reported in the paper to the evaluation artifact and the frozen checkpoint or state required to reproduce it.

## L USE OF LARGE LANGUAGE MODELS

In accordance with the ICLR 2027 policy on the disclosure of AI assistance, we report all uses of large language models (LLMs) and related AI tools in the preparation of this work. The authors take full responsibility for the entire contents of the paper, including any text, code, analysis, or data that was produced with AI assistance; all such material was reviewed, verified, and, where appropriate, corrected by the authors before inclusion.

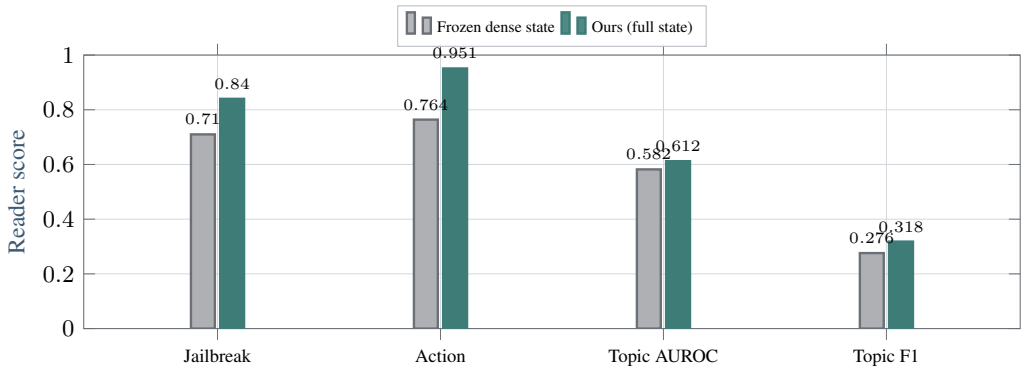



Figure 14: Induced structure versus a capacity-matched, structure-free dense reader on the same frozen encoder (headline numbers, cf. Table 38). The full semantic–relational state retains an advantage on every reader, consistent with the induced graph structure, rather than raw encoder capacity, carrying the transferable signal.

| PAPER RESULT | EVALUATION ARTIFACT | REQUIRED CHECKPOINT / STATE |
|---|---|---|
| Joint corpus 64,888 and M1/M2 intrinsic metrics (Table 3) | `exp1b` (trainer evaluator) | released checkpoint + held-out test partition |
| Jailbreak cross-family 0.894 | `exp-1b-crescendo-clean-encoder.ipynb` | joint checkpoint + `exp1b` jailbreak reader |
| Jailbreak channel occlusion 0.894 / 0.750 / 0.865 | `m3_crescendo_heldout_loadbearing.ipynb` | joint (`exp1b`) state + `exp1b` jailbreak reader |
| Action Guard 0.951 / 0.942 | `m3_action_guard_loco_reproduce.ipynb` | joint (`exp1b`) state + InjecAgent-only LOCO reader |
| Action Guard live-harness ASR 6.1% / utility 74.2% (selected threshold; Table 18) | live same-harness AgentDojo run (Appendix G.11) | joint (`exp1b`) state + InjecAgent-only reader, threshold fixed on InjecAgent validation |
| Action Guard value-add 0.951 vs. 0.528 | `m3_action_guard_edge_ablation_4a_vs_4b.ipynb` | `ag_fullgraph_loco` / `ag_noedge_loco` |
| Action Guard shortcut probes and position matching | `m3_action_guard_loco_reproduce.ipynb` (audit cells) | cross-corpus features only, full held-out set |
| Topic Shift 0.770 / 0.612 | `topic_shift_postproc.ipynb` | joint (`exp1b`) state + predictive `topic_shift` reader |
| Topic Shift boundary F1 0.318 (TIAGE, validation-selected) | `topic_shift_postproc.ipynb` | joint (`exp1b`) state + predictive `topic_shift` reader |
| Topic retrospective 0.896 / 0.641 | `topic_shift_beat_predictive.ipynb` | joint (`exp1b`) state + retrospective `topic_shift` reader |
| Same-backbone FT jailbreak 0.780 (Table 2) | `natural_baseline_crescendo.py` | stock ModernBERT-large, all weights unfrozen; no induced state |
| Same-backbone FT action authorization 50.5% live-harness ASR (Table 18) | `natural_baseline_actionguard.py` | stock ModernBERT-large, all weights unfrozen; no induced state |
| Same-backbone FT topic shift 0.190 | `natural_baseline_topicshift.py` | stock ModernBERT-large, all weights unfrozen; no induced state |
| Frozen dense-state control (Table 38) | `frozen_dense_control_{crescendo,actionguard,topicshift}.py` | frozen stock / joint (`exp1b`) encoder, mean-pooled turns; no induced state |

Table 39: **Mapping from headline paper results to executable artifacts.** Every downstream row uses the same frozen `exp1b` state, induced from the joint corpus of Appendix C.2. A checkpoint and an induced cache from different upstream encoders must never be mixed, and a reader trained with message passing disabled must be scored with it disabled.

**Writing and polishing.** We used LLM-based assistants to edit prose for clarity, concision, and grammar, and to suggest reorganizations of existing text. All scientific claims, framings, and conclusions originate with the authors; AI edits were applied at the level of phrasing and presentation and were checked against the underlying results.

**Retrieval and discovery.** We used LLM-based tools to help locate related work and to surface candidate references during the literature review. Every citation was subsequently verified by the authors against the primary source; no reference was included on the basis of an AI suggestion alone.

**Research ideation and execution.** LLMs were used as a sounding board during ideation (e.g., discussing possible analyses, ablations, and failure modes) and as a coding assistant during the implementation of experiments and analysis scripts. All experimental design decisions, code, and results were authored, run, and validated by the authors.

**Drafting sections.** Initial drafts of some passages, primarily in the appendices and in descriptive portions of the main text, were produced with AI assistance and then substantially revised and fact-checked by the authors. No section was included verbatim from an AI system without author review.

**Synthetic data generation.** Portions of the training and stress-test corpora described in the paper were generated or augmented with LLMs (for example, synthetic multi-turn dialogues and controlled negative examples). The construction, filtering, and validation of these data are described in the relevant sections; the authors are responsible for their quality and for any labels derived from them.

**Uses not applicable.** We did *not* use AI tools to prove mathematical claims, and no use of AI falls outside the categories enumerated above.