Early Detection of Re-Identification Risk in Multi-Turn Dialogues via Entity-Aware Evidence Accumulation

Yeongseop Lee , Seungun Park , Yunsik Son *

Posted Date: 3 March 2026

doi: 10.20944/preprints202603.0209.v1

Keywords: privacy; quasi-identifier; conversational AI; incremental disclosure; entity tracking; confidence gating; provenance; traceability; selective prediction; runtime guardrails

 doi:10.20944/preprints202603.0209.v1

Article

Early Detection of Re-Identification Risk in Multi-Turn Dialogues via Entity-Aware Evidence Accumulation

Yeongseop Lee, Seungun Park, and Yunsik Son *

Department of Computer Science and Artificial Intelligence, Dongguk University, Seoul 04620, Republic of Korea
* Correspondence: sonbug@dongguk.edu

Abstract

In multi-turn conversational AI, individually innocuous personally identifiable information (PII) fragments disclosed across successive turns can accumulate into a re-identification risk that no single utterance reveals on its own. Existing PII detectors operate on isolated utterances and therefore cannot track this cross-turn evidence build-up. We propose a stateful middleware guardrail whose core design principle is *speaker-attributed entity isolation*: every extracted PII fragment is classified by its originating conversational participant (first-person USER vs. incidentally mentioned third parties), and evidence is accumulated in entity-isolated subgraphs that structurally prevent cross-entity contamination. A three-tier extraction pipeline (Tier-0 deterministic regex; Tier-1 Presidio/spaCy NER with zero-shot NER independent verification; Tier-2 independent zero-shot NER; plus rule-based post-processing) refines noisy NER candidates, and an evidence-gated Commit Gate writes only corroborated cues to entity state, firing a re-identification onset signal t_{pred} at the earliest turn where combination-based onset rules grounded in the re-identification uniqueness literature are satisfied. On a 184-record template-synthetic evaluation corpus, the system achieves OW@5 = 70.7% with MAE = 2.442 turns, reducing naïve accumulation MAE by 56% (BL2 MAE = 5.522). We confirm structural robustness on a 300-record mutation stress set and sanity-check RULE_B generalization on the ABCD external corpus (OW@0 = 97.1%, MAE = 0.011). The pipeline requires no modification to the underlying conversational model and serves as a drop-in runtime guardrail for existing dialogue systems.

Keywords: privacy; quasi-identifier; conversational AI; incremental disclosure; entity tracking; confidence gating; provenance; traceability; selective prediction; runtime guardrails

1. Introduction

Autonomous AI agents deployed in customer service and healthcare consultation increasingly operate through long, multi-turn interactions with users. In such settings, users often disclose personally identifiable information (PII) not as a single explicit statement, but as fragmented, low-identifiability cues distributed across turns to facilitate task completion [1–5]. While each cue may appear benign in isolation, their accumulation and combination within the dialogue context can form quasi-identifiers that pose a practical re-identification risk, as discussed in the *k*-anonymity and uniqueness literature. Operationally, we treat a dialogue as a stream of quasi-identifier evidence whose combinations progressively shrink an implicit anonymity set over turns, motivating an onset notion for when cumulative evidence becomes re-identifying.

Despite this risk, widely used PII detection and redaction tools (e.g., Amazon Comprehend PII, Microsoft Presidio [6]) and many commercial guardrail configurations are typically optimized for turn-level processing. These tools and configurations do not, by default, provide cross-turn entity attribution or onset-based cumulative risk assessment. This limitation is particularly critical because meaningful cumulative risk assessment requires attributing extracted cues to the appropriate entity (e.g., the user vs. other individuals mentioned incidentally) over time. As a result, while commercial

 doi:10.20944/preprints202603.0209.v1

agent frameworks are effective at maintaining conversational context, they often lack a dedicated runtime layer for tracking and auditing the cumulative privacy exposure state, creating a common security blind spot in practice.

We propose a stateful middleware pipeline that tracks cumulative re-identification risk in multi-turn dialogue streams and identifies the onset turn t_{pred} with auditable provenance.

Challenges. Three properties of multi-turn dialogue make this task non-trivial: (1) *Fragmentation*—cues for a single individual are scattered across non-adjacent turns; (2) *Cue uncertainty*—off-the-shelf extractors can be inconsistent, making raw cues unreliable for risk decisions without corroboration; (3) *Entity collision*—multiple individuals are discussed in one session, and mis-attribution inflates or masks true risk.

Contributions.

- An *Evidence Graph* that incrementally accumulates entity-level cues across turns and detects the onset t_{pred} via an operational combination rule (Section 3, Section 4);
- A *Commit Gate* that updates entity state only when evidence is corroborated under a conservative commit policy, reducing noise-driven false alarms (Section 4);
- Speaker-attributed entity isolation that maintains separate per-participant state (USER vs. OTHER$_i$), preventing cross-entity risk contamination (Section 4);
- Evaluation on two synthetic corpora and two external benchmarks, demonstrating improved onset accuracy and robustness (Section 5, Section 6).

This work extends our preliminary middleware framework for NER-based PII detection in generative AI prompts [7] and our coreference-resolution pipeline for tracing incremental disclosures across dialogue turns [8] into a unified, stateful onset-detection system with auditable provenance.

Figure 1 illustrates the core phenomenon. Individually innocuous cues disclosed across turns progressively shrink the implicit anonymity set; the system fires the onset signal t_{pred} at the first turn where the committed cue combination satisfies the re-identification threshold.

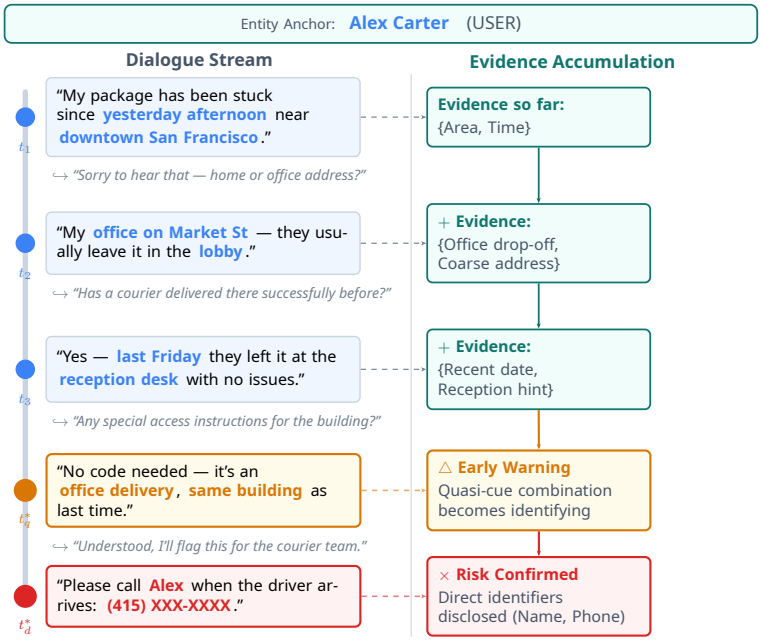

Figure 1. Incremental PII disclosure in a package-delivery support chat: individually harmless fragments can become risky when accumulated, while a direct identifier enables contactability. Here t_q^* denotes the quasi-identifier warning onset (RULE_A) and t_d^* denotes the direct-identifier confirmed onset (RULE_B).

 doi:10.20944/preprints202603.0209.v1

2. Background and Related Work

2.1. Incremental Disclosure and Re-identification Onset

In multi-turn dialogue, users rarely disclose a full identity profile in a single utterance; instead, personally identifying cues are released as low-identifiability fragments distributed across many turns. **Sweeney** [1] showed that {sex, date-of-birth, ZIP code} alone uniquely identify 63–87% of the U.S. population; **Golle** [2] replicated this on updated census data. **De Montjoye et al.** [3] demonstrated that four spatio-temporal points re-identify 95% of 1.5M mobility traces; a follow-up study [4] showed that four credit-card transactions suffice to re-identify 90% of shoppers. **Rocher et al.** [5] estimated that 99.98% of adults are re-identifiable from 15 sparse demographic attributes. These results establish that even a small number of **quasi-identifiers**, once combined, can reduce an individual's anonymity set to near-singleton size.

In a dialogue setting, however, such combinations emerge **incrementally**: each turn may add zero or one new cue, and the transition from safe to risky is rarely signaled explicitly. We refer to this transition point as the **re-identification onset turn**; a formal definition of t_{pred} and the associated combination rules are given in Section 3. For example, a demographic attribute combined with a coarse location cue and an affiliation mention can become identifying once jointly accumulated across turns.

2.2. Turn-Level PII Detection and Redaction

A variety of tools perform PII detection and redaction on individual text segments. **Microsoft Presidio** [6] provides a modular framework combining regex patterns with spaCy named entity recognition (NER) [9]. **Amazon Comprehend PII** [10] offers a managed API for entity-level PII classification and redaction. **GLiNER** [11] provides label-prompted zero-shot NER, making it suitable as a flexible upstream extractor; its multi-task extension [12] adds simultaneous typing and relation extraction. More broadly, transformer-based language models have been successfully applied to reduce false positives in automated security analysis [13], and deep-learning techniques have similarly been employed to evaluate the vulnerability of hiding mechanisms in cyber-physical systems [14], underscoring the growing role of NLP and deep learning in safety-critical detection pipelines.

All of these tools are designed for **turn-level** (or document-level) processing: each input is analyzed independently, and the output is a set of labeled spans. They do not, by default, maintain cross-turn state, attribute detected cues to specific conversational entities, or evaluate whether newly detected cues change a cumulative risk profile. **We do not aim to improve detector accuracy**; we treat existing extractors as interchangeable upstream components and build a stateful monitoring layer on top.

2.3. Entity Attribution in Dialogue

Accurate risk monitoring requires knowing *whose* information is being disclosed. In multi-turn dialogue, a single session may reference multiple individuals: the first-person user, family members, colleagues, or public figures. Misattributing a cue to the wrong entity inflates or masks its true risk profile—the **entity collision** problem identified in Section 1.

Speaker attribution is partially addressed by dialogue act and role tagging in task-oriented systems, where system vs. user turns are structurally distinguished. However, third-party mentions within a single user turn ("my wife works at Samsung") require finer-grained resolution. **Coreference resolution** [15] provides a general mechanism for linking mentions across utterances, but state-of-the-art neural systems require full-document ingestion and are not designed for incremental, turn-by-turn operation. In our prior work [8], we evaluated several coreference models for privacy-oriented dialogue and found that hybrid NER–coreference pipelines can effectively trace incremental disclosures; yet the computational overhead motivated the lightweight entity-isolation approach adopted here.

Our system addresses this gap with a lightweight **entity-isolation** policy: first-person cues are attributed to USER via speaker-role metadata, and third-party mentions are assigned to separate OTHER_i nodes using conservative rule-based linking (last-name match, within-window proximity).

 doi:10.20944/preprints202603.0209.v1

Ambiguous mentions are withheld rather than committed, prioritizing precision over recall in entity attribution. This conservatism is critical because false attribution can trigger an artificially early onset, which is more harmful than a delayed warning in our monitoring setting.

2.4. Stateful Guardrails and Runtime Monitoring

The rapid deployment of agents based on large language models (LLMs) has prompted a growing ecosystem of runtime safety mechanisms. **NeMo Guardrails** [16] provides programmable dialogue rails that intercept and redirect unsafe conversational flows. **Llama Guard** [17] fine-tunes a language model to classify input–output pairs against a safety taxonomy. Open-source frameworks such as **Guardrails AI** [18] validate structured LLM outputs against user-defined schemas. The **OWASP Top 10 for LLM Applications** [19] catalogs risks including prompt injection, data leakage, and excessive agency.

These mechanisms predominantly operate at the **single-turn** or **single-exchange** level: they classify, filter, or block individual inputs or outputs but do not track how information accumulates across turns within a session. In particular, to our knowledge, widely deployed guardrails do not typically maintain an **entity-level evidence state**, apply a commit policy to noisy extractions, or formalize a cumulative onset criterion. **Contextual Integrity** [20] provides a normative framework for appropriate information flow, and **Narayanan and Shmatikov** [21] demonstrated that sparse background knowledge suffices for large-scale de-anonymization; both motivate the need for cross-turn evidence accumulation but do not prescribe a runtime mechanism.

Our pipeline fills this gap by combining entity-isolated evidence accumulation, a confidence-gated commit policy, and an auditable onset criterion that operates alongside existing guardrail and agent frameworks without modifying model weights.

2.5. Summary of Gaps

Table 1 summarizes high-level properties of representative prior work (cross-turn state, runtime intervention, and audit/logging). While some approaches support one or more of these three properties, they do not jointly provide the four capabilities required for entity-attributed onset monitoring.

Table 1. Common methodological properties of representative prior work. ✓ indicates the capability is typically supported as a primary design goal; ✗ indicates it is not typically supported; △ indicates partial or optional support.

Method / Paper	Cross-turn state	Runtime intervention (warn/redact/block)	Audit / logging
MemGPT (B) [22]	✓	✗	△
Recursively Summarizing (B) [23]	✓	✗	✗
NeMo Guardrails (C) [16]	✗	✓	✗
Llama Guard (C) [17]	✗	✓	△
Building Guardrails (C) [24]	✗	✓	△
AgentTrace (D) [25]	✓	✗	✓
Ours	✓	✓	✓

Notes (method families). (B) Memory hygiene / context curation (managing what is stored or fed back across turns); (C) Guardrails (single-turn filtering/blocking); (D) Logging / traceability-first monitoring (observability/auditing).

In summary, existing approaches provide robust context curation, runtime intervention, or session-level logging, but none combines the four differentiating capabilities our pipeline contributes: (i) entity attribution, (ii) onset t_{pred} detection, (iii) confidence-gated state updates, and (iv) auditable per-entity provenance. Table 2 maps each gap to the corresponding mechanism in our system.

 doi:10.20944/preprints202603.0209.v1

Table 2. Design gaps in multi-turn re-identification risk monitoring and our corresponding mechanisms.

Gap	Why prior families miss it	Our mechanism
Entity attribution	No per-entity cumulative state; cue conflation	Entity isolation + linking
Onset t_{pred}	No earliest-satisfaction criterion across turns	Onset rules (Section 4.6)
State-update gating	Extractor noise directly corrupts state	Commit Gate
Onset audit trail	No per-entity provenance to justify t_{pred}	Evidence Graph provenance

Notes. Entity isolation, Commit Gate, and Evidence Graph provenance are defined in Section 4. "Onset audit trail" denotes cross-turn, per-entity provenance sufficient to justify t_{pred} (not merely runtime logs).

3. Threat Model and Problem Formulation

This section formalizes the threat assumptions and the onset task definition used throughout the paper. We present system design in Section 4 and evaluation in Sections 5 and 6.

3.1. Observer Model

Assets. The protected assets are per-entity identity profiles for each conversational participant (USER and zero or more OTHER_i).

Observer. The adversary is a *transcript-only inference observer* who reads dialogue utterances turn by turn and attempts to narrow the anonymity set of a referenced individual to a singleton.

Capabilities. The observer (i) has access to every utterance in the current session, (ii) can aggregate disclosed cues across turns (*cross-turn accumulation*), and (iii) may possess general background knowledge at the level of population distributions or coarse organizational context (e.g., that a given university is located in a particular city), but does not perform person-level lookups.

External linkage. Joining dialogue-derived attributes with external databases, social graphs, or platform registries constitutes a distinct, stronger threat class and is not assumed in this work. We scope the observer to transcript-only inference; external data sources are not part of the assumed observation model. We treat direct identifiers (e.g., phone number, email) as enabling *practical contactability*—the ability to reach the individual through a unique communication channel—rather than full civil-record re-identification, which would require external registry lookups outside our threat scope.

Non-goals. Model-weight attacks, training-data extraction, and direct database intrusion are outside scope. Improving the accuracy of any individual PII detector is likewise a non-goal (consistent with Section 2): we treat upstream extractors as interchangeable components and focus on the cumulative monitoring layer.

Success condition. Re-identification succeeds when accumulated PII fragments for a single entity satisfy a predefined onset rule; the *earliest* such turn is the re-identification onset t_{pred}. The defender's objective is to compute t_{pred} as early and accurately as possible while avoiding premature alerts.

3.2. Task Definition: Re-identification Onset

Entities. A dialogue may reference multiple individuals. We require that each PII cue be attributed to exactly one entity (USER or OTHER_i) to prevent cross-entity risk contamination (the *entity collision* problem noted in Section 1).

Evidence state. Let $\mathcal{F}_t(e)$ denote the set of PII fragments attributed to entity e up to and including turn t under a specified attribution policy and an evidence-admission policy (i.e., which extracted cues are written into state). In Section 4, we operationalize this policy with a conservative commit mechanism; the onset definition itself is independent of any particular admission algorithm.

We partition PII types into three classes: *direct identifiers* (`DIRECT`: EMAIL, PHONE, Social Security number (SSN), ...), *quasi-identifiers* (`QUASI`: OCCUPATION, ORG, LOCATION, AGE, ...), and a *name anchor* (`NAME`: NAME_FULL).

Output contract. The system outputs a single value per entity:

 doi:10.20944/preprints202603.0209.v1

$$t_{\text{pred}}(e) \;=\; \min\{\, t \mid \text{onset\_rule}(\mathcal{F}_t(e)) = 1 \,\} \tag{1}$$

or $\emptyset$ (abstain) if no onset rule fires by the end of the dialogue. Informally, the monitor produces two distinct onset signals: an *IPP-score-based quasi-identifier accumulation signal* (RULE_A) and a *name anchor paired with a direct identifier* (RULE_B). RULE_A fires when the weighted quasi-ID pressure score (Section 4.6) exceeds a calibrated threshold θ; it does not require a name anchor and may operate on an unresolved entity when no name has yet been disclosed. RULE_B requires both a name anchor and a direct identifier: the direct token establishes uniqueness and the name anchor establishes *whose* identity is exposed, completing entity-attributed re-identification. The precise operational form of these signals (score weights and thresholds) is specified in Section 4.6.

3.3. Grounding in the Uniqueness Literature

In the k-anonymity framework [26,27] and its extensions—ℓ-diversity [28] and t-closeness [29]—re-identification occurs when the equivalence-class size $m(Q_t(e))$ for entity e's quasi-identifier set drops below a threshold k_0. In principle, $t^*_{\text{th}} = \min\{\, t \mid m(Q_t(e)) \le k_0 \}$; in practice, $m(\cdot)$ is uncomputable for a transcript-only observer because no reference population is available. The uniqueness literature [1–3] establishes that even a small number of quasi-identifiers in combination can substantially shrink an individual's effective anonymity set— driving linkability risk toward the level of a population singleton without any direct identifier being present. This empirical finding motivates RULE_A as an *identifiability-pressure early-warning signal*: when the IPP-weighted quasi-ID accumulation crosses the threshold θ, the quasi-attribute combination signals non-trivial re-identification risk regardless of whether a name anchor or direct identifier has yet appeared. RULE_B addresses the complementary case: when a name anchor and a direct identifier co-occur, the direct token establishes uniqueness and the name anchor identifies *whose* identity is exposed, completing confirmed re-identification. Our RULE_A/B instantiation (Section 4.6) is consistent with this empirical guidance while remaining conservative in operation: the system is designed to suppress premature alerts under extraction noise and ambiguous attribution, even at the cost of abstentions.

In Section 4, we instantiate the above task contract with an entity-isolated Evidence Graph and a conservative commit policy.

4. Proposed System

4.1. Overview

We propose a **three-tier hybrid extraction pipeline** combining deterministic pattern matching (Tier-0), a neural weak-candidate generator with zero-shot NER independent verification (Tier-1 + Recheck), and independent zero-shot NER extraction (Tier-2), plus a rule-based post-processor to extract PII fragments from multi-turn dialogue. Figure 2 illustrates the full pipeline.

At a high level, the method has six interacting components: (1) multi-extractor cue refinement (Regex → Presidio → zero-shot NER recheck/independent extraction → rule post-processing → span-priority deduplication), (2) deterministic canonicalization of extracted values, (3) an entity attribution filter that routes each fragment to the correct entity graph by speaker role and suppresses agent-sourced evidence from the user state, (4) entity-isolated state accumulation via an Evidence Graph, (5) an evidence-gated state-update policy (Commit Gate) that promotes only corroborated or sufficiently reliable cues to entity state via a four-policy commit mechanism, and (6) optional dictionary-based anaphora resolution that augments implicit references ("my email", "same as before") with previously committed values. This combination is designed to reduce early false alarms (overshoot) caused by noisy cues. We refer to an *overshoot* as an onset prediction earlier than the ground-truth onset ($t_{\text{pred}} < t^*_{\text{oracle}}$).

Each turn is processed through a five-stage pipeline: *Extract* → *Normalize* → *Link* (attribution filter; optional anaphora augmentation) → *Commit* → *Risk evaluation*. Committed fragments accumulate across turns to build an *evidence graph* $\mathcal{G}_t$ per conversational entity. The **Commit Gate** monitors $\mathcal{G}_t$

 doi:10.20944/preprints202603.0209.v1

and fires the onset signal t_{pred}, computed according to Equation (1), as soon as one of the combination rules defined in Section 4.6 is satisfied.

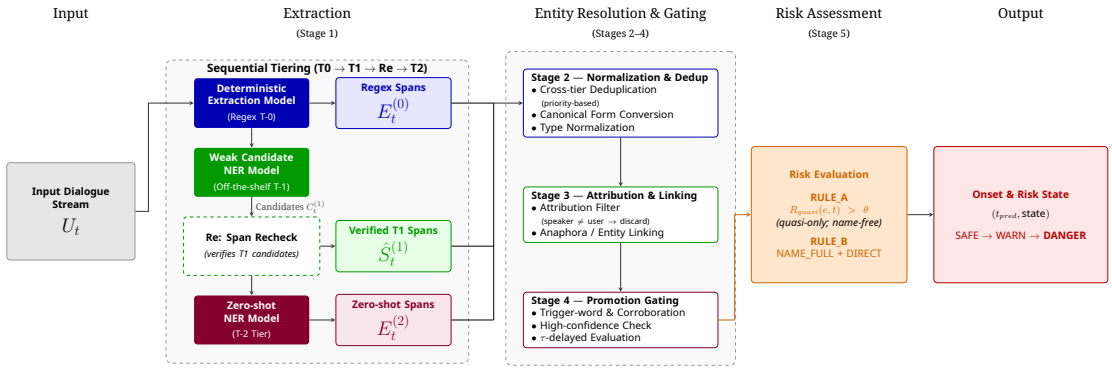

Figure 2. Overview of the proposed multi-turn PII extraction and onset-risk pipeline. Stage 1 applies sequential tiering (T0→T1→Re→T2) to extract span sets, which are consolidated via normalization, attribution/linking, and promotion gating (Stages 2–4). Stage 5 evaluates onset risk (RULE_A/RULE_B) and outputs the onset turn t_{pred} and the final risk state.

4.2. Extraction Pipeline

4.2.1. Tier-0: Deterministic Regex

Tier-0 applies hand-crafted regular expressions to extract syntactically unambiguous structured identifiers. The full Tier-0 type inventory is: *Direct identifiers*: EMAIL, PHONE, SSN, CREDIT_CARD, US_PASSPORT, US_DRIVER_LICENSE; *Quasi-identifiers*: ZIP5 (5-digit U.S. postal code with contextual prefix), DOB (date-of-birth verbal and literal patterns), and GENDER (explicit self-identification phrases). The quasi-identifier patterns (ZIP5, DOB, GENDER) are included because they carry the highest IPP attribute weights (DOB= 3.0, ZIP5= 2.5; Section 4.6). Because all Tier-0 types are syntactically unambiguous, every span receives a fixed confidence of 1.0 and bypasses the zero-shot NER recheck entirely.

4.2.2. Tier-1: Presidio/spaCy-sm (Weak Candidates)

Tier-1 uses Microsoft Presidio [6] backed by spaCy NER [9] to detect NAME_FULL, ORG, and LOC spans. All Tier-1 outputs receive a deliberately *weak* fixed confidence (Table A1), positioning this tier as a **candidate generator** rather than a final classifier. Small-model NER exhibits non-trivial false-positive rates on occupational and organizational titles; the weak weight prevents premature commitment while preserving recall. All spans whose attribute type belongs to $\mathcal{R} = \{\texttt{NAME\_FULL}, \texttt{NAME\_PART}, \texttt{ORG}, \texttt{LOC}\}$ are forwarded to zero-shot NER recheck.

4.2.3. Zero-shot NER Independent Verification (Recheck)

For each Tier-1 candidate in $\mathcal{R}$, we query GLiNER [11,12] with a ten-label prompt covering person, occupation, organization, location, and related attribute types. A candidate is **confirmed** if a GLiNER span overlaps the Tier-1 span by at least one character; its confidence is then upgraded to $\max(c_{\text{presidio}}, c_{\text{gliner}})$, which floors the effective weight at the Presidio default of 0.55 while allowing the GLiNER score to dominate when it is higher. Candidates with no overlapping GLiNER span are **dropped** as false positives. This two-model independent verification design provides stronger false-positive suppression than score thresholding alone.

4.2.4. Tier-2: Zero-shot NER Independent Extraction

In the same forward pass used for the recheck (cached per turn), Tier-2 captures PII types outside the Presidio NER vocabulary: OCCUPATION, ORG_UNIT (department), AFFILIATION, ALIAS, LOCATION, AGE_BAND, and RELATIONSHIP. Non-overlapping Tier-2 spans are added to the fragment list with their raw GLiNER confidence score.

 doi:10.20944/preprints202603.0209.v1

4.2.5. Post-Processor: Rule-Based Heuristics

A lightweight rule pass is applied after merging Tier-0/1/2 fragments. All score adjustments are listed in Table A1.

1. **Subtype demotion**: A one-token NAME_FULL span (no whitespace) is demoted to NAME_PART with reduced confidence, suppressing the common first-name-only false commit. Multi-token all-capitalized spans receive a small boost.
2. **ORG surface verifier**: ORG spans containing a known suffix (Inc., LLC, University, Department, . . .) are boosted; single-token ORG spans with no suffix are penalized.
3. **ALIAS context verifier**: ALIAS spans whose local context contains alias markers (*aka*, *also known as*, *goes by*, . . .) are boosted.

The post-processor modifies only confidence scores and subtype labels; it does not add or remove spans.

4.2.6. Span-Priority Deduplication

After the post-processor, overlapping spans from different tiers are resolved by a strict priority rule: Tier-0 (regex) > Tier-1 (Presidio) > Tier-2 (zero-shot NER). For each pair of overlapping spans, the lower-priority fragment is discarded. This ensures that syntactically confirmed Tier-0 identifiers are never overridden by probabilistic NER outputs.

4.3. Canonicalization

Before fragments enter the Commit Gate, all extracted values undergo deterministic canonicalization to ensure consistent matching and prevent duplicate entity entries. The pipeline applies Unicode NFKC normalization, type-specific case-folding and suffix removal (e.g., stripping corporate suffixes from ORG values), and format normalization for structured types (phone digits, email lowercase).

All downstream comparisons—including entity linking, fact deduplication, and onset-rule evaluation—operate on canonical values. For entity linking, ambiguous matches (e.g., abbreviation or substring overlap) are withheld rather than merged, preventing over-merging of distinct entities.

4.4. Entity Attribution Filter

In multi-turn dialogue, fragments extracted from *agent* turns—job titles, department names, organizational affiliations—can be mistakenly accumulated into the *user* entity state, producing spurious early onsets. We prevent this with a deterministic **entity attribution filter** applied at the Link stage of every turn.

The filter routes each extracted fragment to the entity graph of its *speaker*: fragments with `speaker_id` $\neq$ ‘‘`user`’’ are excluded from the user entity entirely. This speaker-gated linking ensures that agent-sourced Quasi-ID fragments (e.g., the agent's own title or employer) never pollute the user's evidence graph, regardless of surface lexical overlap. Unlike a naive speaker-label filter that only excludes agent turns, the attribution filter is integrated with entity-isolated state accumulation and is applied before evidence admission and onset-rule evaluation.

The attribution filter described above is always active in every system variant reported in this paper. In ablation experiments, we optionally enable an *anaphora-resolution module* (Section 4.5): **Anaphora-T0** denotes the rule-based anaphora tier only, and **Anaphora-T1** denotes Anaphora-T0 plus the neural resolver. The **no-anaphora** configuration disables the anaphora module entirely while keeping the attribution filter enabled. **Naming.** The anaphora tiers are numbered independently of the extraction tiers in Section 4: extraction Tier-0/1/2 refer to the regex, Presidio, and GLiNER stages, whereas Anaphora-T0/T1 refer to the rule-based and neural anaphora resolvers, respectively. We use the “Anaphora” prefix throughout to distinguish these tiers from the extraction tiers and from full-document coreference resolution [15], which our system does not perform.

 doi:10.20944/preprints202603.0209.v1

4.5. Anaphora Resolution (Optional)

Speakers frequently use anaphoric references ("the company", "my email", "same as before") that implicitly refer to previously disclosed PII values. A naïve accumulator that ignores these references may miss evidence that would advance the onset. We provide an optional two-tier anaphora resolver that augments the fragment stream *after* the attribution filter and *before* the Commit Gate.

4.5.1. Tier-0: Rule-Based Anaphora

Tier-0 is a deterministic, dictionary-based resolver that operates over user turns only (agent turns are already filtered by the attribution filter; Section 4.4). It covers 12 anaphora categories—including locative ("that place" → LOCATION), organizational ("my employer" → ORG), and self-referential patterns ("my email" → EMAIL)—each defined by a curated phrase set. Two guard mechanisms suppress false positives: a *negation guard* that detects retraction cues in the local context, and an *update-signal guard* that suppresses re-propagation when a speaker correction is detected. Resolved references produce *virtual fragments* that enter the normal Commit Gate pipeline.

4.5.2. Tier-1: Neural Anaphora Resolver

After Tier-0, we optionally run a lightweight neural coreference model [30] over a sliding window of recent turns augmented with a PII summary anchor that keeps previously committed values within the model's context window.

4.5.3. Hybrid Orchestrator

The hybrid resolver runs Tier-0 first, then invokes Tier-1 only for attributes not yet covered by Tier-0 (lazy evaluation). When both tiers resolve the same attribute, the fragment with higher confidence wins. This design ensures that the deterministic tier handles the common cases (pronoun carry, locative reference) at negligible cost, while the neural tier fills gaps in long-range coreference chains when enabled. We evaluate the anaphora resolver's effect via ablation in Appendix A.2.

4.6. Commit Gate

The Commit Gate operationalizes the onset contract of Section 3.2 through a two-stage process: *Stage 1* promotes raw candidates to committed facts, and *Stage 2* evaluates onset rules over the committed fact set.

4.6.1. Stage 1: Candidate-to-Fact Promotion

An extracted fragment f enters the entity's candidate pool upon first observation at turn t_{obs}. Promotion to a committed fact is governed by four policies, evaluated in the order listed:

1. **Trigger-word commit**: explicit confirmation tokens in the utterance promote the candidate immediately.
2. **Corroboration commit**: agreement across ≥ 2 distinct extractor tiers promotes the candidate.
3. **High-confidence immediate commit**: candidates exceeding a confidence threshold are promoted without delay.
4. **τ-delayed commit**: remaining candidates are held for a fixed number of turns before promotion.

All threshold values are listed in Table A1. A **supersession** mechanism handles corrections: when a new candidate for the same attribute carries sufficient confidence or has waited the delay period, it replaces the existing fact, operationalizing belief revision in the Evidence Graph (Section 4.7).

4.6.2. Stage 2: Onset Rule Evaluation

After each promotion event, the gate evaluates two onset rule families over the effective evidence set for entity e—committed facts $\mathcal{F}_t(e)$ augmented with high-confidence candidates as detailed at the end of this subsection:

 doi:10.20944/preprints202603.0209.v1

$$\textbf{RULE\_A}: \quad R_{\text{quasi}}(e,t) > \theta \tag{2}$$

$$\textbf{RULE\_B}: \ \exists f_1 \in \texttt{NAME},\ f_2 \in \texttt{DIRECT} \text{ s.t. } f_1, f_2 \in \mathcal{F}_t(e) \tag{3}$$

RULE_A operationalizes quasi-based early-warning onset via the Identifiability Pressure Proxy (IPP) score defined in Equation (4): onset fires when the weighted quasi-attribute accumulation exceeds threshold θ. IPP is computed per attributed entity; a name anchor is *not* required—quasi-ID evidence may accumulate under an unresolved (UNK) entity when no name has been disclosed, enabling a name-free, quasi-identifier-based warning [31]. RULE_B fires when a name anchor and a direct identifier (EMAIL, PHONE, SSN) are jointly present; the direct token establishes uniqueness and the name anchor establishes *whose* identity is exposed, completing confirmed re-identification. Here `NAME` denotes a full-name anchor (`NAME_FULL`), and `QUASI` refers to typed quasi-identifiers (e.g., `ORG`, `OCCUPATION`, `LOCATION`, `AGE`). The IPP score is defined as:

$$R_{\text{quasi}}(e,t) = \sum_{a \in \mathcal{A}(e,t)} W_a \cdot r_a + \mu\, \mathbf{1}[d \geq 2] + \nu\, \mathbf{1}[d \geq 3] \tag{4}$$

where $\mathcal{A}(e,t)$ is the set of committed quasi-attribute *(type, value) pairs* for entity e at turn t (when the same attribute type is updated, only the latest value is retained via supersession), W_a is the re-identification weight of attribute type a (a *policy-informed prior* informed by NISTIR 8053 [31] sensitivity categories but not a standardized value; sensitivity to weight choices is reported separately): DOB= 3.0, ZIP5= 2.5, LOCATION= 2.0, OCCUPATION/ORG/SCHOOL= 1.5, AGE= 1.0, GENDER= 0.5, $r_a = r_a(v_a) \in [0,1]$ is a *persona/corpus-conditioned rarity proxy* for the most recently committed value v_a of attribute type a (formal definition below), $d = |\mathcal{A}(e,t)|$ is the attribute diversity (number of distinct committed types), and μ=0.50, ν=1.00 are diversity bonuses. Onset fires when $R_{\text{quasi}} > \theta$, calibrated at $\theta = 4.5$ via record-type stratification within the calibration partition: the in-scope records are partitioned into quasi-first direct-present records (D_{pos}) and direct-dominant records (D_{neg}), and θ is chosen to maximize recall on D_{pos} subject to a $\leq 10\%$ false-positive rate on D_{neg}, where a false positive is an IPP alarm before the direct identifier appears; ties favor the smaller θ. Because θ is a single scalar threshold (not a learned parameter vector) and the weakest satisfying combination (LOC+OCC+AGE) scores ≈ 5.2, leaving a 0.7-point margin above θ, the calibration is empirically stable under moderate perturbations of the calibration partition. The rarity proxy $r_a(v)$ is defined as:

$$p_a(v) = \frac{\text{count}_a(v) + \alpha}{N_a + \alpha |V_a|}, \qquad \tilde{r}_a(v) = -\log p_a(v), \qquad r_a(v) = \min\left(1,\ \frac{\tilde{r}_a(v)}{\tilde{r}_a^{\max}}\right) \tag{5}$$

where $N_a = \sum_v \text{count}_a(v)$, V_a is the post-normalization unique value set from the development persona pool (canonical values per pipeline Stage 2; Section 4.3), α=1.0 (Laplace smoothing), and $\tilde{r}_a^{\max}$ caps the score at the observed pool maximum.

Before rule matching, attribute names are alias-normalized to ensure that extraction-tier naming variations do not prevent rule satisfaction. Specifically, the high-confidence candidates referenced above (Table A1) are included in $\mathcal{F}_t(e)$ for rule-matching purposes, reflecting that near-certain evidence can enable re-identification even before formal commitment.

4.6.3. Three-State Risk Model

The gate maintains a three-state risk model per entity: SAFE ($\mathcal{A}(e,t) = \emptyset$: no committed quasi-attributes), WARN ($\mathcal{A}(e,t) \neq \emptyset$ and $R_{\text{quasi}} \leq \theta$: quasi-evidence accumulating but below threshold; the presence or absence of a name anchor does *not* determine this boundary), and DANGER ($R_{\text{quasi}} > \theta$ via RULE_A, or NAME_FULL + DIRECT jointly present via RULE_B). Only the DANGER state triggers the onset signal t_{pred}; WARN is logged as a diagnostic indicator but does not produce a prediction. Note that RULE_B (NAME_FULL + DIRECT) does not require any committed quasi-attribute: if a name

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 March 2026 doi:10.20944/preprints202603.0209.v1

anchor and a direct identifier are committed before any quasi-ID, the entity transitions directly from SAFE to DANGER, bypassing WARN entirely. When RULE_A fires before the Hybrid oracle onset within an in-scope dialogue, the lead-time $\Delta_{\text{lead}} = t^*_{\text{oracle}} - t_{\text{pred}}$ (Section 5.2) quantifies the advance-warning window provided ahead of the oracle onset.

On firing, the gate emits t_{pred} together with an auditable onset record containing the triggering rule, contributing fragment set, and full metadata (tier, origin turn, confidence).

Figure 3 illustrates the three-state Commit Gate evolution for a representative dialogue. At $t = 3$, only NAME_FULL is present (gate locked). At $t = 7$, a third-party PHONE is committed to a separate OTHER entity; the USER gate is entirely unaffected (entity-isolation guarantee). At $t = 11$, OCCUPATION, ORG, and LOCATION are committed to USER; RULE_A fires and the system outputs $t_{\text{pred}} = 11$.

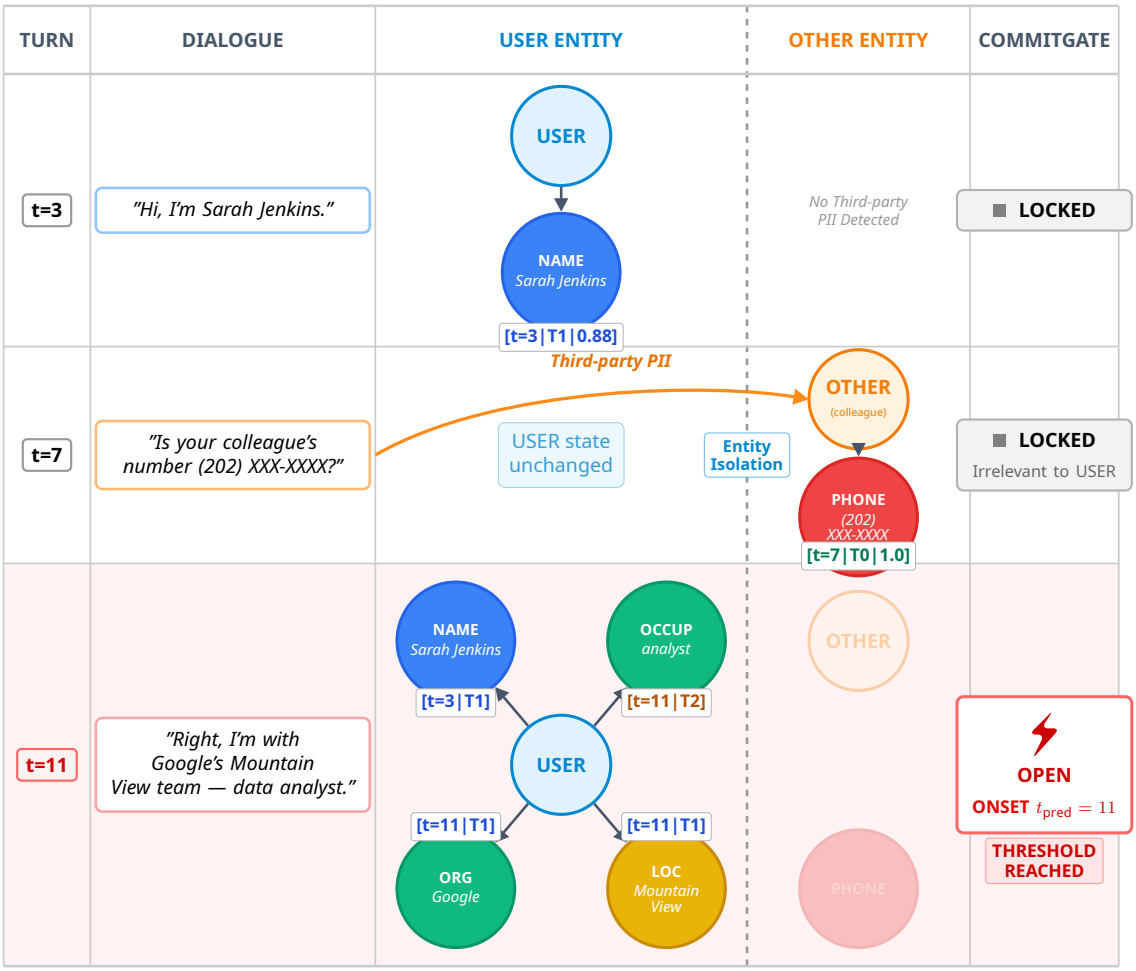

Figure 3. Example of commit-gate state evolution in a multi-turn dialogue. Evidence for the user entity accumulates across turns and is kept isolated from non-user entities by attribution filtering. At turn $t = 11$, sufficient quasi-identifiers are promoted, triggering RULE_A and producing the re-identification onset signal $t_{\text{pred}} = 11$.

4.7. Evidence Graph

The **Evidence Graph** $\mathcal{G}_t = (V, E)$ is the core belief-state data structure. Each conversational entity e_i is a node in V; each committed PII fragment is a leaf node connected by a directed metadata-bearing edge. Each fragment node carries a metadata tuple:

$$m_f = (value,\ t_{\text{commit}},\ tier,\ conf) \tag{6}$$

where t_{commit} is the turn at which the fragment is promoted to a committed fact, $tier \in \{0, 1, 2\}$ encodes extraction provenance, and $conf \in [0, 1]$ is the post-processed confidence. This metadata

 doi:10.20944/preprints202603.0209.v1

supports two key functions: (1) *Exact onset computation*: the gate queries t_{commit} across RULE-satisfying fragments to determine the precise t_{pred}; (2) *Belief revision*: when a speaker corrects a fact, the tuple m_f is overwritten with the newer $(t_{\text{commit}}, conf)$, keeping the belief state current without duplication.

Separating entities into individual subgraphs (USER vs. OTHER$_i$) provides a structural guarantee: PII edges belonging to a third party can never directly trigger the USER's Commit Gate, regardless of how many attributes the third party has disclosed. Figures 4 and 5 show the evidence graph and the PII reveal timeline, respectively, for a five-entity team-introduction dialogue (28 turns).

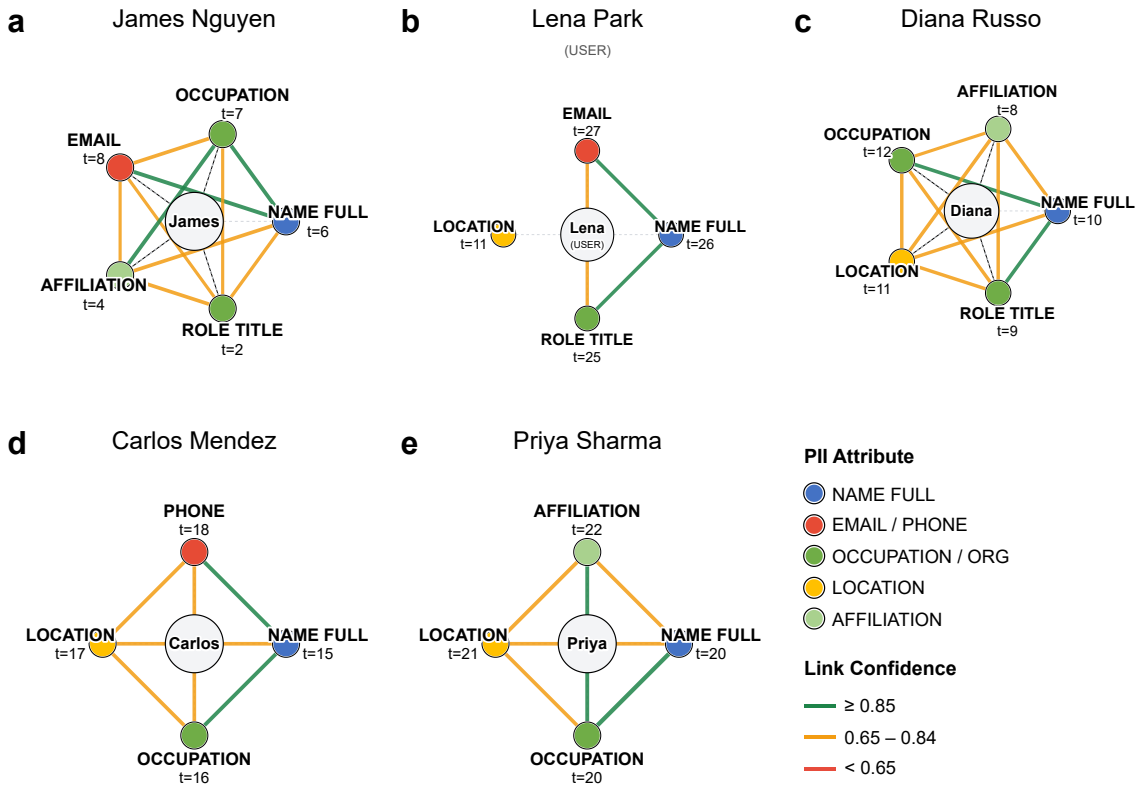

Figure 4. Evidence graph (entity PII subgraphs) for a five-entity team-introduction dialogue (28 turns). Each panel shows the radial PII subgraph for one entity; edges are colored by link confidence ℓ (green ≥ 0.85, yellow 0.65–0.84, red < 0.65).

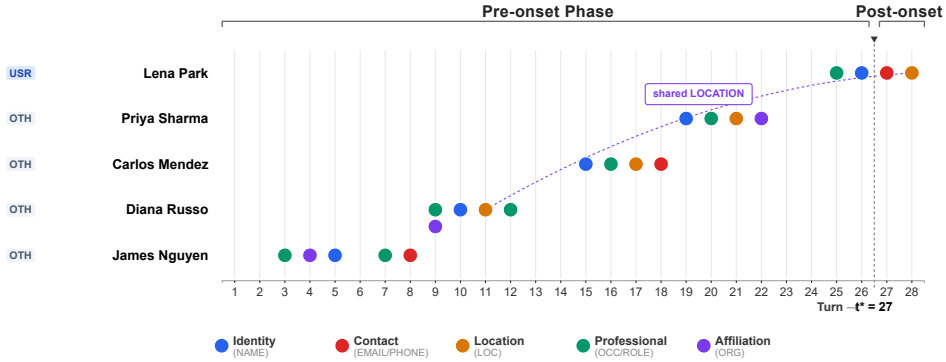

Figure 5. PII reveal timeline and cross-entity links for the same five-entity dialogue. Each dot marks the turn at which a PII attribute was first disclosed; the dashed arc marks a cross-entity shared LOCATION co-occurrence. The vertical dashed line marks the onset $t^*_{\text{oracle}} = 27$.

 doi:10.20944/preprints202603.0209.v1

4.7.1. Intra-entity Link Confidence

For qualitative analysis, each intra-entity PII pair is annotated with a link confidence score based on turn distance and type compatibility. Pairs appearing in the same turn receive the highest score; confidence decays with increasing turn separation. Link scores are used only for visualization and case studies; they are **not** used as input to the Commit Gate.

4.8. Wait-for-B: Selective Prediction via Evidence-Type Gating

A frequent failure mode in multi-turn monitoring is *A-only* early commitment, where quasi-identifier evidence triggers onset without any subsequent direct identifier. We introduce the **Wait-for-B** abstention policy to mitigate this failure mode by gating commitments on evidence type. Wait-for-B is an optional policy layer and is **disabled** in all primary results reported in Section 6; its timing–coverage trade-off is analyzed separately in Section 7.7.

When RULE_A fires on an *A-only* configuration, instead of committing immediately the system stores the pending trigger turn $\hat{t}_A$ and continues observing. If a direct identifier that completes the RULE_B condition (given the existing name anchor) is subsequently committed at turn t_B, the system retrospectively confirms the onset at $t_{\text{pred}} = \hat{t}_A$. If the dialogue terminates without such a direct identifier, the system abstains (returns $t_{\text{pred}} = \emptyset$).

This is a **prospective, online** policy: no lookahead is required, decisions are updated incrementally, and the zero-early-prediction guarantee is structural rather than probabilistic. The policy is an instance of evidence-gated selective prediction: commitment is withheld until a minimum evidence-type requirement (presence of ≥ 1 direct-identifier fragment completing RULE_B) is satisfied, distinct from score-thresholding in that the abstention criterion is based on evidence type rather than confidence level. We report quantitative coverage–precision results in Section 6.3.

4.9. Baseline Systems

We include baselines that isolate the effect of (i) cross-turn accumulation and (ii) confidence-gated state updates. Table 3 describes the two baseline systems used for comparison.

Table 3. Baseline system descriptions.

System	Description
BL1 (Single-turn)	Processes each utterance independently; reports onset only when a single turn satisfies RULE_A or RULE_B. Cross-turn evidence is not accumulated.
BL2 (Naive Accum)	Accumulates all extractor outputs without confidence gating (NER outputs are immediately promoted to committed facts); reports onset at the earliest satisfaction of the operational onset rules (RULE_A/B). The entity attribution filter (Section 4.4) is active, as in all system variants.
Ours	Full three-tier pipeline (Tier-0/1/2) with zero-shot NER recheck, post-processor, entity attribution filter, and Commit Gate. Wait-for-B (Section 4.8) disabled; anaphora module disabled.

4.10. Runtime Overhead

To verify that the pipeline operates within interactive-session timescales, we profile the full system (Tier-0 regex + Tier-1 Presidio + Tier-2 GLiNER + Commit Gate; anaphora module disabled) on the full Track-W corpus ($N = 200$ dialogues; first 3 excluded as warm-up; 197 evaluated, mean 25.9 turns/dialogue). All measurements are taken on a desktop workstation with an NVIDIA RTX 3090

 doi:10.20944/preprints202603.0209.v1

GPU. The per-turn latency is **292 ms** (median) / **297 ms** (mean) / **347 ms** (P95), and the end-to-end per-dialogue latency is **6.4 s** (median) / **7.7 s** (mean). The Tier-2 GLiNER [11] neural forward pass accounts for the majority of per-turn cost; Tier-0 regex and Tier-1 Presidio together add <15 ms. Because typical LLM response generation takes 1–5 s per turn, the privacy monitor can run *in parallel* with generation and complete within the LLM latency window, adding negligible additional delay to the dialogue.

5. Datasets and Evaluation Methodology

Research Question

We organize the evaluation around one core research question: *Does the pipeline generalize across template-synthetic dialogues (Track-W, $N=184$), structurally mutated dialogues (Track-S, $N=300$), and real external task-oriented corpora (ABCD [32], $N=279$; MultiDoGO [33], $N=370$)?* As a supplementary analysis, we apply the pipeline to the Contextual Privacy benchmark [34] (Section 7.8) to characterize the definitional gap between re-identification onset and contextual-integrity norm violation; this is not a performance comparison track. Table 4 summarizes the evaluation tracks.

Table 4. Evaluation tracks and dataset overview. N_{in}: in-scope dialogues (onset ground truth available).

Track	Dataset	N_{in}	Origin	Role
W	woz_synthetic	184	template-synth	primary
S	woz_stress	300	mutated-synth	primary
R_A	ABCD test	279	real (external)	ext. (NAME+DIRECT)
R_D	MultiDoGO test	370	real (external)	ext. (NAME+DIRECT)

5.1. Datasets

woz_synthetic (Track-W, $N = 200$).

A synthetic corpus using MultiWOZ [35] and SpokenWOZ [36] as *structural scaffolds only* (no WOZ utterance text is retained; only dialogue length and speaker-alternation structure are borrowed). Utterance content is replaced with domain templates from five service scenarios (customer service, tech support, medical intake, HR onboarding, insurance inquiry). PII profiles are sampled from curated entity pools spanning U.S. cities, occupational titles, universities, and phone area codes. Following the privacy-preserving synthetic data philosophy advocated in related work [37,38], all PII values are purely fictitious and no real personal data appear in the corpus. The corpus is balanced across 100 NAME+DIRECT records (full name and phone/email both present; Hybrid oracle $= \max(t_{\text{name}}, t_{\text{direct}})$) and 100 quasi-only records (no name or direct identifier; three trigger quasi-identifiers injected; Hybrid oracle $= t_{\text{quasi\_complete}}$, the turn at which all three trigger attributes are committed). 184 of 200 records are in-scope (16 near-miss—dialogues in which the intended trigger PII is never fully disclosed, yielding $t^*_{\text{oracle}} = \emptyset$—are excluded), comprising 89 NAME+DIRECT and 95 quasi-only dialogues; all in-scope records are scored under the **Hybrid oracle** (Section 5.2).

woz_stress (Track-S, $N = 300$).

A robustness diagnostic set applying three mutation operators to Rule×Length-stratified seeds from woz_synthetic. Table 5 details the three operators.

 doi:10.20944/preprints202603.0209.v1

Table 5. Track-S mutation operators. Bucket A preserves t^*_{oracle}; Bucket B shifts t^*_{oracle} and provides relabeled evidence.

Op.	n	**Name**	t^*_{oracle} **effect**	**Bucket**
C1	110	order_swap	unchanged	A
C2	150	dist+k	$+k$ filler turns	B
C3	40	cross_entity	+2 confuser cues	B

C1 tests Commit Gate order-invariance; C2/C3 inject temporal displacement and entity confusion. Bucket B cases are retained as negative evidence rather than discarded. C2 comprises six sub-variants with $k \in \{6, 8, 10, 12, 15, 20\}$ filler turns (30, 30, 25, 35, 14, 16 records respectively), totaling 150 records; C3 comprises 40 records.

External Evaluation Suite

To address the inherent limitation of evaluating only on self-generated data, we include two external corpora whose dialogue structures and PII distributions are independent of our pipeline design. Table 6 summarizes label availability.

Table 6. External-dataset label availability. t^*_{onset} = re-identification onset ground truth.

Dataset	N_{total}	N_{in}	t^*_{onset}	**Evaluation mode**
ABCD test [32]	1,004	279	auto (Hybrid)	Track-R
MultiDoGO test [33]	2,848	370	auto (Hybrid)	Track-R

Neither corpus provides manually annotated onset labels; t^*_{oracle} is computed automatically via the Hybrid oracle ($\max(t_{\text{name}}, t_{\text{direct}})$, which reduces to this form for NAME+DIRECT records). A dialogue is **in-scope** if and only if both a customer name and at least one direct identifier (email, phone, or SSN) appear in the transcript; all other dialogues are excluded ($t^*_{\text{oracle}} = \emptyset$). Both corpora annotate only name and direct identifiers; no quasi-identifier labels (occupation, organization, etc.) are available. Consequently, all in-scope onsets follow the RULE_B pathway and RULE_A is never exercised.

ABCD (Action-Based Conversations Dataset [32]). A customer-service corpus of 10,042 human–agent dialogues collected by ASAPP, covering order status, shipping, account access, and other retail service intents. We use the official test split (1,004 dialogues); after filtering for in-scope records with at least one identifiable onset, $N = 279$ dialogues remain. The corpus annotates customer name, email, and phone number; no quasi-identifier labels (organization, occupation, location, etc.) are provided. All in-scope onsets therefore follow the RULE_B pathway; RULE_A combinations cannot be evaluated.

MultiDoGO (Multi-Domain Goal-Oriented Dialogues [33]). A multi-domain corpus of 14,215 Wizard-of-Oz dialogues spanning six service domains (airline, fastfood, finance, insurance, media, software), collected by Amazon. We use the test split (2,848 dialogues); 370 are in-scope. Similar to ABCD, the annotated PII slots are limited to name, email, phone, and SSN; quasi-identifier labels are absent. All in-scope onsets follow the RULE_B pathway. The domain diversity (e.g., insurance, finance) nonetheless provides useful coverage of direct-ID surface patterns rarely seen in our synthetic data.

5.2. Ground-Truth Oracle Definition

The system evaluates *both* RULE_A and RULE_B (Equations (2)–(3)) at runtime and reports t_{pred} at whichever rule fires first. To *score* predictions we need a ground-truth onset turn t^*_{oracle}, computed from the raw dialogue transcript independently of the system.

Notation.

To distinguish oracle-level transcript properties from system outputs, we use the following oracle-level turns (all derived from the raw transcript independently of the system):

- t_{name} — first turn containing the user's **full name** in the text;
- t_{direct} — first turn containing a **direct identifier** (EMAIL, PHONE, or SSN);
- $t_{\text{quasi\_complete}}$ — turn at which the third trigger quasi-attribute is injected (quasi-only records; data-intrinsic oracle property);

 doi:10.20944/preprints202603.0209.v1

- $t_{\text{name+direct}} = \max(t_{\text{name}}, t_{\text{direct}})$ — turn when both name and direct identifier co-occur (NAME+DIRECT records).

System output is denoted t_{pred}; oracle onset is t^*_{oracle}. These are *never* used interchangeably.

Hybrid oracle (primary).

Oracle onset is computed per record type, reflecting the qualitatively distinct re-identification events each represents:

$$t^*_{\text{oracle}}(\text{Hybrid}) = \begin{cases} \max(t_{\text{name}}, t_{\text{direct}}) & \text{NAME+DIRECT records} \\ t_{\text{quasi\_complete}} & \text{quasi-only records} \end{cases} \tag{7}$$

For NAME+DIRECT records, onset is the turn when both a full name and a direct identifier (EMAIL, PHONE, or SSN) have appeared—confirming contactability. For quasi-only records, onset is $t_{\text{quasi\_complete}}$, the turn at which the third trigger quasi-attribute is committed—confirming identifiability-pressure completion. **In-scope condition**: t^*_{oracle} is defined (equivalently, `t_oracle_onset is not null` in the data): 11 near-miss NAME+DIRECT records and 5 near-miss quasi-only records are excluded, leaving $N_{\text{in}} = 184$. The two oracle cases correspond to different risk events (contactability vs. identifiability-pressure completion); stratified results by record type supplement overall figures.

OR criterion (sensitivity analysis only).

$$t^*_{\text{oracle}}(\text{OR}) = \min(t_{\text{quasi\_complete}}, t_{\text{name+direct}}) \tag{8}$$

where $t_{\text{name+direct}} = \max(t_{\text{name}}, t_{\text{direct}})$ and $t_{\text{quasi\_complete}}$ is the trigger quasi-ID completion turn. Onset is the earlier of quasi-ID completion or NAME+DIRECT confirmation; used for sensitivity analysis only.

Lead-time (derived metric).

When RULE_A fires before the Hybrid oracle onset on an in-scope dialogue, the system provides advance warning. We define **lead-time** as:

$$\Delta_{\text{lead}}(i) = t^*_{\text{oracle},i}(\text{Hybrid}) - t_{\text{pred},i} \qquad (\Delta_{\text{lead}} > 0 \text{ when system fires before oracle}) \tag{9}$$

Under the OR criterion, Category I cases (quasi-first direct-present records where $t_{\text{quasi\_complete}} \leq t_{\text{pred}} < t_{\text{direct}}$) resolve to $\Delta t \approx 0$; the OR–Hybrid MAE gap reflects their mean lead-time (Section 7.7).

All primary results (Section 6) use the **Hybrid oracle**. A concrete example illustrating the Hybrid/OR divergence, an OR-criterion sensitivity analysis, and policy recommendations are provided in Section 7.3.

5.3. Evaluation Metrics

OW@k (Overall Within-k).

OW@k $= |\{i : |t_{\text{pred},i} - t^*_{\text{oracle},i}| \leq k\}| \;/\; N_{\text{in}}$, where the denominator N_{in} includes *all* in-scope dialogues: abstentions always count as errors against every k threshold (Track-W: $N_{\text{in}} = 184$). Primary metric: OW@5.

SW@k (Selective Within-k).

Identical to OW@k but restricted to non-abstaining predictions (denominator $= N_{\text{pred}}$). Measures accuracy conditional on the system making a prediction. SW@$k \geq$ OW@k by construction; the gap reflects the abstention rate.

 doi:10.20944/preprints202603.0209.v1

MAE

Mean absolute error $|t_{\text{pred},i} - t^*_{\text{oracle},i}|$ computed over non-abstaining cases only. Abstentions are excluded from the MAE numerator but penalized through the OW@k denominator, preventing a system from inflating MAE by selectively abstaining on hard cases.

AURC

Area under the risk–coverage curve, computed via a $|\delta|$-based confidence proxy (where $\delta_i = t_{\text{pred},i} - t^*_{\text{oracle},i}$) and trapezoidal integration [39]. Lower AURC indicates better selective prediction quality. A system that universally abstains achieves AURC $= 0$ at coverage $= 0\%$ (trivial-rejector failure mode [39]), which is uninformative. AURC values are provided in the companion repository; the SW@k–OW@k gap in the main tables serves as the primary coverage–accuracy characterization.

F1@k (Timing-Aware F1)

We additionally compute F1@k by combining timing-aware precision and recall, but omit it from the main tables because it does not materially change the overall conclusions compared to OW@k/SW@k under our evaluation protocol; full F1@k tables are provided in Appendix A.3.

5.4. Statistical Protocol and Reproducibility

Bootstrap Confidence Intervals

All CIs use $B = 2{,}000$ percentile bootstrap resamples. For the primary tracks (W, S, R) we report point estimates due to scale; paired bootstrap CIs are available from the authors upon reasonable request.

Reproducibility

All random seeds are fixed and the full pipeline is deterministic given a fixed model snapshot. Generation scripts and reproduction instructions are provided in the companion repository.

6. Results

6.1. Track-W: Template-Synthetic Evaluation

Table 7 reports detection performance on the Track-W split ($N_{\text{in}} = 184$; 16 near-miss records excluded, $\bar{t}_{\text{oracle}} = 19.02$ turns).

Table 7. Track-W detection results on *woz_synthetic* ($N = 184$ in-scope). OW@k = Overall Within-k (%); SW@k = Selective Within-k (%); MAE = Mean Absolute Error (turns); Cov = Coverage (%); early/exact/late = turn-accuracy counts.

System	Cov	OW@0	OW@1	OW@3	OW@5	SW@0	SW@5	MAE	early	exact	late
BL1 (Single-turn)	63.0	60.9	60.9	63.0	63.0	96.6	100.0	0.086	4	112	0
BL2 (Naive Accum)	85.3	42.9	46.7	57.1	65.8	50.3	77.1	5.522	74	82	1
Ours	**79.9**	**54.4**	**55.4**	**66.3**	**70.7**	**68.0**	**88.4**	**2.442**	42	100	5

All rows include the entity attribution filter (Section 4.4). Anaphora ablation shows no measurable effect (Table A2, Appendix A.2).

Ours achieves OW@5 $= 70.7\%$ and MAE $= 2.442$ turns, reducing MAE relative to the naive-accumulation baseline BL2 (5.522) by 56%.

Table 8 stratifies these results by record type. BL1 achieves OW@5 $= 87.6\%$ on NAME+DIRECT records—where a single turn containing both name and direct identifier satisfies RULE_B immediately—but covers only 40.0% of quasi-only records, where cross-turn quasi-ID accumulation is required. Ours improves quasi-only coverage to 72.6% via IPP score accumulation and reduces NAME+DIRECT MAE to 1.513 relative to BL2 (4.469), confirming that the Commit Gate delay is not costly on the RULE_B pathway.

 doi:10.20944/preprints202603.0209.v1

Table 8. Track-W results stratified by record type (NAME+DIRECT: $n = 89$; quasi-only: $n = 95$). The two record types exercise distinct detection pathways (RULE_B vs. RULE_A); Table 7 reports their weight-averaged combination.

System	Record type	n	Cov	OW@0	OW@5	MAE
BL1 (Single-turn)	NAME+DIRECT	89	87.6	86.5	87.6	0.038
	quasi-only	95	40.0	36.8	40.0	0.184
BL2 (Naive Accum)	NAME+DIRECT	89	91.0	50.6	76.4	4.469
	quasi-only	95	80.0	35.8	55.8	6.645
Ours	NAME+DIRECT	89	**87.6**	**67.4**	**83.1**	**1.513**
	quasi-only	95	**72.6**	**42.1**	**58.9**	**3.493**

BL1 Cov = 40.0% on quasi-only records reflects the structural limitation of single-turn detection: RULE_A requires three quasi-IDs to be committed within a single turn—an event that occurs in only 40% of quasi-only dialogues. Ours dominates BL2 on MAE across both record types while maintaining comparable or higher coverage.

6.2. Track-S: Structural Stress Robustness

Table 9 reports detection performance on the Track-S split ($N = 300$; $\bar{t}_{\text{oracle}} = 24.25$; 190/300 records are Bucket-B relabeled).

Table 9. Track-S detection results on *woz_stress* ($N = 300$; 110 Bucket-A + 190 Bucket-B).

System	Cov	OW@0	OW@1	OW@3	OW@5	SW@0	SW@5	MAE	early	exact	late
BL1 (Single-turn)	66.7	64.0	64.0	65.3	65.3	96.0	98.0	0.350	8	192	0
BL2 (Naive Accum)	86.3	41.3	43.0	49.0	55.0	47.9	63.7	8.884	129	124	6
Ours	**82.0**	**52.7**	**53.3**	**58.3**	**62.3**	**64.2**	**76.0**	**5.118**	82	158	6

All rows include the entity attribution filter (Section 4.4). Anaphora ablation shows no measurable effect (Table A2, Appendix A.2).

Under structural stress mutations, Ours achieves MAE = 5.118 turns and OW@5 = 62.3%, reducing MAE relative to BL2 (8.884) by 42%. The higher absolute MAE compared with Track-W (2.442) reflects the deliberate severity of Track-S: 63% of records (190/300) undergo Bucket-B mutations that inject filler turns, entity confusers, and attribute scatterers, systematically displacing evidence and inflating oracle onset positions (see C2 discussion in Section 7). Track-W remains the primary benchmark; Track-S demonstrates robustness under adversarial conditions.

Table 10 provides the corresponding stratification by record type. NAME+DIRECT records suffer more under stress—BL2 MAE rises from 4.469 (Track-W) to 7.154 (Track-S)—confirming that filler-turn injection and entity-confuser cues primarily harm evidence timing on the RULE_B pathway. Ours reduces BL2 MAE by 47% on NAME+DIRECT (7.154→3.791) and by 38% on quasi-only (10.797→6.705), while maintaining a quasi-only coverage advantage over BL1 (74.7% vs. 44.0%).

Table 10. Track-S results stratified by record type (NAME+DIRECT: $n = 150$; quasi-only: $n = 150$).

System	Record type	n	Cov	OW@0	OW@5	MAE
BL1 (Single-turn)	NAME+DIRECT	150	89.3	88.0	88.7	0.134
	quasi-only	150	44.0	40.0	42.0	0.788
BL2 (Naive Accum)	NAME+DIRECT	150	90.7	50.7	64.0	7.154
	quasi-only	150	82.0	32.0	46.0	10.797
Ours	NAME+DIRECT	150	**89.3**	**64.7**	**72.7**	**3.791**
	quasi-only	150	**74.7**	**40.7**	**52.0**	**6.705**

BL2 quasi-only MAE (10.797) substantially exceeds NAME+DIRECT MAE (7.154), reflecting deeper overshoot on IPP-path dialogues under stress mutations. The residual gap between Track-W and Track-S quasi-only MAE (3.493 vs. 6.705) is consistent with larger temporal displacements in Bucket-B mutations.

Figure 6 visualizes OW@k as a function of tolerance k for both tracks, illustrating that Ours dominates BL2 at all $k \geq 2$ while BL1 saturates at $k = 3$.

 doi:10.20944/preprints202603.0209.v1

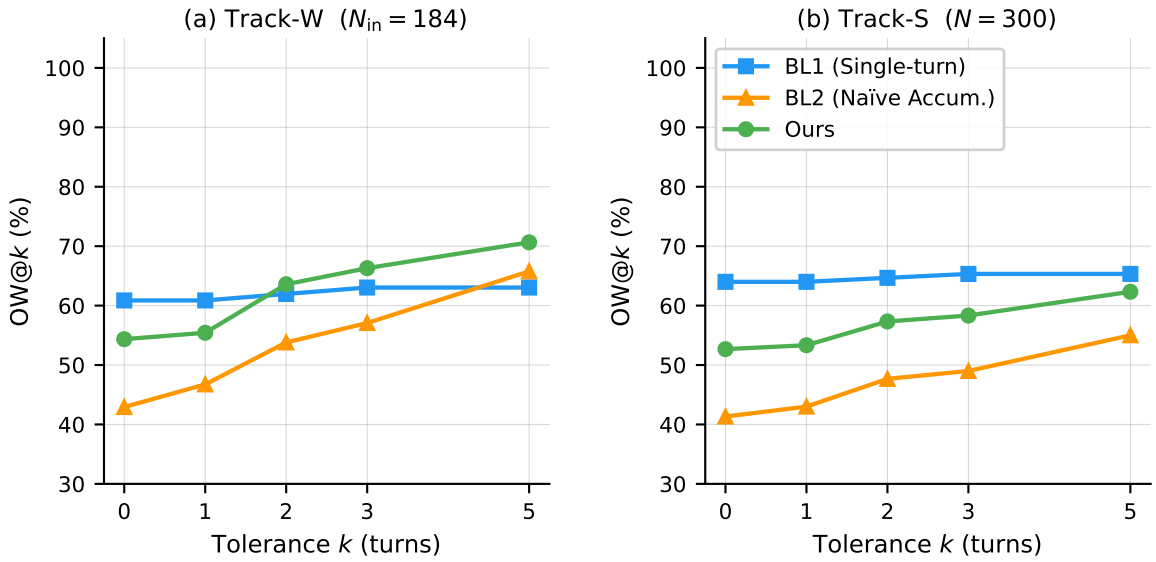

Figure 6. Overall Within-k accuracy (OW@k, %) as a function of tolerance k for (a) Track-W ($N = 184$) and (b) Track-S ($N = 300$). Abstentions count as errors for all k. BL1 (Single-turn) saturates at $k = 3$; BL2 (Naïve Accumulation) achieves higher coverage but large timing spread; Ours achieves the highest OW@5 on both tracks.

6.3. Wait-for-B: Coverage–Precision Trade-Off

Table 11 quantifies the trade-off introduced by enabling the Wait-for-B policy (Section 4.8). The coverage drop from ≈80% to ≈43% is attributable entirely to quasi-only records: Wait-for-B structurally abstains on all quasi-only dialogues because no direct identifier is present to confirm RULE_B. Among NAME+DIRECT records alone, coverage remains 87.6% (Track-W) and 89.3% (Track-S), and every issued prediction is exact (SW@5 = 100%, MAE = 0.000). Operators requiring zero prediction error—e.g., audit-logging contexts where incorrect onset timestamps are inadmissible—can enable Wait-for-B, accepting full abstention on quasi-only records in exchange for an error-free prediction record on NAME+DIRECT dialogues.

Table 11. Effect of enabling Wait-for-B on Track-W and Track-S. Default configuration has Wait-for-B disabled. SW@5 and MAE are computed over non-abstaining predictions only.

Track	Configuration	Cov	OW@0	OW@5	SW@5	MAE
W	**Ours** (default)	79.9	54.4	70.7	88.4	2.442
	+Wait-for-B	42.4	42.4	42.4	**100.0**	**0.000**
S	**Ours** (default)	82.0	52.7	62.3	76.0	5.118
	+Wait-for-B	44.7	44.7	44.7	**100.0**	**0.000**

The ≈37-point coverage drop reflects complete abstention on quasi-only records (Track-W: 95/184; Track-S: 150/300): no direct identifier is present in these dialogues, so Wait-for-B can never receive the RULE_B confirmation signal. On NAME+DIRECT records, coverage is 87.6%/89.3% (Track-W/S) and MAE = 0, matching oracle onset exactly. OW@5 < BL1 overall (42.4% vs. 63.0% on Track-W) because BL1 does predict on some quasi-only records; SW@5 = 100% vs. BL1 SW@5 = 100% are comparable, but the quasi-only coverage gap makes this a different operating point rather than a worse system.

6.4. Track-R: External Sanity Check

As noted in Section 5.1, ABCD and MultiDoGO annotate only name and direct identifiers; all in-scope onsets therefore follow the RULE_B pathway exclusively.

On ABCD ($N_{in} = 279$), the pipeline achieves OW@0 = 97.1%, OW@5 = 98.2%, and MAE = 0.011 turns (Cov = 98.2%; SW@5 = 100%), confirming that RULE_B detection generalizes to real customer-service dialogue without dataset-specific tuning.

On MultiDoGO ($N_{in} = 370$), coverage drops to 8.9% (33 of 370 in-scope records predicted; the remaining 337 return `INSUFFICIENT_EVIDENCE`). Two structural properties of the MultiDoGO corpus

 doi:10.20944/preprints202603.0209.v1

account for this low figure. First, the corpus's crowdsourced annotation pipeline tokenized all email addresses before storage (e.g., `dheepa 23 gmailcom` instead of `dheepa23@gmail.com`); no in-scope record contains an `@` symbol anywhere in the dialogue text. Consequently, 172 of 370 in-scope records (46.5%) are email-only and undetectable under the current Tier-0 regex layer, which requires a standard `user@domain` format; the theoretical coverage ceiling is therefore 53.5%, not 100%. A corpus-adapted fuzzy matcher (e.g., `<word> <digits> <domain>` patterns) could recover a portion of these records; we leave this corpus-specific preprocessing step as future work. Second, among the 198 records with regex-detectable phone numbers, the conversations are short (mean 8.7 turns, user utterances only—agent responses are stripped from the published dataset) and formulaic, providing insufficient cross-turn context for the Commit Gate to accumulate confidence above τ. The 33 successfully predicted records—all phone-based—achieve SW@5 $= 100\%$ and MAE $= 0.273$, confirming that when the Commit Gate does fire, it fires accurately. We therefore treat ABCD as the primary external reference and report MultiDoGO for transparency.

7. Discussion

7.1. Principled Uncertainty Management via Attribution and Commit Gating

A key design principle of our system is that *abstention is not failure*—it is an explicit signal that available evidence is insufficient for a confident commitment. This distinguishes our gated pipeline from naive accumulation (BL2), which treats every NER output as equally reliable and commits immediately. Three categories of evidence from Tracks W and S illustrate why this distinction matters in practice.

Category 1: Name-Collision Abstention (P1 Pattern)

On Track-W, 6 records and on Track-S, 12 records trigger an `NAME_COLLISION` abstention: two distinct entities share the same name, making it structurally impossible to determine which entity is the re-identification target. BL2 *predicts* on all of these—it has no mechanism to detect the ambiguity—producing predictions that may be attributed to the wrong entity. Our system withholds judgment entirely. In an operational deployment, a false re-identification alert on the wrong person is a qualitatively different failure from a missed detection; the pipeline's explicit `NAME_COLLISION` signal (emitted at the attribution layer before onset-rule evaluation) enables downstream handling (e.g., human review) rather than silent propagation of an incorrect attribution.

Category 2: Shared-Attribute Late Prediction (P2 Pattern)

The five late predictions on Track-W (and six on Track-S) all originate from `P2_shared_attribute` records, in which two entities share a common Quasi-ID attribute (e.g., the same employer). BL2 triggers onset early because it accumulates shared attributes without entity isolation; our system correctly routes the shared attribute to the entity graph and awaits disambiguating evidence. The $\Delta t = +3$–$+5$ turn latency is the cost of correct entity attribution: the system is slow precisely because it is careful.

Category 3: Insufficient-Evidence Abstention

On Track-W, 11 records that BL2 predicts are abstained by our pipeline with reason `INSUFFICIENT_EVIDENCE`; on Track-S, 15 such records exist. In 9 of these 11 Track-W cases and 12 of 15 Track-S cases, BL2 exhibits large negative timing offsets (see Tables 7–9 and Table 12), on evidence sets whose per-attribute confidence scores fall below the Commit Gate threshold $\tau = 0.5$. These are structurally *weak* evidence configurations—records involving long-range coreference (P4), indirect expressions (P7), or entity switching (P8) that scatter quasi-ID attributes across many turns with reduced per-mention confidence. This abstention reflects a calibrated judgment that the evidence does not warrant an alert, trading coverage for precision.

 doi:10.20944/preprints202603.0209.v1

Summary

Across Tracks W and S, our pipeline exercises restraint on three structurally distinct grounds: entity-ambiguity abstention (P1), shared-attribute deferral (P2), and sub-threshold confidence abstention. BL2 predicts in all three cases, achieving higher coverage at the cost of alerts that are ambiguously attributed, prematurely committed, or marginally supported. The resulting lower coverage is therefore not a performance deficiency but a *transparency property*: the system communicates the structure of its uncertainty rather than suppressing it.

7.2. Qualitative Case Studies

To ground the quantitative findings, we examine four representative dialogues from Track-W (Table 12) that illustrate the principal error modes. For each case we show the critical dialogue turns, the BL2 commit turn (t_{BL2}), the ground-truth onset turn (t^*_{oracle}), and the pipeline decision.

Case A—P8 (Entity Switch): Wrong-Entity Early Commitment

`woz_synthetic_0064`, *medical_intake*, $t^*_{\text{oracle}} = 11$, BL2 $\Delta t = -1$, Ours $\Delta t = 0$.

T09 [user]: *"I'm an Event Planner based in Toledo, by the way."*
T11 [user]: *"You might have my* ***colleague's*** *info pulled up—please switch to mine:* ***Ryan Chavez****. Email is chavez.r@brownwadea.com."*

BL2 commits after T09, attributing OCCUPATION (*Event Planner*) and LOCATION (*Toledo*) to a profile that belongs to the *colleague*, not to the calling user. T11 reveals the entity switch and delivers the actual target's full name and email. Our pipeline defers commitment until T11, yielding an exact prediction ($\Delta t = 0$). BL2's one-turn advantage in timing corresponds to a *fundamentally incorrect entity attribution*: not a timing error, but a wrong-person alert.

Case B—P1 (Name Collision): Disambiguation Before Commitment

`woz_synthetic_0116`, *hr_onboarding*, $t^*_{\text{oracle}} = 21$, BL2 $\Delta t = -3$, Ours $\Delta t = 0$.

T17 [user]: *"...I work for Turner Ltd, for reference."*
T19 [user]: *"...I'm a Social Media Manager based in Greensboro."*
T21 [user]: *"There might be* ***another Amanda*** *in your system, but I'm* ***Amanda Allen****. My number is 414-708-6688."*

At T17–T19, BL2 has accumulated ORG, OCCUPATION, and LOCATION, committing three turns before the user discloses that a name-collision exists. T21 provides both the collision signal (*"another Amanda"*) and the disambiguating full-name and phone anchor. Our pipeline withholds prediction until T21; BL2's earlier alert would potentially target the wrong Amanda.

Case C—P5 (Revision): Deferral Pending Correction

`woz_synthetic_0129`, *customer_service*, $t^*_{\text{oracle}} = 23$, BL2 $\Delta t = -4$, Ours $\Delta t = 0$.

T19 [user]: *"...I work for Morton, Medina and Webb..."*
T21 [user]: *"...I'm a Database Administrator based in Colorado Springs..."*
T23 [user]: *"My name is* ***Aimee Santos****. My number—wait, I gave you the wrong one earlier. The correct one is* ***402-945-3147****."*

BL2 predicts at T19, before the user's name or phone number has been provided, and four turns before she corrects a previously stated number. Our pipeline awaits a full-name disclosure and resolves the revision, committing exactly at T23. An early BL2 alert here would carry an *uncorrected phone number*.

Case D—P4 (Evidence Sparsity): Abstention on Absent Direct Identifier [*Out-of-Scope Illustrative Case*]

`woz_synthetic_0163`, *hr_onboarding*. No direct identifier present; $t^*_{\text{Hybrid}} = \emptyset$ (Hybrid-oracle out-of-scope and no quasi-ID trigger completed). BL2 $\Delta t = -10$ (relative to OR-oracle $t^*_{\text{OR}} = 33$); Ours: ABSTAIN (`INSUFFICIENT_EVIDENCE`).

 doi:10.20944/preprints202603.0209.v1

T13 [user]: "…*I'm based in* ***Irvine****.*"
T24 [user]: "…*I work as a* ***Mechanical Engineer****.*"
T33 [user]: "…*As I mentioned—I'm* ***54****, a Mechanical Engineer in Irvine.*"

Three Quasi-ID attributes (LOCATION, OCCUPATION, AGE) are dispersed across 20 turns in a 46-turn dialogue. No name or direct identifier (email, phone) appears anywhere. BL2 commits at T23, ten turns early, based on a spurious ORG extraction that BL2 mistakenly co-accumulates with LOCATION and OCCUPATION. The pipeline's per-attribute confidence scores for the extracted quasi-ID fragments do not individually reach the commit threshold $\tau = 0.5$—attributable to low-confidence GLiNER outputs on sparse, multi-sentence context—and no cross-tier corroboration event occurs across the 46-turn dialogue, triggering an `INSUFFICIENT_EVIDENCE` abstention. In a real deployment, this abstention prevents a low-confidence alert attributed to a partially-incorrect evidence set.

Table 12. Case study summary: four representative Track-W dialogues. $\Delta t = t_{\text{pred}} - t^*_{\text{oracle}}$; negative = early, positive = late. Each row shows the Commit Gate decision and error type.

Case	Record	Pattern	t^*_{oracle}	BL2 Δt	BL2 error	Ours
A	0064	P8 entity switch	11	-1	wrong entity	$\Delta t = 0$
B	0116	P1 name collision	21	-3	wrong person	$\Delta t = 0$
C	0129	P5 revision	23	-4	stale value	$\Delta t = 0$
D†	0163	P4 sparse evidence	$\emptyset$	$(-10^\star)$	spurious ORG	ABSTAIN

All four dialogues are from Track-W (`woz_synthetic.jsonl`). BL2 error types describe the *qualitative* nature of the incorrect prediction beyond the timing offset $|\Delta t|$. Cases A–C are among the 25 records where OURS predicts exactly ($\Delta t = 0$) but BL2 commits early (mean BL2 $\Delta t = -11.6$ across 25 cases, Track-W). †Case D is an out-of-scope illustrative case (no direct identifier; $t^*_{\text{Hybrid}} = \emptyset$) and is *not* included in the 11 in-scope abstentions in Table 7. $^\star\Delta t$ is relative to the OR-oracle $t^*_{\text{OR}} = 33$ (illustration only). The 11 in-scope `INSUFFICIENT_EVIDENCE` abstentions display mean BL2 $\Delta t = -8.1$ (early 9/11).

Summary of Case Studies

In Cases A–C, BL2's earlier timing does not represent a more accurate prediction—it represents commitment to wrong-entity attribution (A), wrong-person disambiguation (B), or uncorrected values (C), all of which are qualitatively more harmful than a timing offset. Case D demonstrates that abstention on genuinely weak evidence is a *feature*: the system declines to fire when the evidence is structurally insufficient, rather than emitting a low-quality alert that would require downstream correction. These observations reinforce the *transparency property* framing: the system communicates the structure of its uncertainty in forms that enable principled downstream handling.

Figure 7 shows the full distribution of Δt for Ours and BL2 on both tracks, confirming that BL2 skews strongly negative (premature commitment) while Ours concentrates near $\Delta t = 0$.

 doi:10.20944/preprints202603.0209.v1

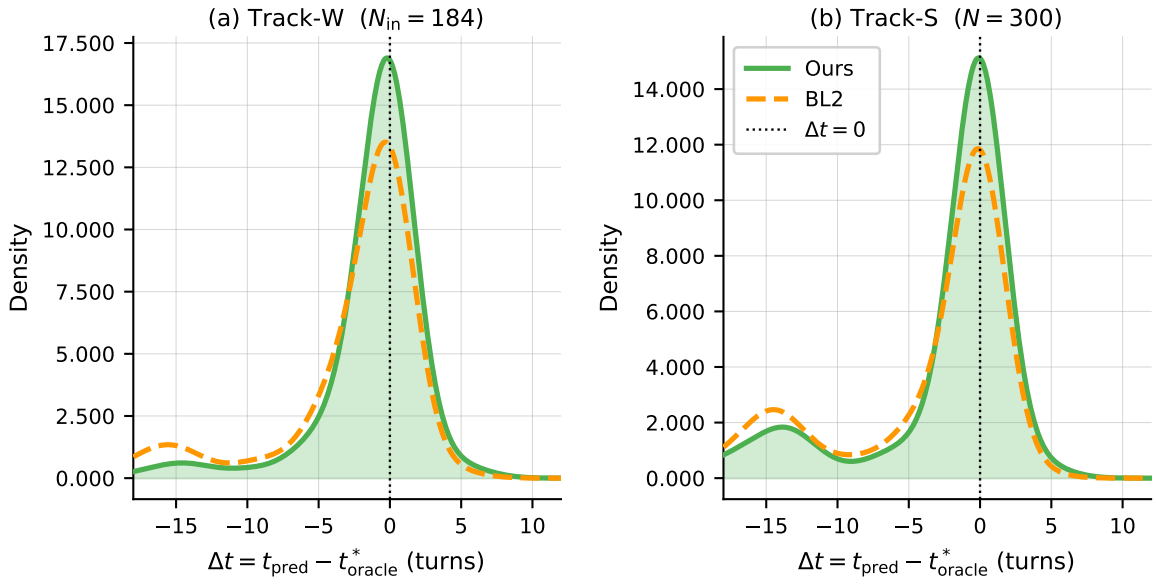

Figure 7. Distribution of prediction offset $\Delta t = t_{\text{pred}} - t^*_{\text{oracle}}$ (turns) for BL2 (Naïve Accumulation) and Ours on (a) Track-W and (b) Track-S (non-abstaining predictions only). Negative values indicate early commitment; zero is exact. BL2 skews strongly negative due to premature commitment; Ours concentrates near $\Delta t = 0$.

7.3. Implications of the Hybrid/OR Evaluation Criterion

Illustrative example.

Consider a dialogue where the user discloses NAME + ORG + OCCUPATION at turn 10 (satisfying RULE_A) and then provides an email at turn 20 (satisfying RULE_B). If the system detects the Quasi-ID combination and predicts $t_{\text{pred}} = 10$:

- **Hybrid**: $t^*_{\text{oracle}} = 20$ (email turn, NAME+DIRECT oracle), so $\Delta t = 10 - 20 = -10$ (10-turn early error);
- **OR**: $t^*_{\text{oracle}} = 10$ (quasi completion turn), so $\Delta t = 10 - 10 = 0$ (exact match).

In Track-W, the 42 early predictions divide into two structurally distinct categories: **(i)** 24 Category I cases—quasi-first direct-present records where $t_{\text{quasi\_complete}} \leq t_{\text{pred}} < t_{\text{direct}}$ (definitional gap resolved by OR oracle); and **(ii)** 18 Category II cases—any record type where $t_{\text{pred}} < t_{\text{quasi\_complete}}$ (genuine advance warning before quasi-ID completion). Under the Hybrid oracle, both categories contribute to MAE (= 2.442); under OR, category (i) resolves to $\Delta t \approx 0$ but category (ii) remains early, yielding MAE(OR) = 0.946.

Sensitivity Analysis

Switching from the Hybrid oracle (Eq. (7)) to the OR criterion (Eq. (8)) reduces MAE substantially on both tracks by eliminating definitional-gap errors (category i), but a portion of early predictions (category ii) survive under OR—reflecting genuine IPP-driven advance warning ahead of the oracle definition (Table A4, Appendix A.4). The residual early predictions under OR are not algorithmic failures; they are **genuine lead-time events** in which the IPP identifies a risk-sufficient quasi-ID combination before all attributes required by the oracle rule are accumulated.

Policy Implications

The choice between AND and OR is a policy decision dependent on the attacker model and regulatory framework. If a user discloses three Quasi-ID attributes (address, occupation, date of birth) before any direct identifier appears, has re-identification already occurred? From a k-anonymity perspective, three correlated Quasi-IDs may suffice for population-unique identification [27], making the OR criterion defensible in high-risk domains (healthcare, finance). The Hybrid oracle, by contrast, provides a stronger, legally recognized threshold anchored to direct identifiers (for NAME+DIRECT

 doi:10.20944/preprints202603.0209.v1

records) and to identifiability-pressure completion (for quasi-only records). We recommend that future evaluations report both criteria: Hybrid oracle for conservative compliance assessment, OR for risk-sensitive early-warning systems.

7.4. Commit Gate vs. Naive Accumulation

As reported in Tables 7–9, the Commit Gate substantially reduces MAE on both tracks relative to naive accumulation (BL2). The primary mechanism is *premature-commitment suppression*: without a confidence gate, spurious NER outputs (job titles misclassified as names, abbreviations captured as ORG spans) are immediately accumulated as onset evidence, triggering t_{pred} too early. The two-model independent verification (Tier-1 + zero-shot NER recheck) and subtype-demotion post-processor filter these false positives, producing a reduction in overshoot rather than mere onset delay. Notably, BL2 achieves a higher binary F1 ($F1_{\text{bin}}$) than Ours (0.921 vs. 0.888 on Track-W; Table A3, Appendix A.3) precisely because it never abstains—every in-scope dialogue receives a prediction regardless of evidence quality. However, binary F1 treats a prediction attributed to the wrong entity (Case A), an ambiguous person (Case B), or an uncorrected value (Case C) identically to an exact-match prediction. Once timing tolerance is introduced, the ranking reverses: Ours leads on F1@k for all $k \leq 5$ on both tracks, confirming that Commit Gate selectively suppresses the low-quality predictions that inflate $F1_{\text{bin}}$.

7.5. Abstention Policy

On both tracks the system abstains on roughly one fifth of in-scope dialogues (Track-W: 20.1%; Track-S: 18.0%) (Tables 7–9). This is deliberate: we prioritize zero contamination over complete coverage, motivated by the asymmetric costs of false re-identification signals in operational deployments. At forced full coverage, selective accuracy drops substantially (the SW@k–OW@k gap in Tables 7–9 quantifies this trade-off), confirming that abstention targets genuinely harder cases.

7.6. Anaphora Resolution: Null Effect and Its Implications

Table A2 (Appendix A.2) shows that the optional anaphora resolver produces zero measurable effect on both Track-W and Track-S ($\Delta\text{early} = 0$ across all ablation steps). The entity attribution filter (Section 4.4)—which routes each fragment to the correct entity graph by speaker role—already suppresses the dominant error source: agent-turn Quasi-ID fragments leaking into the user state. The null result separates two concerns that are superficially similar: (i) *agent contamination suppression* is handled entirely by the attribution filter at the Link stage; and (ii) *anaphoric evidence augmentation* (resolving "my email" to a previously disclosed address) does not alter onset timing on either dataset, because the Commit Gate's τ-delayed promotion already captures the underlying facts before anaphoric re-mention. Long-range neural coreference (Tier-1) [15,30] adds no benefit beyond the rule-based dictionary resolver (Tier-0) for this task.

7.7. RULE_A Lead-Time and the Wait-for-B Trade-Off

Early predictions in both tracks divide into two structurally distinct categories under the IPP onset path. **Category I** (definitional-gap early): defined exclusively on **quasi-first direct-present records**, where $t_{\text{quasi\_complete}} \leq t_{\text{pred}} < t_{\text{direct}}$: the system fires after quasi completion but before the direct identifier appears. The Hybrid oracle ($\max(t_{\text{name}}, t_{\text{direct}})$) counts this as early, but OR oracle ($t_{\text{quasi\_complete}}$) resolves these to $\Delta t \approx 0$. Track-W yields 24 such cases; Track-S yields 45. **Category II** (genuine lead-time early): **record-type-agnostic**; defined as $t_{\text{pred}} < t_{\text{quasi\_complete}}$ (system fires before quasi-ID completion, regardless of record type). Category I and II are mutually exclusive: $t_{\text{quasi\_complete}}$ is the boundary. Track-W yields $n = 18$ such cases (mean $\Delta_{\text{lead}} = 6.6$ turns); Track-S yields $n = 37$ (mean $\Delta_{\text{lead}} = 13.7$ turns). These are not algorithmic failures: they are genuine *lead-time events*, in which the IPP correctly identifies a re-identification risk window before the oracle definition's final attribute arrives. The trade-off is a *policy decision*: operators who prioritize confirmed-onset precision

 doi:10.20944/preprints202603.0209.v1

(Hybrid oracle scoring) accept higher MAE; those who prioritize early-warning coverage retain the IPP trigger and measure value via lead-time.

7.8. Re-identification vs. Contextual Integrity

To probe the boundary between our re-identification onset task and norm-based privacy detection, we applied the pipeline to 100 LLM-agent simulation trajectories from the Contextual Privacy benchmark [34] (SALT-NLP, MIT), which annotates contextual-integrity norm-violation labels rather than re-identification onset. Table 13 summarizes the results.

Table 13. Definitional gap analysis on the Contextual Privacy dataset [34] ($N = 100$; 40 leak / 60 no-leak records).

Metric	Value	Notes
Recall (re-ID onset)	60.0%	t^*_{pii} vs. `leak_label`
Precision	42.9%	NAME+EMAIL typically co-occurs
F1	50.0%	
False Alarm Rate	53.3%	FP = contextually appropriate PII
FP: `privacy_violation_rate=0`	100% (32/32)	all FP preserved privacy
FP: `non_shareable_shared=0`	100% (32/32)	no norm violated
Dominant rule (TP)	RULE_B 100%	NAME+EMAIL in turn 1
$\bar{t}^*_{\text{pii}}$ (TP cases)	turn 1.38	early EMAIL co-occurrence
$\bar{t}^*_{\text{norm}}$ (TP cases)	turn 5.96	contextual-integrity violation onset
Δt $(t^*_{\text{pii}} - t^*_{\text{norm}})$	-4.58 turns	re-ID precedes norm violation

The 53.3% false alarm rate is a definitional artifact: re-identification onset (our task) and contextual-integrity norm-violation onset (the benchmark's label) are distinct privacy event types with different turn distributions. All 32 FP cases have `privacy_violation_rate=0` (contextual integrity preserved), confirming they are definitional artifacts, not algorithm failures. The negative Δt indicates re-identification risk emerges on average 4.6 turns before a detectable contextual-integrity norm violation.

Two findings are noteworthy. First, the negative $\Delta t = -4.58$ turns shows that re-identification risk emerges systematically *earlier* than contextual-integrity norm violations: the pipeline detects an identity-anchoring combination (typically NAME + EMAIL in turn 1) before any contextually inappropriate information flow occurs. This suggests that a re-identification monitor could serve as an *early-warning pre-filter* ahead of contextual-integrity monitors [20]. Second, all 32 false positives have `privacy_violation_rate = 0`: these are cases where PII was shared but *contextual integrity was preserved* (e.g., a customer providing their email to a legitimate service agent). This confirms a genuine *definitional gap*—the two tasks (re-identification onset vs. contextual-integrity norm violation) target fundamentally different privacy event types. Any future attempt to unify them would require joint annotation of both onset types within the same corpus.

7.9. Scope Limitations

LOCATION/DATE Granularity

The system currently maps all LOCATION and DATE attributes to QUASI_ID regardless of granularity. HIPAA Safe Harbor §164.514 [31,40] distinguishes LOCATION_FINE (city, ZIP) and FULL_DATE (month+day, birth date) as direct-ID equivalents. Granularity-aware attribute typing is a natural extension path.

UNIQUE_ID Out of Scope

Account numbers, URLs, IP addresses, and device identifiers (HIPAA Safe Harbor items 10–16) are not covered by RULE_A/B and constitute a direct future-work extension.

 doi:10.20944/preprints202603.0209.v1

Pseudonym and Handle

Usernames and display names are excluded from RULE_A/B per the Article 29 Working Party (WP29) Opinion 05/2014 [41] boundary: cross-platform handle resolution requires an external-source attacker outside our threat model.

Dataset Scale

The primary tracks (W, $N = 184$; S, $N = 300$) are template-synthetic and mutation-synthetic respectively. ABCD ($N = 279$) serves as the primary external reference (Section 6.4); MultiDoGO provides domain diversity but achieves only 8.9% prediction coverage. As detailed in Section 6.4, 46.5% of MultiDoGO's in-scope records contain only tokenized email addresses (no @ symbol) that the current Tier-0 regex layer cannot match; a corpus-adapted fuzzy-email rule could recover these records without architectural changes. The remaining phone-bearing records lack the cross-turn conversational context required by the Commit Gate. Larger-scale evaluation on PrivacyBench [42] is planned for future work.

C2 Filler-Turn Displacement

The C2 stress operator inserts k content-free filler turns between existing dialogue turns, displacing the oracle onset by exactly k. For large k ($k \geq 15$), MAE spikes because the system fires at approximately the original evidence-accumulation point while the oracle has shifted forward. This effect is neutralizable by a simple operator-level preprocessing heuristic: filtering empty or content-free turns before pipeline ingestion eliminates the displacement entirely. The confound between k and dialogue length (large k is structurally assigned only to longer dialogues, which independently accumulate more pre-oracle quasi-ID evidence) makes dialogue-length-controlled evaluation of C2 a natural future-work extension.

RULE_A-Only Dialogues Outside Hybrid-Oracle Scope

The Hybrid-oracle design (Section 5.2) defines in-scope as dialogues where t^*_{oracle} is defined: NAME+DIRECT records require at least one direct identifier; quasi-only records require the trigger quasi-ID set to be complete. Dialogues in which RULE_A fires but neither condition is met—the *A-only population*—are excluded from Hybrid-oracle scoring by construction: they carry no $t^*_{\text{oracle}}(\text{Hybrid})$ and are neither scored as correct nor penalized as false positives. In such dialogues the quasi-ID combination constitutes the *entire* re-identification event, and evaluating RULE_A's precision/recall would require OR-oracle labels. Constructing an evaluation corpus with annotated A-only cases—i.e., conversations where quasi-ID accumulation reaches the RULE_A threshold but no direct identifier ever appears—is a prerequisite for directly measuring RULE_A performance as an independent detector, and is left as future work.

No External Corpus for RULE_A Evaluation

To the best of our knowledge, no publicly available dialogue corpus provides turn-level annotations for both *direct identifiers* (e.g., email, phone) *and* multiple *quasi-identifier* types (e.g., occupation, organization, age) within the same conversation. ABCD and MultiDoGO—the two corpora used in Track-R—annotate name plus direct-ID slots only; consequently the RULE_A pathway (IPP quasi-ID accumulation; $R_{\text{quasi}} > \theta$) cannot be exercised on any existing external dataset (Section 5.1). RULE_A is therefore validated exclusively on the synthetic tracks (Track-W, Track-S), where quasi-ID diversity is controlled by design. This annotation gap itself motivates the synthetic-corpus methodology adopted in this work: until real-world dialogue datasets include richer PII-type annotations, synthetic generation remains the only reproducible way to evaluate combination-based re-identification rules.

8. Conclusions

We presented a stateful evidence accumulation pipeline for detecting re-identification onset in multi-turn dialogue. The system's three-tier extraction pipeline, zero-shot NER independent

 doi:10.20944/preprints202603.0209.v1

verification (instantiated with GLiNER), and Commit Gate policy achieve OW@5 = 70.7% with MAE = 2.442 turns on a 184-record synthetic evaluation corpus (Track-W), reducing BL2's MAE by 56%. We confirm structural robustness on a 300-record mutation stress set (Track-S, MAE = 5.118) and externally sanity-check RULE_B generalization on the ABCD corpus (OW@0 = 97.1%, MAE = 0.011).

The system produces auditable evidence graphs with turn-level provenance, enabling transparent diagnosis of re-identification risk in production conversational AI deployments. The combination-based onset definition and the stateful evidence accumulation architecture are designed for operational deployment as a middleware layer in existing dialogue pipelines, requiring no modification to the underlying conversational model.

Future directions include: (1) granularity-aware LOCATION/DATE classification aligned with HIPAA Safe Harbor [40], GDPR Article 4(1) [43], and CCPA § 1798.100 [44]; (2) UNIQUE_ID extension (IP addresses, account numbers); (3) large-scale evaluation on PrivacyBench [42]; and (4) adaptive Commit Gate policies incorporating domain-specific k-anonymity population estimates.

Author Contributions: Conceptualization, Y.L. and Y.S.; methodology, Y.L. and Y.S.; software, Y.L.; validation, Y.L. and S.P.; formal analysis, Y.L. and Y.S.; investigation, Y.L. and S.P.; resources, Y.L.; data curation, Y.L.; writing—original draft preparation, Y.L.; writing—review and editing, Y.L., S.P. and Y.S.; visualization, Y.L.; supervision, Y.S.; project administration, Y.S.; funding acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2026-RS-2020-II201789) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This work was supported by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ICAN(ICT Challenge and Advanced Network of HRD) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2026-RS-2023-00260248). This work was supported by the Commercialization Promotion Agency for R&D Outcomes(COMPA) grant funded by the Korea government(Ministry of Science and ICT) (2710086167). This work was supported by the Commercialization Promotion Agency for R&D Outcomes(COMPA) grant funded by the Korea government(Ministry of Science and ICT) (RS-2025-02412990).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The synthetic evaluation corpora (woz_synthetic and woz_stress) were generated using dialogue scaffolds from MultiWOZ [35] (https://github.com/budzianowski/multiwoz, accessed on 27 February 2026) and SpokenWOZ [36] (https://spokenwoz.github.io/SpokenWOZ-github.io/, accessed on 27 February 2026). The external corpora used for Track-R evaluation are publicly available: ABCD [32] (https://github.com/asappresearch/abcd, accessed on 27 February 2026) and MultiDoGO [33] (https://github.com/awslabs/multi-domain-goal-oriented-dialogues-dataset, accessed on 27 February 2026). All synthetic corpora are fully reproducible from the generation scripts with fixed random seeds.

Acknowledgments: A preliminary version of this work was presented at the 26th International Symposium on Advanced Intelligent Systems (ISIS 2025), Cheongju, Korea, under the title "Coreference Resolution for Privacy Protection: Tracing Incremental Disclosures in Multi-turn Dialogue." This article substantially extends the conference paper through a redesigned pipeline architecture, new evaluation metrics, additional synthetic and external evaluation corpora, and comprehensive sensitivity analysis.

Conflicts of Interest: The authors declare no conflicts of interest.

 doi:10.20944/preprints202603.0209.v1

Abbreviations

PII	Personally Identifiable Information
NER	Named Entity Recognition
LLM	Large Language Model
SSN	Social Security Number
AURC	Area Under the Risk–Coverage Curve
OW@k	Overall Within-k Accuracy
SW@k	Selective Within-k Accuracy
MAE	Mean Absolute Error
CI	Confidence Interval
HIPAA	Health Insurance Portability and Accountability Act
GDPR	General Data Protection Regulation
CCPA	California Consumer Privacy Act
QUASI	Quasi-Identifier
DIRECT	Direct Identifier

Appendix A. Supplementary Tables

Appendix A.1. Extraction Pipeline Constants (Table A1)

Table A1. Extraction pipeline confidence constants and post-processor score adjustments.

Tier / Rule	Parameter	Value	Rationale
Tier-0 (Regex)	Base confidence	1.00	Syntactically unambiguous
Tier-1 (Presidio)	Base confidence	0.55	Weak-candidate generator
Tier-2 (GLiNER)	Base confidence	raw score	Zero-shot NER output
Commit Gate	τ_f (neural)	0.50	Default commit threshold
Commit Gate	τ_f (regex)	1.00	Tier-0 pass-through
Commit Gate	High-conf. immediate	≥ 0.90	Bypass delay
Commit Gate	τ-delay (turns)	3	Pending-candidate wait
Commit Gate	Supersession	≥ 0.80	Replace existing fact
Commit Gate	Candidate rule-match	≥ 0.85	Pre-commit rule participation
PP: Subtype demotion	1-token NAME cap	≤ 0.45	Suppress first-name-only
PP: Subtype demotion	Multi-token caps boost	$+0.05$	Favor full-name spans
PP: ORG verifier	Known-suffix boost	$+0.10$	Inc., LLC, University, …
PP: ORG verifier	No-suffix penalty	-0.15	Single-token ambiguous ORG
PP: ALIAS verifier	Context-match boost	$+0.15$	$\pm$60-char alias marker
IPP quasi-ID attribute weights (Eq. 4)			
DOB	$W_{\texttt{DOB}} = 3.0$	—	Date of birth (NISTIR 8053 high-risk)
ZIP5	$W_{\texttt{ZIP5}} = 2.5$	—	5-digit postal code
LOCATION	$W_{\texttt{LOC}} = 2.0$	—	City / region
OCCUPATION/ ORG/ SCHOOL	$W = 1.5$	—	Professional/ institutional context
AGE	$W_{\texttt{AGE}} = 1.0$	—	Age or age band
GENDER	$W_{\texttt{GEN}} = 0.5$	—	Gender marker

PP = post-processor. All thresholds were set via qualitative error analysis on development dialogues; no grid search was performed.

 doi:10.20944/preprints202603.0209.v1

Appendix A.2. Anaphora Ablation

Table A2. Anaphora resolver ablation on Track-W and Track-S. T0 = rule-based dictionary resolver; T1 = T0 + neural resolver. Δearly = change in early-prediction count.

Track	Configuration	Cov	OW@0	MAE	Δearly
W	Ours (attrib. only)	79.9	54.4	2.442	—
	+Anaphora-T0	79.9	54.4	2.442	0
	+Anaphora-T1	79.9	54.4	2.442	0
S	Ours (attrib. only)	82.0	52.7	5.118	—
	+Anaphora-T0	82.0	52.7	5.118	0
	+Anaphora-T1	82.0	52.7	5.118	0

Zero change on both tracks (Δearly = 0 everywhere); interpretation in Section 7.

Appendix A.3. Timing-Aware F1 Tables

Table A3. Timing-aware F1@k on Track-W ($N = 184$) and Track-S ($N = 300$). F1@k combines precision and recall where a true positive requires both an in-scope prediction and $|\delta t| \leq k$; abstentions count as false negatives. Conclusions are consistent with Tables 7–9.

Track	System	Cov	F1@0	F1@3	F1@5	$F1_{\text{bin}}$
W	BL1 (Single-turn)	63.0	0.747	0.773	0.773	0.773
	BL2 (Naive Accum)	85.3	0.463	0.616	0.710	0.921
	Ours	**79.9**	**0.604**	**0.737**	**0.785**	0.888
S	BL1 (Single-turn)	66.7	0.768	0.784	0.784	0.800
	BL2 (Naive Accum)	86.3	0.444	0.526	0.590	0.927
	Ours	**82.0**	**0.579**	**0.641**	**0.685**	0.901

$F1_{\text{bin}}$ = binary F1 ($k = \infty$, timing ignored). BL2 achieves the highest $F1_{\text{bin}}$ (0.921 / 0.927) despite the lowest F1@0, illustrating that binary detection F1 obscures timing accuracy; see Section 7 for discussion. Minor rank fluctuations (e.g., Track-S F1@3) may occur due to harmonic-mean coupling of coverage and timing accuracy.

Appendix A.4. Hybrid vs. OR Oracle Sensitivity

Table A4. Hybrid vs. OR oracle sensitivity (Track-W and Track-S). Under the OR criterion, Category I early predictions (quasi-first direct-present records with $t_{\text{quasi\_complete}} \leq t_{\text{pred}} < t_{\text{direct}}$) become near-exact matches and MAE drops substantially.

stem	Cov	OW@0 (Hybrid)	OW@0 (OR)	MAE (OR)
2	85.3	42.9	64.1	2.376
ırs	79.9	54.4	67.4	0.946
2	86.3	41.3	65.0	3.803
ırs	82.0	52.7	67.7	2.163

OR oracle reduces MAE substantially by eliminating Category I definitional-gap errors (quasi-first direct-present records where RULE_A fires before the direct identifier). Residual MAE under OR reflects genuine Category II IPP-driven advance warning: the IPP score crosses θ before $t_{\text{quasi\_complete}}$, constituting real lead-time detection rather than algorithmic failure.

References

1. Sweeney, L. Simple Demographics Often Identify People Uniquely. *Carnegie Mellon University, Data Privacy Working Paper* **2000**, *3*, 1–34. Available online: https://dataprivacylab.org/projects/identifiability/paper1.pdf (accessed on 27 February 2026).

 doi:10.20944/preprints202603.0209.v1

2. Golle, P. Revisiting the Uniqueness of Simple Demographics in the US Population. In Proceedings of the 5th ACM Workshop on Privacy in Electronic Society (WPES). ACM, 2006, pp. 77–80. https://doi.org/10.1145/1179601.1179615.
3. de Montjoye, Y.A.; Hidalgo, C.A.; Verleysen, M.; Blondel, V.D. Unique in the Crowd: The Privacy Bounds of Human Mobility. *Scientific Reports* **2013**, *3*, 1376. https://doi.org/10.1038/srep01376.
4. de Montjoye, Y.A.; Radaelli, L.; Singh, V.K.; Pentland, A.S. Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata. *Science* **2015**, *347*, 536–539. https://doi.org/10.1126/science.1256297.
5. Rocher, L.; Hendrickx, J.M.; de Montjoye, Y.A. Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models. *Nature Communications* **2019**, *10*, 3069. https://doi.org/10.1038/s41467-019-10933-3.
6. Microsoft. Microsoft Presidio: Data Protection and Anonymization SDK, 2023. Documentation: https://microsoft.github.io/presidio/.
7. Lee, Y.; Park, S.; Kim, J.; Kim, H.; Son, Y. Framework for Detecting Personally Identifiable Information Risks in Generative AI Prompts via Time-Series Analysis. In Proceedings of the MITA 2025 Conference, 2025. ISSN 1975-4736.
8. Lee, Y.; Son, Y. Coreference Resolution for Privacy Protection: Tracing Incremental Disclosures in Multi-turn Dialogue. In Proceedings of the 26th International Symposium on Advanced Intelligent Systems (ISIS 2025), Cheongju, Korea, November 2025.
9. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. spaCy: Industrial-Strength Natural Language Processing in Python, 2020. https://doi.org/10.5281/zenodo.1212303.
10. Amazon Web Services. Amazon Comprehend PII Detection, 2023. Available online: https://docs.aws.amazon.com/comprehend/latest/dg/how-pii.html (accessed on 27 February 2026).
11. Zaratiana, U.; Tomeh, N.; Holat, P.; Charnois, T. GLiNER: Generalist Model for Named Entity Recognition Using Bidirectional Transformer. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2024. https://doi.org/10.18653/v1/2024.naacl-long.300.
12. Knowledgator. GLiNER2: An Efficient Multi-Task Information Extraction System with Schema-Driven Interface, 2025, [2507.18546].
13. Nguyen, D.H.; Seo, A.; Nnamdi, N.P.; Son, Y. False Alarm Reduction Method for Weakness Static Analysis Using BERT Model. *Applied Sciences* **2023**, *13*, 3502. https://doi.org/10.3390/app13063502.
14. Park, S.; Seo, A.; Cheong, M.; Kim, H.; Kim, J.; Son, Y. Evaluating the Vulnerability of Hiding Techniques in Cyber-Physical Systems Against Deep Learning-Based Side-Channel Attacks. *Applied Sciences* **2025**, *15*. https://doi.org/10.3390/app15031548.
15. Lee, K.; He, L.; Lewis, M.; Zettlemoyer, L. End-to-End Neural Coreference Resolution. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 188–197. https://doi.org/10.18653/v1/D17-1018.
16. Rebedea, T.; Dinu, R.; Sreedhar, M.; Parisien, C.; Cohen, J. NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails, 2023, [arXiv:cs.CL/2310.10501].
17. Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; et al. Llama Guard: LLM-Based Input–Output Safeguard for Human–AI Conversations, 2023, [arXiv:cs.CL/2312.06674].
18. Guardrails AI, Inc.. Guardrails AI: Adding Guardrails to Large Language Models, 2024. Available online: https://github.com/guardrails-ai/guardrails (accessed on 27 February 2026).
19. OWASP Foundation. OWASP Top 10 for Large Language Model Applications, version 2025, 2025. PDF: https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-v2025.pdf.
20. Nissenbaum, H. Privacy as Contextual Integrity. *Washington Law Review* **2004**, *79*, 119–158. Available online: https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/ (accessed on 27 February 2026).
21. Narayanan, A.; Shmatikov, V. Robust De-anonymization of Large Sparse Datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy (S&P), 2008, pp. 111–125. https://doi.org/10.1109/SP.2008.33.
22. Packer, C.; Wooders, S.; Lin, K.; Fang, V.; Patil, S.G.; Stoica, I.; Gonzalez, J.E. MemGPT: Towards LLMs as Operating Systems, 2023, [arXiv:cs.AI/2310.08560].
23. Qian, H.; Ai, Z.Z.; Jian, Y.B.; Wang, F. Recursively Summarizing Enables Long-Term Dialogue Memory in Large Language Models, 2023, [arXiv:cs.CL/2308.15022].

 doi:10.20944/preprints202603.0209.v1

24. Dong, Y.; Mu, R.; Jin, G.; Qi, Y.; Hu, J.; Zhao, X.; Meng, J.; Fu, W.; Huang, X. Building Guardrails for Large Language Models, 2024, [arXiv:cs.CL/2402.01822].
25. AlSayyad, A.; Huang, K.Y.; Pal, R. AgentTrace: A Structured Logging Framework for Agent System Observability, 2026, [arXiv:cs.SE/2602.10133]. AAAI 2026 Workshop LaMAS, https://doi.org/10.48550/arXiv.2602.10133.
26. Samarati, P.; Sweeney, L. Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression. In Proceedings of the IEEE Symposium on Research in Security and Privacy, 1998. Available online: https://epic.org/wp-content/uploads/privacy/reidentification/Samarati_Sweeney_paper.pdf (accessed on 27 February 2026).
27. Sweeney, L. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **2002**, *10*, 557–570. https://doi.org/10.1142/S0218488502001648.
28. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. ℓ-Diversity: Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data* **2007**, *1*, 3–es. https://doi.org/10.1145/1217299.1217302.
29. Li, N.; Li, T.; Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity. In Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE), 2007, pp. 106–115. https://doi.org/10.1109/ICDE.2007.367856.
30. Otmazgin, S.; Cattan, A.; Goldberg, Y. 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 (EACL), 2023, pp. 2752–2760. https://doi.org/10.18653/v1/2023.eacl-main.202.
31. Garfinkel, S.L. NIST Internal Report 8053: De-Identification of Personal Information. Technical Report NIST IR 8053, National Institute of Standards and Technology, 2015. https://doi.org/10.6028/NIST.IR.8053.
32. Chen, D.; Chen, H.; Yang, Y.; Lin, A.; Yu, Z. Action-Based Conversations Dataset: A Corpus for Building More In-Depth Task-Oriented Dialogue Systems. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Association for Computational Linguistics, 2021, pp. 3002–3017. https://doi.org/10.18653/v1/2021.naacl-main.239.
33. Peskov, D.; Clarke, N.; Krone, J.; Fodor, B.; Zhang, Y.; Youssef, A.; Diab, M. Multi-Domain Goal-Oriented Dialogues (MultiDoGO): Strategies Toward Curating and Annotating Large Scale Dialogue Data. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, 2019, pp. 4526–4536. https://doi.org/10.18653/v1/D19-1460.
34. Wen, Y.; Zhang, Y.; Lian, J.; Yi, X.; Xie, X.; Yang, D. Contextualized Privacy Defense for LLM Agents, 2025. Preprint. Available online: https://github.com/SALT-NLP/contextual_privacy_defense (accessed on 27 February 2026).
35. Budzianowski, P.; Wen, T.H.; Tseng, B.H.; Casanueva, I.; Ultes, S.; Ramadan, O.; Gašić, M. 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 (EMNLP). Association for Computational Linguistics, 2018, pp. 5016–5026. https://doi.org/10.18653/v1/D18-1547.
36. Si, S.; Ma, W.; Gao, H.; Wu, Y.; Lin, T.E.; Dai, Y.; Li, H.; Yan, R.; Huang, F.; Li, Y. SpokenWOZ: A Large-Scale Speech-Text Benchmark for Spoken Task-Oriented Dialogue Agents. *Advances in Neural Information Processing Systems (NeurIPS)* **2023**, *36*. Available online: https://arxiv.org/abs/2305.13040 (accessed on 27 February 2026).
37. Kim, J.; Park, S.; Cha, J.; Son, E.; Son, Y. Novel Synthetic Dataset Generation Method with Privacy-Preserving for Intrusion Detection System. *Applied Sciences* **2025**, *15*, 10609. https://doi.org/10.3390/app151910609.
38. Alabdulwahab, S.; Kim, Y.T.; Son, Y. Privacy-Preserving Synthetic Data Generation Method for IoT-Sensor Network IDS Using CTGAN. *Sensors* **2024**, *24*, 7389. https://doi.org/10.3390/s24227389.
39. Geifman, Y.; El-Yaniv, R. Selective Classification for Deep Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS)* **2017**, *30*. Available online: https://papers.nips.cc/paper/7073-selective-classification-for-deep-neural-networks (accessed on 27 February 2026).
40. U.S. Department of Health and Human Services. Standards for Privacy of Individually Identifiable Health Information: Final Rule (HIPAA Safe Harbor §164.514). Technical report, U.S. Department of Health and Human Services, 2002. Available online: https://www.law.cornell.edu/cfr/text/45/164.514 (accessed on 27 February 2026).

 doi:10.20944/preprints202603.0209.v1

41. Article 29 Data Protection Working Party. Opinion 05/2014 on Anonymisation Techniques. Technical Report WP 216, European Commission, 2014. Available online: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (accessed on 27 February 2026).
42. Mukhopadhyay, S.; Reddy, S.; Muthukumar, S.; An, J.; Kumaraguru, P. PrivacyBench: A Conversational Benchmark for Evaluating Privacy in Personalized AI. *arXiv preprint arXiv:2512.24848* **2025**. https://doi.org/10.48550/arXiv.2512.24848.
43. European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation), 2016. Available online: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed on 27 February 2026).
44. State of California. California Consumer Privacy Act (CCPA), Cal. Civ. Code §§ 1798.100–1798.199.100, 2018. Available online: https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=CIV&sectionNum=1798.100. (accessed on 27 February 2026).