Lexicon: Asynchronous A2A Collaboration for Proactive and Privacy-Centric Speech-Based AI

Anonymous ACL submission

Abstract

We introduce *Lexicon*, a privacy-aware asynchronous assistant-to-assistant (A2A) collaboration framework for proactive speech-based personal AI agents. As assistants evolve from reactive systems to always-listening companions, they face a significant privacy barrier to real-world social deployment: recording non-consenting speakers during everyday interactions. To address this, we propose and implement *Lexicon*, which enforces owner-only speech capture via real-time speaker verification, producing a one-sided transcript that preserves privacy but inevitably incurs missing conversational context. We demonstrate that *Lexicon* can recover necessary context through (1) spatio-temporal context resolution, (2) information gap detection, and (3) minimal A2A queries governed by relationship-aware disclosure policies. Rather than relying on hallucination-prone inference, *Lexicon* treats context recovery as a negotiated privacy-conscious exchange between assistants. We evaluate *Lexicon* on both a large semi-synthetic conversational dataset and a human-annotated multi-user conversational dataset containing natural group discussions and overlapping interactions. Across these evaluations, *Lexicon* achieves 91.4% recall in information-gap detection, 98% relationship classification accuracy, and up to a 97% true negative rate in privacy-sensitive disclosure decisions. By reframing always-listening AI as a coordination problem between privacy-preserving agents, *Lexicon* offers a practical path toward socially deployable proactive conversational assistants.

1 Introduction

Artificial Intelligence (AI) has enabled voice assistants to evolve from reactive query–response models to proactive support that anticipates intent and resolves ambiguities during live interaction (Chen et al., 2025; Ebert et al., 2025; Acikgoz et al., 2025). Always-listening assistants can accelerate this shift with personalization, but they raise serious privacy concerns in social settings, where speech capture inadvertently records non-consenting participants. Because speech is deeply personal and identity-revealing, this tension poses a central challenge to deploying proactive assistants in practice.

In this work, we present *Lexicon*, a framework which attempts to answer the following: *How can proactive conversational assistants recover missing context through coordinated follow-ups between assistants?* *Lexicon* is a first-of-its-kind framework that aims to unlock privacy-centric AI assistants through a 3-pronged approach. First, each assistant continuously captures only the device owner's speech via real-time speaker verification, ensuring strict boundaries. This one-sided capture, however, introduces missing conversational context. Second, *Lexicon* detects and resolves these context gaps using spatio-temporal reasoning, inferring information when possible from available context or issuing targeted queries to interlocutors' agents. Third, *Lexicon* enables privacy-aware assistant-to-assistant (A2A) communication governed by a relationship-based trust model that controls what information may be safely shared. For example, in a conversation between Bob and Alice, Bob's *Lexicon* operates in the background, listening only to Bob. Asynchronously, it identifies gaps, resolves local references, and formulates targeted queries to Alice for the missing information. Alice's *Lexicon* checks these queries against relationship-based permissions and responds only with information that can be shared. Through this A2A exchange, Bob's agent can reconstruct missing context, example: "..meet at 8 AM", without Alice's audio.

Our approach reframes always-on AI not as a surveillance problem, but as a coordination problem between privacy-preserving assistants, striking a practical balance between continuous intelligence and consent-driven data sharing.

2 The Lexicon Architecture

Lexicon's foundation is privacy-centric real-time speaker verification, which results in a one-sided transcript for the authenticated user (henceforth referred to as user transcript). Following this, the objective of *Lexicon* is to detect information gaps in this transcript, and resolve them through relationship and sensitivity-driven A2A collaboration. Figure 1 shows our system pipeline through an example. The following subsections delve into the details of this approach and implementation.

2.1 Speaker Verification

Speaker verification and identification are well-established research areas with extensive prior work (Wang et al., 2023; Liu et al., 2023). We implement speaker verification using the ECAPA-TDNN architecture (Desplanques et al., 2020) at the voice recording end. We selected this model for its lightweight design and strong performance. It operates as the primary privacy filter for *Lexicon*, ensuring isolation and recording of only the authorized user. For on-device deployment, we implemented the model on Android (Samsung S20 Ultra) using TFLite optimizations. The system achieves a Real-Time Factor (RTF) of 0.08, enabling real-time background execution with minimal computational overhead. In our framework, false positives are critical, as recording an unauthorized speaker may lead to privacy violations. Hence, we prioritize minimizing false positives, even at the cost of a lower True Positive Rate (TPR). To enforce this constraint, we fix the False Positive Rate (FPR) at 1% and select the corresponding decision threshold.

2.2 Context Resolution and Information Gap Detection

This stage resolves implicit references and identifies missing information required from the interlocutor for completing actionable items. We perform this in two steps, both implemented using few-shot GPT-4.1 with a low temperature ($T = 0.1$) to prioritize deterministic outputs (Renze, 2024).

■ **Local Reference Resolution**: This step combines the user transcript with mobile metadata (e.g.,location, time, calendar) to resolve implicit references. We first resolve ambiguous linguistic expressions such as pronouns, temporal phrases, and spatial mentions by grounding them in prior dialogue turns and real-world metadata (including GPS location, Wi-Fi SSID, and calendar events). For example, an utterance such as "I'll go there" links "there" to a specific venue inferred from recent dialogue and current location, while phrases such as "after that" or "tomorrow" are mapped to concrete calendar events when possible. This step determines whether a missing entity can be fully resolved using only locally available conversational and mobile data information. *Lexicon* prioritizes local resolution before initiating any A2A communication, minimizing unnecessary requests.

■ **Information Gap Detection**: After attempting local resolution, *Lexicon* identifies entities that remain unresolved from the user transcript and available metadata. We perform turn-level analysis using a sliding window over the current turn and the previous five turns. This focuses gap detection on the most relevant conversational context while limiting interference from distant dialogue history.

Rather than asking the model whether a sentence contains sufficient detail to resolve context in a user transcript, we define explicit attribute requirements for each common entity type. (a) Medical entities require name, dosage, and frequency; (b) Temporal entities require event-level specificity beyond relative references; (c) Spatial entities require building-, floor-, or room-level identifiers; (d) General coordination entities require identifying attributes for people, objects, or tasks.

Entities that still fail these requirements are labeled as "information gaps". The system then queries the interlocutor's assistant for missing attributes under the disclosure policy described next. At this stage, we prioritize recall to avoid missing critical entities, allowing later stages to suppress unnecessary queries. Together, these stages separate locally recoverable context from information that requires explicit A2A coordination.

2.3 Privacy Management

Next, *Lexicon* initiates A2A collaboration to fill these gaps by assessing interlocutor trust levels and enforcing social rules for safe information sharing.

■ **Relationship Assessment**. We categorize interactions into three trust levels modelled after sociological research on social distance, which organizes human interactions based on degree of intimacy (Guerdelli et al., 2023). To distinguish between these levels, we use three primary indicators: (i) *Shared reference density*, which measures the ratio of implicit to explicit references to identify a shared history between speakers; (ii) *Conversational formality*, which analyzes sentence structure

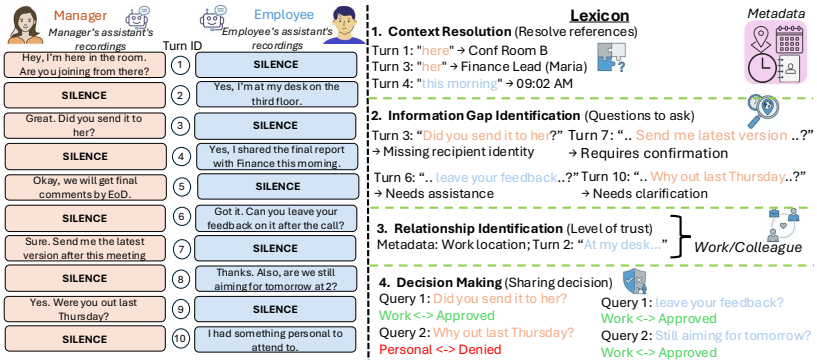

Figure 1: *Lexicon*: one-sided transcript processing pipeline. (Left) Multi-turn two-user dialogues, where each user's assistant only records their authenticated user. (Right) The system performs spatio-temporal reference resolution and identifies information gaps (2.2), then applies a decision filter to approve queries based on the relationship (2.3).

Relation Level	Proximity	Linguistic Indicators	Disclosure Policy by Sensitivity		
			Low	Mid	High
L_1	Intimate (High)	Implicit references to private spaces ("the bedroom"); elliptical commands ("grab that"); relational nouns ("mom", "babe"); high ambiguity tolerance due to shared grounding.	Direct Reveal	Direct Reveal	Direct Reveal
L_2	Social (Moderate)	Balanced shared history; first-name address; conversational but more explicit entity references; moderate formality.	Direct Reveal	Approval Loop	Suppress
L_3	Professional (Low)	Formal register; honorifics (*Dr., Mr., Ms.*); task-oriented vocabulary; distancing modal verbs (*could you, would you*).	Approval Loop	Suppress	Suppress

Table 1: Unified framework: relationship level inference (left) and sensitivity-aware disclosure policy (right).

to detect shifts from formal to casual phrasing; and (iii) *Naming conventions*, which tracks the use of title or honorifics versus endearment or first names as a marker of social distance. Table 1 outlines the specific indicators used for each relationship level.

Trust Level Assignment: We determine the relationship level by aggregating linguistic markers over a sliding conversational window. While our current categories are preliminary, they are anticipated to grow as the system learns through broader deployment. We design the framework to follow a strict privacy-first logic. To prevent accidental disclosure, if the system detects a threshold of distance indicators, the interaction is immediately locked into Level 3 (Professional). Conversely, we prioritize safety by requiring multiple intimacy indicators before promoting an interaction to Level 1 (Intimate). This ensures that isolated collective terms such as "our" do not trigger high-trust sharing prematurely since such language can appear in non-intimate settings too. These linguistic signals serve as high-confidence indicators of trust that operate in addition to any user-defined permissions.

■ **Hybrid Relationship-Based Disclosure** We use a two-stage disclosure filter that balances strict privacy guarantees with social flexibility. High-risk data is protected by deterministic hard locks that override all social reasoning, while lower-risk interactions incorporate contextual relationship analysis to enable helpful, situation-appropriate responses. We propose privacy as a dynamic social negotiation rather than a static setting, adapting disclosure to the current context without compromising sensitive information.

Deterministic Hard Lock: This filter acts as a high-security safeguard for data with zero tolerance for social disclosure, such as banking credentials or government IDs. We use a deterministic override to prevent the system from even considering social norms for these critical items:

$$D_{\text{gate}}(E, L) = \begin{cases} \text{Abort} & \text{if } \sigma(E) > L \\ \text{Proceed} & \text{otherwise} \end{cases}$$

If the requested information is classified as high-sensitivity (Level L_1) and the interlocutor is not an intimate contact (assigned to L_1), the system immediately terminates the proactive prompt to prevent any risk of exposure.

Fuzzy Social Norms Matrix: Requests that pass the initial security filter are evaluated against a 3×3 Social Sharing Matrix (Table 1 (right)). This matrix maps our three relationship levels (L_1, L_2, L_3) against the sensitivity of the information (Low, Mid, High) to select a disclosure protocol. A Privacy Intent Detector provides an additional safeguard:

deliberate abstraction, self-censoring language, or avoidance cues by the user (e.g., "I had some personal stuff") automatically elevate sensitivity. This causes the assistant to mirror the user's caution, shifting from direct responses to permission-based or no responses. The final decision yields (Table 1)) one of four outcomes: (1) Direct Reveal: Full information is shared, (2) Partial Reveal with sensitive entities hidden, (3) Approval Loop: via an explicit private signal (haptic buzz) before speaking. (4) Suppression: silence to protect owner's privacy.

3 Dataset and Baselines

3.1 Dataset

Lexicon requires two types of datasets: (i) a speaker verification dataset, and (ii) a turn-based dialogue dataset involving conversations between two users. We use speech between two interlocutors as a preliminary exploration case.

Dataset for speech verification: We use the VoxConverse (Chung et al., 2020) dataset, which comprises 50 hours of multi-speaker, turn-based conversational audio extracted from YouTube videos. We employ the ECAPA-TDNN model pre-trained on the VoxCeleb (Nagrani et al., 2017) dataset for target speaker verification.

Dialogue dataset generation: Existing dialogue datasets (Li et al., 2017) are not suited for studying context resolution, information gap detection, relationship-aware reasoning and decision-making. While several synthetic and semi-synthetic resources exist (Wu et al., 2024; Chen et al., 2025; Ge et al., 2025), they are optimized for response generation or task completion and lack explicit, fine-grained annotations for conversational resolution and information gaps. In particular, they do not encode missing temporal cues, spatial references, contextual dependencies, or implicit assumptions about relationships. As a result, they do not provide the ground truth necessary to evaluate structured gap detection and relationship-based decisions.

To address this limitation, we construct a structured synthetic dataset with explicit control over contextual dependencies and information gaps. Rather than relying on post-hoc annotation, we design an event-based generation pipeline in which the model first produces structured answers and gap specifications, and later generates dialogues conditioned on these constraints. This reduces hallucinations and ensures faithful realization of intended resolution requirements. Each dialogue contains explicit annotations for missing entities, required resolutions, and relationship-aware disclosure decisions. The resulting dataset contains approximately 5,700 dialogues spanning nine relationship types across high-stakes, workplace, academic, and informal settings (Table 3). We validate annotation quality through evaluation against 500 sampled dialogues independently labelled by three students, achieving a Cohen's Kappa score of 0.78 (good agreement) (Li et al., 2024a).

Human-annotated conversational dataset: Beyond synthetic data, we evaluate *Lexicon* on a real multi-party conversational dataset collected under IRB approval. The dataset consists of 20 participants divided into four groups of five users each. Before the group sessions, we conducted informal one-on-one conversations with each participant to capture lightweight contextual information such as hobbies, routines, interests, and living situations, approximating the type of longitudinal memory future egocentric assistants may gradually accumulate over time. Each group then participated in two conversational sessions consisting of collaborative story reading and open-ended group discussions involving topics such as food preferences, travel planning, outdoor activities, and pet ownership. In total, the dataset contains ≈15,000 spoken words across natural multi-user interactions. The story-reading sessions were designed to evaluate speaker verification robustness under structured conversational turn-taking between multiple speakers. In contrast, the group discussion sessions were intentionally unstructured and non-turn-based, containing overlapping speech, interruptions, conversational drift, humor, disagreement, and spontaneous coordination behaviors. These interactions better approximate realistic conversational environments where multiple speakers dynamically enter and leave discussions without explicit conversational boundaries.

To analyse these interactions, we annotated the multi-user conversational dataset for metrics defined in Section 4.2. Similar to our synthetic dataset evaluation, three independent student annotators labelled the conversations following a shared annotation protocol. We measure inter-annotator agreement using Cohen's Kappa and obtain a score of 0.83, indicating strong agreement of the collected annotations for conversational analysis. Unlike conventional meeting-analysis datasets that assume continuous full-room recording, our setting follows the same privacy-centric assumption as *Lexicon*: the assistant only observes the authenticated user's

egocentric speech stream. As a result, the assistant reasons over partial conversations rather than globally captured conversations.

3.2 Baselines

We evaluate *Lexicon* using three widely adopted reasoning-based resolution techniques: *Chain-of-Thought* (CoT) (Wei et al., 2022), *Chain-of-Thought with Self-Reflexion* (CoT-SR) (Shinn et al., 2023), and *Tree-of-Thought* (ToT) (Yao et al., 2023). **CoT** allows the LLM to generate intermediate reasoning steps before producing the final answer. **CoT-SR** method extends CoT performance by enabling the model to evaluate and refine its reasoning through self-reflection. **ToT** method allows the model to evaluate alternative reasoning paths before selecting a final output (see §C for details).

We further evaluate *Lexicon* on other LLM variants- GPT-4o, Claude-3.5 Sonnet, and Gemini-1.5. We limit our evaluation to these variants, compared to (Zhang et al.; Lyu et al., 2023; Mo and Xin, 2024; Chen et al.) due to their ease of generalization and consistent performance in diverse conversational scenarios.

Also, we evaluate *Lexicon* on two inference only models- One-Sided Conversation (1SC) (Ebert et al., 2025) and Infilling Language Model (ILM) (Donahue et al., 2020). Both the models infer user dialogues and contextual gaps using prior turn dialogues and conversation metadata.

4 Evaluation

We evaluate speaker verification followed by *Lexicon*'s four primary components: local reference resolution, information gap detection, relationship identification, and final disclosure decision (Table 2). Our evaluation measures *Lexicon*'s ability to identify and resolve actionable gaps while maintaining privacy-aware information disclosure.

4.1 Speaker Verification Accuracy

We evaluate the speaker verification model on the VoxConverse test dataset, which comprises multi-speaker audio extracted from 232 YouTube videos. The dataset spans 2,612 minutes and includes 8,268 annotated speaker turns. The model is applied to an audio signal using a sliding window with a length of 2 seconds and an overlap of 0.5 seconds (empirically optimized). This configuration results in ~450 evaluated audio segments per recording on average, yielding ~104,400 evaluated segments across the dataset. The model achieves a 10.8% False Negative Rate (FNR) and a 89.2% True Positive Rate (TPR). For story reading sessions, we achieve similar results, 92.4% TPR and 7.6% FNR. We intentionally use a conservative recording threshold to reduce bystander audio capture, even at the cost of occasional missed user recordings.

4.2 Component-Level and End-to-End Evaluation

We evaluate *Lexicon* at both the component and end-to-end levels. Component-level evaluation isolates the contribution of each module under controlled conditions, while end-to-end evaluation measures the robustness of the complete privacy-aware coordination pipeline under accumulated errors. We first show isolated component results.

Spatio-Temporal Context Resolution: We first evaluate the system's ability to identify and ground ambiguous references within the (partial) user transcript. On the semi-synthetic dataset, the resolution engine achieves a 78% TPR in identifying turns that require context resolution, while achieving 72.6% TPR on the human-annotated conversational dataset. For these identified references, the system produces resolutions with a 58% semantic similarity compared to the ground truth annotations on the semi-synthetic dataset and 52.8% on the human-annotated dataset. While linguistic variation in location descriptions (e.g., "Clinic Hallway" vs. "North Wing Corridor") limits exact-match performance, the consistently high TPR across both settings ensures that the majority of information gaps are identified before downstream reasoning. Although the SOTA baseline achieves slightly higher semantic similarity, its substantially lower TPR shows that it resolves fewer ambiguous references overall, likely focusing only on high-confidence cases. Table 4 shows detailed results.

Information Gap Detection: We prioritize recall in gap detection to ensure no critical (e.g., medical or logistical) detail is overlooked. On the semi-synthetic dataset, the system achieves a 91.4% TPR while maintaining a low 6.8% FPR. On the human-annotated conversational dataset, *Lexicon* achieves an 88.5% TPR with a 10.3% FPR despite increased conversational variability, overlapping speech patterns, and unstructured interactions. We consider this an acceptable trade-off, as false positives only trigger additional clarification requests that the interlocutor's agent may decline or refine. This en-

Framework Component	Evaluation Metric	Semi-Synthetic Dataset			Human-Annotated Conv.		
		Zero-shot LLM	SOTA (CoT-SR)	*Lexicon*	Zero-shot LLM	SOTA (CoT-SR)	*Lexicon*
Local Reference Resolution	True Positive Rate (TPR % ↑)	28.4	35.5	**78.3**	24.1	33.2	**72.6**
	Sentence Similarity (% ↑)	34.9	**65.7**	58.1	31.6	**55.2**	52.8
Info Gap Detection	True Positive Rate (TPR % ↑)	46.2	45.8	**91.4**	42.0	44.1	**88.5**
	False Positive Rate (FPR % ↓)	22.8	19.5	**6.8**	25.4	20.1	**10.3**
Relationship Assessment	Classification Accuracy (% ↑)	95.0	**97.0**	96.4	93.1	95.8	**98**
Hybrid Decision Gate	True Negative Rate (TNR % ↑)	66.5	94.0	**97.0**	63.8	92.3	**93.8**
	True Positive Rate (TPR % ↑)	56.3	12.0	**86.0**	50.2	10.7	**88.4**

Table 2: Performance of *Lexicon* across different framework components evaluated on semi-synthetic and human-annotated conversational datasets. Across both settings, *Lexicon* achieves strong information-gap detection and decision gating performance while maintaining high safety integrity.

ables *Lexicon* to function as an active conversational agent rather than a passive recorder.

Implicit Relationship Identification: The relationship inference block achieves 96.4% classification accuracy on the semi-synthetic dataset and 98% accuracy on the human-annotated conversational dataset across the three trust levels. We observe that misclassifications are predominantly directed toward more restrictive relationship levels (e.g., classifying an L_2 Social contact as an L_3 Professional). This intentional *safety bias* reflects the system's conservative privacy design: under ambiguous conversational evidence, the model defaults to stricter access boundaries to minimize unintended information disclosure. Since interlocutors can subsequently provide additional context to re-establish trust, we prioritize conservative relationship estimation over permissive inference.

Hybrid Decision Gate: We evaluate the system's ability to execute the appropriate disclosure policy based on the resolved conversational context and inferred interpersonal relationship. On the semi-synthetic dataset, *Lexicon* achieves an 86% TPR, indicating that the system correctly shares information in the majority of scenarios where disclosure is socially appropriate. On the human-annotated conversational dataset, the system achieves an 88.4% TPR despite increased conversational ambiguity and unstructured interactions. From a privacy-preservation perspective, *Lexicon* achieves a 97% TNR on the semi-synthetic dataset and 93.8% TNR on the human-annotated dataset, showing that the system reliably suppresses sensitive information in scenarios where disclosure would violate inferred social boundaries. These results indicate that the deterministic hard-locks and the fuzzy social matrix effectively guard against privacy breaches.

Lexicon* in comparison to inference only models:** We compare *Lexicon* against two inference-only models. The 1SC (Ebert et al., 2025) inference model masks the non-target user's dialogue and predicts it from the target user's prior dialogues. We found that the 1SC model often skips reference resolution dialogue section. For example, in the dialogue, *"Hi, Emily, can you pull up the meeting details *****here** "*, 1SC model predicts *"Hi, Emily, can you show the meeting points"*. For the semi-synthetic and human-annotated conversations data, 1SC achieves a 20% and 18.7% median semantic similarity for reference resolution respectively. The ILM (Donahue et al., 2020) resolves ambiguous phrases using conversation metadata and local dialogue context. For the semi-synthetic and human-annotated conversations data, ILM achieves 11% and 12.7% semantic similarity. On the other hand, *Lexicon* achieves 78.3% resolution with a semantic similarity of 58.1% for semi-synthetic data and 52.8% for real-world data.

***Lexicon* End-to-End Evaluation:** Beyond isolated component evaluation, we evaluate *Lexicon* end-to-end to measure the effect of error propagation across the pipeline. We observe moderate degradation under full pipeline evaluation. On the human-annotated conversational dataset, local reference resolution decreases from 72.6% to 70.8% TPR, while sentence similarity decreases from 52.8% to 50.3%. These upstream errors partially propagate into information-gap detection, reducing TPR from 88.5% to 83.9% and increasing FPR from 10.3% to 15.6%. Nevertheless, the sys-

tem maintains high recall and identifies actionable gaps. Also, relationship assessment remains comparatively robust, decreasing only from 98% to 97% classification accuracy despite upstream imperfections. This is because relationship inference mainly depends on linguistic style, tone, and naming conventions. Table 4 shows detailed results.

5 Preliminary Multi-Party Conversational Observations

Beyond quantitative evaluation, we use our preliminary multi-user conversational dataset to study how privacy-preserving assistant-to-assistant coordination may support multiple applications.

■ **Conversational Alignment and Social Dynamics** One discussion centered around differing opinions on pineapple pizza, food authenticity, and cultural expectations. Although superficially casual, the interaction showed recurring patterns, disagreement styles, and humor-driven dynamics. Participants repeatedly reinforced, challenged, or softened one another's positions through exaggeration, joking, and conversational turn-taking. For example- "*...pizza should not be sweet. That is a sin against pizza...*" and "*I think it's about balance. Some pizzas work well with pineapple.*" Even under partial egocentric observation, these interactions show how privacy-preserving assistants may infer social engagement patterns. Rather than modeling isolated food preferences, the discussion shows how conversational context can reveal higher-level social signals such as playful disagreement, mediation behavior, and recurring alignment between speakers across interactions.

■ **Cross-Session Continuity and Lifestyle Reasoning** Another discussion focused on preferences for cats versus dogs, gradually evolving into conversations about travel frequency, work routines, and long-term responsibility management. Few participants initially described themselves as "dog people", but later justified a preference toward cats due to practical constraints such as frequent travel, apartment restrictions, or inconsistent caregiving availability. Example statements, "*I have always been a dog person, but due to the maintenance and walking... I'm planning to probably adopt a cat soon*" and "*When you travel a lot... you always have to find somebody to take care of the dog.*" Interestingly, some participants had previously mentioned during earlier one-on-one interactions that they lived alone. When viewed across conversations, these later group discussion statements provide additional contextual grounding for understanding how lifestyle constraints influence long-term preferences and decisions. This highlights how privacy-preserving assistants may gradually accumulate longitudinal conversational understanding across repeated egocentric interactions, even under partial conversational observability and without continuous full-room recording.

■ **Collaborative Coordination and Group Activity Formation** Another discussion explored preferences between board games, hiking, and outdoor activities, gradually evolving from individual preference sharing into collaborative group coordination. Over the course of the interaction, the conversation transitioned from abstract preference discussion toward identifying shared interests and mutually compatible activities. Example statements,"*Board games you can always play, but someone inviting you to a hike is more unusual*" and "*You train your body and your brain. I like both.*" Interestingly, several participants later coordinated outside the session and organized a group pickleball activity. This interaction illustrates how privacy-preserving assistants, if permitted to collaboratively coordinate through A2A communication could suggest activities or form shared plans in the background while still preserving privacy-centric sensing constraints.

6 Related Work

With Large Language Models (LLMs), assistant-centric frameworks have started to evolve and interact with our daily lives. For example, LLAMAPIE (Chen et al., 2025), DailyLLM (Tian et al., 2025), show how agents can be useful for conversation guidance, or automated daily journaling. However, current agent-based systems often assume the availability of reliable and complete user information. In practice, privacy constraints make such contextual data difficult to obtain.

Limitations of Current Access Protocols (MCP). LLMs have shown potential to support daily-life activities through automated logging of personal information and personalized assistance (Tian et al., 2025; Li et al., 2024b; Xu et al., 2024). Model Context Protocol (MCP) (Hou et al., 2025) enables this further by interfacing AI systems with external tools, APIs, and sensor-driven resources. Prior efforts (Guo et al., 2025; Yang et al., 2025) show how MCP-integrated LLMs working with IoT devices

and contextual data streams can enable automation. However, these access protocols primarily emphasize connectivity and tool invocation. They do not explicitly model interpersonal relationships, interaction stakes, or socially grounded access constraints. As a result, decisions about information retrieval and sharing are typically governed by static permissions rather than dynamic reasoning.

One-Sided Conversation Reconstruction: Risks of Inference-Based Context Recovery. Recent studies on contextual and temporal reasoning (Liu et al.; Ruiz et al., 2025; Wang and Zhao, 2024; Ge et al., 2025; Ebert et al., 2025) show that LLMs can infer missing or implicit information in conversational settings. Existing approaches often rely on long conversational histories (Liu et al.), persistent user memory (Ge et al., 2025), or stored personal data to fill contextual gaps.While these systems improve contextual continuity, they implicitly assume that missing information can be safely inferred from available signals.

Also, current systems fall short for A2A collaboration and either retrieve information without considering relations or infer missing details without guarantees of correctness. *Lexicon* allows for collaborative information access and models human relationships and the constraints they impose.

Speaker verification and A2A communication: Prior HCI-based work on privacy-aware speaker verification, such as MeMic (Zulfikar et al., 2024) and Malkin et al. (Malkin et al., 2022), focuses on hardware- and permission-level enforcement of data access. In contrast, *Lexicon* address context gaps through controlled and policy-aware information exchange. The recent Google A2A protocol (Google Blog, 2025) proposes a standardized approach for agent communication. *Lexicon* is designed to add a privacy layer to such systems.

7 Conclusion

By framing missing context as information gaps and enforcing relationship-aware disclosure policies, *Lexicon* demonstrates that proactivity, privacy, and societal usability can coexist. Through evaluations on both semi-synthetic and human-annotated multi-party conversational datasets, we show that *Lexicon* can identify actionable conversational gaps, selectively recover missing context, and enforce privacy-sensitive disclosure policies while preserving interlocutor privacy. *Lexicon* highlights a shift in how future proactive assistants may operate in socially realistic environments. As always, listening AI systems become increasingly embedded in everyday interactions, privacy-preserving coordination, minimal disclosure, and partial conversational reasoning will become system requirements rather than optional constraints.

8 Limitations

Lexicon represents a preliminary step toward privacy-preserving proactive conversational assistants built on one-sided speech capture, information-gap identification, context resolution, and assistant-to-assistant (A2A) coordination. We identified the following future research challenges:

Limitations of current conversational evaluation: While we evaluate with human-annotated multi-party conversational dataset, a large scale evaluation on large-scale evaluation remains due to the difficulty of collecting privacy-sensitive conversational interactions at scale. Real-world conversations exhibit substantially greater variability, including overlapping speech, topic drift, sarcasm, and interruptions, which are difficult to fully capture in synthetic data.

Open challenges in conversational reasoning: Several capabilities explored in *Lexicon* require more study before deployment in socially realistic settings. Multi-speaker conversational environments introduce dynamic participation, overlapping speech, and fragmented conversational context that complicate reliable assistant coordination. Similarly, extending context recovery beyond speech through multimodal signals such as vision, email, calendars, and digital activity remains an important direction for improving task grounding and contextual understanding.

Deployment requirements for seamless operation: Beyond conversational reasoning itself, practical deployment of *Lexicon* requires robust infrastructure for participant discovery, asynchronous A2A communication, longitudinal memory management, and graceful degradation under partial assistant adoption. Importantly, the speaker verification pipeline already operates in real time and the remaining latency is currently dominated primarily by cloud-based LLM inference and A2A coordination requests. However, recent advances in compact on-device LLMs (Cheng et al., 2026; Allal et al., 2025) suggest that future versions of *Lexicon* may increasingly support low-latency or fully local privacy-preserving infrastructure.

References

Emre Can Acikgoz, Cheng Qian, Hongru Wang, Vardhan Dongre, Xiusi Chen, Heng Ji, Dilek Hakkani-Tür, and Gokhan Tur. 2025. A desideratum for conversational agents: Capabilities, challenges, and future directions. *arXiv preprint arXiv:2504.16939*.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, and 1 others. 2025. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*.

Tuochao Chen, Nicholas Scott Batchelder, Alisa Liu, Noah A Smith, and Shyamnath Gollakota. 2025. Llamapie: Proactive in-ear conversation assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13801–13824.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Wen Cheng, Tuochao Chen, Karim Helwani, Sriram Srinivasan, Luke Zettlemoyer, and Shyamnath Gollakota. 2026. Micro language models enable instant responses. *arXiv preprint arXiv:2604.19642*.

Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman. 2020. Spot the conversation: Speaker diarisation in the wild. In *INTERSPEECH*.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2492–2501.

Victoria Ebert, Rishabh Singh, Tuochao Chen, Noah A Smith, and Shyamnath Gollakota. 2025. Reading between the lines: The one-sided conversation problem. *arXiv preprint arXiv:2511.03056*.

Yubin Ge, Salvatore Romeo, Jason Cai, Raphael Shu, Yassine Benajiba, Monica Sunkara, and Yi Zhang. 2025. Tremu: Towards neuro-symbolic temporal reasoning for llm-agents with memory in multi-session dialogues. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18974–18988.

Google Blog. 2025. Announcing the agent2agent protocol (a2a). https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/. Accessed: 2026-02-20.

Hajer Guerdelli, Claudio Ferrari, and Stefano Berretti. 2023. Interpersonal relation recognition: a survey. *Multimedia Tools and Applications*, 82(8):11417–11439.

Yunqi Guo, Guanyu Zhu, Kaiwei Liu, and Guoliang Xing. 2025. Sensormcp: A model context protocol server for custom sensor tool creation. In *Proceedings of the 23rd Annual International Conference on Mobile Systems, Applications and Services*, pages 747–752.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. 2025. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *CoRR*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, and 1 others. 2024b. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.

Tianchi Liu, Kong Aik Lee, Qiongqiong Wang, and Haizhou Li. 2023. Disentangling voice and content with self-supervision for speaker recognition. *Advances in Neural Information Processing Systems*, 36:50221–50236.

Yanming Liu, Xinyue Peng, Jiannan Cao, Shi Bo, Yanxin Shen, Tianyu Du, Sheng Cheng, Xun Wang, Jianwei Yin, and Xuhong Zhang. Bridging context gaps: Leveraging coreference resolution for long contextual understanding. In *The Thirteenth International Conference on Learning Representations*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

Nathan Malkin, David Wagner, and Serge Egelman. 2022. Runtime permissions for privacy in proactive intelligent assistants. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 633–651.

Shentong Mo and Miao Xin. 2024. Tree of uncertain thoughts reasoning for large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12742–12746. IEEE.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *INTERSPEECH*.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 7346–7356.

Alfredo Garrachón Ruiz, Tomás de la Rosa, and Daniel Borrajo. 2025. On the temporal question-answering capabilities of large language models over anonymized data. *arXiv preprint arXiv:2504.07646*.

Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, and 1 others. Multi-turn reinforcement learning from preference human feedback, may 2024. *URL http://arxiv.org/abs/2405.14655*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Ye Tian, Xiaoyuan Ren, Zihao Wang, Onat Gungor, Xiaofan Yu, and Tajana Rosing. 2025. Dailyllm: context-aware activity log generation using multimodal sensors and llms. In *2025 IEEE 22nd International Conference on Mobile Ad-Hoc and Smart Systems (MASS)*, pages 372–380. IEEE.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. 2023. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yuqing Wang and Yun Zhao. 2024. Tram: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*.

Huatao Xu, Panrong Tong, Mo Li, and Mani Srivastava. 2024. Autolife: Automatic life journaling with smartphones and llms. *arXiv preprint arXiv:2412.15714*.

Ningyuan Yang, Guanliang Lyu, Mingchen Ma, Yiyi Lu, Yiming Li, Zhihui Gao, Hancheng Ye, Jianyi Zhang, Tingjun Chen, and Yiran Chen. 2025. Iot-mcp: Bridging llms and iot systems through model context protocol. In *Proceedings of the ACM Workshop on Wireless Network Testbeds, Experimental evaluation & Characterization*, pages 73–80.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Wazeer Deen Zulfikar, Cayden Pierce, and Pattie Maes. 2024. Memic: Towards social acceptability of user-only speech recording wearables. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9.

A Dataset generation

A.1 Keywords generation

For generating synthetic key words for a target conversation, we use the publicly available dataset for that relationship type. Some of the public available dataset are- DailyDialogy, ChatDoctor. We analyse the dialogue dataset and derive contextually relevant keywords

A.1.1 Keywords generation prompt

- **Keywords generation from publicly available datasets for teacher–student interaction**
  1. Analyse the conversation between the teacher and the student.
  2. Generate keywords based on the content of the conversation.
  3. Extract a list of educationally and conversationally relevant keywords or short phrases from the conversation. These keywords may include learning topics, academic concerns, study strategies, assignments, feedback, classroom activities, performance discussions, or instructional guidance provided by the teacher. Generic or filler words (e.g., "the," "and," or "okay") must be excluded.
  4. For each extracted keyword, assign a score from 1 to 10 indicating how commonly

the keyword appears in real-world teacher–student conversations, where:

  - 1 = Rarely used in teacher-student academic discussions
  - 10 = Very frequently used and standard in classroom or academic conversations

  The assigned ranking should reflect how commonly the keyword appears in typical educational interactions rather than how often it appears in the specific conversation.

5. Do not assume or infer academic problems or learning issues beyond what is explicitly stated or strongly implied in the conversation.
6. Use neutral, professional, and education-appropriate language.
7. Arrange the extracted keywords in descending order based on their assigned rank such that:
  - Keywords with a rank of 10 appear first
  - Keywords with lower ranks follow in decreasing order (9 → 1)
  - Keywords with the same rank may appear in any order relative to each other
8. Generate exactly 500 keywords.

- **Keywords generation with explicit prompt for housemates interaction**

1. Consider a conversation occurring between companions such as couples, flatmates, siblings, or individuals sharing a close personal or living relationship.
2. Generate keywords based on the semantic and conversational content of the dialogue.
3. Extract a list of lifestyle-focused, relationship-oriented, household coordination, emotional support, and daily life interaction keywords or short phrases from the conversation. Keywords may include, but are not limited to:
  - Household responsibilities or shared chores
  - Financial or expense planning discussions
  - Daily routine coordination or schedule planning
  - Emotional support or wellbeing discussions
  - Relationship communication or personal interaction topics
  - Living arrangement coordination or household management
  - Social or family event planning
  - Personal habits or lifestyle adjustment discussions
  - Shared resource or space usage coordination
  - Decision-making related to travel, purchases, or activities
  - Conflict resolution or compromise discussions
  - Personal health, wellness, or daily care coordination
  - Communication, trust, or relationship maintenance topics

  Generic or filler words (e.g., "the," "and," "okay," "yes") must be excluded.
4. For each extracted keyword, assign a score from 1 to 10 indicating how commonly the keyword appears in real-world companion or close personal relationship conversations, where:
  - 1 = Rarely used in everyday companion or shared living conversations
  - 10 = Very frequently used and standard terminology in household or close relationship interactions

  The assigned ranking should reflect typical conversational frequency in companion relationships rather than how often the keyword appears within the specific dialogue.
5. Do not infer or fabricate relationship conflicts, emotional distress, or personal issues beyond what is explicitly stated or clearly implied in the conversation.
6. Use neutral, respectful, and socially appropriate conversational language.
7. Arrange the extracted keywords in descending order based on their assigned rank such that:
  - Keywords with a rank of 10 appear first
  - Keywords with lower ranks follow in decreasing order (9 → 1)
  - Keywords with the same rank may appear in any order relative to each other
8. Generate exactly 500 keywords or short phrases.

  Note: The number of keywords is fixed at 500 to minimize the likelihood of hal-

lucinated outputs and the inclusion of non-essential or irrelevant terms.

A.1.2 Example keywords

- **Sample keywords generated after analysing a publicly available dataset (Teacher–student (Shani et al.))**
  Discussion, experiment, lecture, assessment, activities, timeline, question, answer, group work, interactive, hands-on, assignment, feedback, quiz, summary, example, role-play, classroom, lesson, research, presentation, worksheet, class, student, teacher, understanding, concept, topic, learning, instructions, activity, study, accomplishments, challenges, creative, motivation, problem, solution, performance, review, application, context, questions, history, science, art, method, group, strategy, critical thinking, politics, democracy, freedom, revolution, war, government, civil rights, independence, conflict, equality, leader, values, beliefs, social, economic, oppression, rights, historical, engineering, architecture, society, environment.
- **Sample keywords generated by explicit prompt to an LLM (Housemates)**
  Chores, cleaning, cooking, groceries, dishes, laundry, bills, rent, budget, dinner plans, schedule, bedtime, wake-up time, work hours, shopping list, trash, takeout, breakfast, lunch, dinner, pets, errands, cleaning schedule, shared expenses, grocery shopping, household supplies, utilities, Wi-Fi, dishes rotation, bathroom cleaning, vacuuming, recycling, laundry day, meal prep, food preferences, cooking rotation, shared meals, grocery run, fridge space, food storage, cleaning products, rent

A.2 Event generation

We utilize the generated keywords to generate fictional events related to that relationship.

A.2.1 Event generation prompt

- Analyse the provided keywords and construct potential user profiles.
- Generate ten events based on the keywords, modelling interactions between a teacher and a student.
- Carefully analyse each keyword and curate suitable teacher and student profiles such that their roles, academic goals, and learning challenges logically align with the provided keywords.
- Ensure that each generated event reflects realistic classroom environments, academic expectations, mentoring relationships, and authentic learning experiences.
- Output the events in the following structured JSON format:

```
{
    { ... },
    { ... },
    ...
    { ... }
}
```

- Each event must be written as a narrative story.
- Each event must begin with the following profile paragraph, followed by a continuous narrative:

  [Name] is a [age]-year-old [occupation] living in [location].

  Story Style Guidelines

  - Write the narrative in third-person perspective.
  - The story should naturally integrate:
    * Student thoughts, motivations, learning concerns, and academic pressures
    * Teacher expectations, mentoring approaches, and instructional strategies
    * Classroom dynamics, peer collaboration, and educational environment
  - Avoid including:
    * Explicit academic challenges or learning conflicts
    * Feedback, grading, or performance evaluation discussions
    * Instructional or academic decision-making processes
    * Student growth, skill development, or resolution of challenges
    * Classroom-, institutional-, or peer-level outcome summaries
  - Additional formatting constraints:
    * Do not use dialogue labels such as “Teacher:” or “Student:”

    - Do not use bullet points or section headings within the narrative
    - Avoid formal evaluation or administrative report writing style
    - Maintain a natural, coherent storytelling format

- Each generated event must contain a minimum of 500 words.

A.2.2 Example event

Evelyn is a 14-year-old student living in Boston. The classroom was alive with energy as the students gathered for the hands-on science lesson. Today, Evelyn's teacher, Ms. Garcia, had carefully prepared an experiment to help her students better understand the concept of buoyant force and density. Evelyn watched as Ms. Garcia demonstrated Archimedes' Principle using a tub of water, various objects, and a spring scale. The teacher's approach was interactive, inviting students to predict which objects would float and which would sink before testing their hypotheses. Evelyn felt a surge of excitement; she had always been curious about the science behind everyday phenomena. As her group worked through the experiment, she collaborated with her peers to record displacement and calculate weight, eager to see if their predictions matched the results. The teacher moved from group to group, observing their progress and offering encouragement, fostering a supportive environment where curiosity was celebrated. Evelyn felt motivated by the teamwork and the practical nature of the activity. The lesson ended with students sharing their findings, and Evelyn reflected on how the experiment deepened her understanding of the scientific method. She left class looking forward to more opportunities for discovery and hands-on learning.

A.3 Dialogue generation

Core Logic and Constraints

1. The Core Logic (Read Carefully)

- **The "Gap":** Assistant A records **only** User A. Assistant B records **only** User B. Each assistant cannot hear the other user.

- **The Problem:** When User B says "Let's meet at *Joe's Pizza*," Assistant A hears silence, followed by User A saying "Okay." Assistant A therefore misses the location entity.

- **The Goal:** Generate the full conversation along with the specific **Protocol Queries** required to resolve the missing contextual gaps.

2. Strict Constraints

- **Length:** Approximately 2,000 words (approximately 15 minutes of dialogue).

- **High Density Requirement:** Include at least 15 local context resolutions and 8 inter-agent queries.

  Required Resolution Types (Critical):

  - **Spatial** references (e.g., here, there, this place, that building, the café, that intersection, etc.)
  - **Temporal** references (e.g., now, then, later, today, tomorrow, at that time, etc.)

- **Query Quality Filter (Critical for Training):** Generate two types of potential protocol queries:

  - **HIGH_VALUE Queries** (8+ queries): Queries that involve missing information that is either:
    - Actionable (e.g., time, location, contact name)
    - Sensitive (e.g., privacy-related information)
  - **LOW_VALUE Queries** (3–5 queries): Queries that involve missing information that is not important, such as:
    - Jokes
    - Weather comments
    - Casual observations
    - Opinions or small talk
  - Each generated query must include the field:

    `query_quality_check: "HIGH_VALUE"`

    or

    `query_quality_check: "LOW_VALUE"`
  - **Purpose:** The LOW_VALUE queries are used to evaluate whether the Information Gap Detection system can correctly reject irrelevant or non-actionable queries.

A.3.1 Generation steps

Step 1: The World State

Define a complex scenario (e.g., organizing a

protest, planning a secret surprise, managing a crisis).

Step 2: The Metadata

Generate valid JSON metadata for **both** users (e.g., GPS, Wi-Fi, calendar) that explains the local context. For example, specify why a spatial reference such as "here" corresponds to a specific location (e.g., "Central Park").

Step 3: The Conversation

Write the dialogue while adhering to the following requirements:

- **Force Ambiguity:** Users should rely on implicit context (e.g., "I'm *here*", "Did you ask *him*?")
- **Force Gaps:** User B should propose plans or introduce names, while User A should agree or respond without explicitly repeating the missing entity.
- **Spatial References (Required):** Include expressions such as "here", "there", "this place", "that building", "the gym", and "that corner".
- **Temporal References (Required):** Include expressions such as "now", "then", "later", "at that time", and "tomorrow morning".
- **Person References (Required):** Include expressions such as "him", "her", "them", "that person", and "the coordinator".
- **Object References (Required):** Include expressions such as "this", "that", "it", "the package", and "that document".

Step 4: The Answer Keys (Protocol Outputs)

- **Resolutions:** Map ambiguous references to explicit contextual entities (e.g., map "here" → GPS coordinates).
- **Queries:** Generate structured protocol payloads for retrieving missing contextual information.

A.3.2 Output Format

```
{ "dataset_id": "scenario_protocol_***",
"backstory": {
"summary": "...",
"relationship": "Colleague | Spouse | Doctor"
},
"mobile_context_snapshot": {
"user_a": {
"location_semantic": "...",
"gps_coords": "...",
"wifi_ssid": "...",
"calendar_next": "..."
},
"user_b": {
"location_semantic": "...",
"wifi_ssid": "..."
}
},

"conversation_transcript": [
{ "turn_id": 1,
"speaker": "User A",
"text": "..." },
{ "turn_id": 2,
"speaker": "User B",
"text": "..." }
],

"ground_truth_resolutions": [
{ "trigger_turn_id": 1,
"ambiguous_phrase": "...",
"resolved_entity": "...",
"resolution_source": "User A GPS | User A
Wifi | User A Calendar" }
],

"required_protocol_queries": [
{
"trigger_turn_id": 3,
"query_quality_check": "LOW | MEDIUM | HIGH",
"reason": "...",
"protocol_payload": {
"intent": "...",
"target_slot / target_attribute": "...",
"urgency": "NONE | ROUTINE | IMMEDIATE"
},
"natural_language_fallback": "..."
}
] }
```

A.3.3 Synthetic dialogue

- **Doctor–patient interaction dialogue**
  - **Turn 1 — User A:** Hi, Dr. Sharma, I'm here for my appointment.
  - **Turn 2 — User B:** Hello, Samantha, please have a seat. How are you feeling now compared to the weekend?
  - **Turn 3 — User A:** I still have a headache, but it's less intense than it was.

- **Turn 4 — User B:** Did you notice the dizziness again after you were at that trail?
- **Turn 5 — User A:** Yes, right after I finished, especially when I stood up by the benches.
- **Turn 6 — User B:** And how about now?
- **Turn 7 — User A:** Today, the dizziness is mild. I've been working from my studio, so mostly sitting.
- **Turn 8 — User B:** Were you able to get enough sleep last night?
- **Turn 9 — User A:** Not really, I woke up around 3 AM and felt the headache pounding.
- **Turn 10 — User B:** Did anything help relieve it then?
- **Turn 11 — User A:** I drank some water and tried breathing exercises, but the pain just eased a little.
- **Turn 12 — User B:** Did you use any medication this time?
- **Turn 13 — User A:** No, I didn't want to overdo it on painkillers. I kept some Advil in my bag, just in case.
- **Turn 14 — User B:** Good, moderation is important. Can you point to which side you feel the pain more now?
- **Turn 15 — User A:** It's mostly behind my right eye, radiating to my temple.
- **Turn 16 — User B:** When you say 'that pain', does it feel sharp or dull?
- **Turn 17 — User A:** Kind of a throbbing ache, not really stabbing.
- **Turn 18 — User B:** Okay. Last time, you said you had light sensitivity. Is that happening now, here in the office?
- **Turn 19 — User A:** Not as much now, but at the trail and in my studio yesterday, the sunlight was almost unbearable.
- **Turn 20 — User B:** Can you tell me how long 'it' lasted yesterday after work?
- **Turn 21 — User A:** The headache lasted until around 8 PM, so several hours after I left.
- **Turn 22 — User B:** Did you stay hydrated throughout your ride and afterward?
- **Turn 23 — User A:** I tried, but I might've slacked a bit. I finished one bottle on the trail, and lost the second in the parking lot.
- **Turn 24 — User B:** Was that the usual route you took or something new?
- **Turn 25 — User A:** It was a new loop—I went through Barton Creek, took the creek crossing, and ended up by the old railroad bridge.
- **Turn 26 — User B:** Were there any moments when you felt you needed to stop or rest more than usual?
- **Turn 27 — User A:** Yes, especially at that shady spot beneath the large oak. I took five minutes then.
- **Turn 28 — User B:** And when you noticed 'it' starting, was it before or after you met up with your biking group?
- **Turn 29 — User A:** After. I joined them late at the far trailhead.
- **Turn 30 — User B:** Have you had any other episodes like this, say, at work or during a normal day?
- **Turn 31 — User A:** Yes, once last Thursday, when I was finishing a large project. The headache came on suddenly.
- **Turn 32 — User B:** Was it similar in intensity to what you felt on the trail?
- **Turn 33 — User A:** Less intense at work, but still distracting. I had to take a break and step outside.
- **Turn 34 — User B:** Who was around at the office when that happened?
- **Turn 35 — User A:** Just my coworker Alex. She offered me some tea.
- **Turn 36 — User B:** And that seemed to help?
- **Turn 37 — User A:** Only a little, but I appreciated her help.
- **Turn 38 — User B:** Earlier, you mentioned 'that folder' had stress-related items. Can you bring it next time?
- **Turn 39 — User A:** Of course, it's in my laptop bag.
- **Turn 40 — User B:** Let's talk about the testing. Your blood test is scheduled at the lab today, correct?
- **Turn 41 — User A:** Yes, I'll head there after this.
- **Turn 42 — User B:** Was the appointment for 2:00 PM or later?

- **Turn 43 — User A:** 2:00 PM at Texas Health Lab.
- **Turn 44 — User B:** And for the MRI, you have a slot at that center I recommended, right?
- **Turn 45 — User A:** Yes, it's booked for tomorrow, 4:30 PM at Medical Imaging Center.
- **Turn 46 — User B:** Good. After the results come in, let's touch base here or over the phone.
- **Turn 47 — User A:** Sounds great, I'll call as soon as I get them.
- **Turn 48 — User B:** If you notice 'it' worsening, or if there's any new symptom, contact me immediately.
- **Turn 49 — User A:** Absolutely. Thanks for listening and explaining everything.
- **Turn 50 — User B:** Of course. By the way, which bike did you use for the ride?
- **Turn 51 — User A:** My teal Trek. The old one's still in the shop.
- **Turn 52 — User B:** I see. Is 'it' heavier than the last?
- **Turn 53 — User A:** About the same, but it's got a new hydration system built in.
- **Turn 54 — User B:** That should help. How long did you ride on that last loop?
- **Turn 55 — User A:** Almost two hours, with a few stops.
- **Turn 56 — User B:** Did the terrain seem tougher than usual there?
- **Turn 57 — User A:** A bit. The creek crossing was muddy, and getting over that bridge took extra energy.
- **Turn 58 — User B:** Were you riding alone when you felt the worst of it?
- **Turn 59 — User A:** Yes, for a few miles before the others caught up.
- **Turn 60 — User B:** Okay. Let's get you ready for the tests. If you need to reschedule, call the front desk or let me know directly.
- **Turn 61 — User A:** Will do, thanks again.
- **Turn 62 — User B:** No problem. See you soon, and take care until then.
- **Turn 63 — User A:** Thank you, Dr. Sharma.
- **Turn 64 — User B:** You're welcome, Samantha. Enjoy the rest of your day.
- **Turn 65 — User A:** (smiling) You too. Hopefully the sun won't be so brutal later.
- **Turn 66 — User B:** (chuckles) Austin weather, can't predict it.

• **Labelled context resolution**

- **Turn 1** — ambiguous_phrase: "here"; resolved_entity: "Dr. Sharma's Clinic, West Avenue, Austin"; resolution_source: "User B GPS + Wifi".
- **Turn 4** — ambiguous_phrase: "that trail"; resolved_entity: "Barton Creek Greenbelt"; resolution_source: "User A Calendar".
- **Turn 5** — ambiguous_phrase: "benches"; resolved_entity: "Picnic Benches near Creek Crossing, Barton Creek Greenbelt"; resolution_source: "User A GPS".
- **Turn 6** — ambiguous_phrase: "now"; resolved_entity: "10:30 AM (Time of Appointment)"; resolution_source: "User A Calendar".
- **Turn 7** — ambiguous_phrase: "studio"; resolved_entity: "Design Studio, East 6th Street, Austin"; resolution_source: "User A GPS + Wifi".
- **Turn 9** — ambiguous_phrase: "3 AM"; resolved_entity: "03:00 AM, User A's Home, Holly Street"; resolution_source: "User A Calendar".
- **Turn 11** — ambiguous_phrase: "water"; resolved_entity: "Filtered Water Bottle, 20 oz"; resolution_source: "User A Object Log".
- **Turn 13** — ambiguous_phrase: "Advil in my bag"; resolved_entity: "Ibuprofen (Advil), 200 mg tablet, Laptop Bag pocket"; resolution_source: "User A Object Log".
- **Turn 14** — ambiguous_phrase: "which side"; resolved_entity: "Right eye, right temple"; resolution_source: "User A Self-report".
- **Turn 15** — ambiguous_phrase: "my right eye"; resolved_entity: "Orbital area, right side"; resolution_source: "User A Self-report".

- **Turn 16** — ambiguous_phrase: "that pain"; resolved_entity: "Headache radiating from right temple, throbbing ache"; resolution_source: "User A Symptom Log".
- **Turn 18** — ambiguous_phrase: "here in the office"; resolved_entity: "Dr. Sharma's Clinic, Patient Room 2"; resolution_source: "User B GPS + Calendar".
- **Turn 19** — ambiguous_phrase: "the trail"; resolved_entity: "Barton Creek Greenbelt"; resolution_source: "User A Calendar".
- **Turn 19** — ambiguous_phrase: "my studio"; resolved_entity: "Design Studio, East 6th Street, Austin"; resolution_source: "User A GPS + Wifi".
- **Turn 20** — ambiguous_phrase: "it"; resolved_entity: "Headache episode"; resolution_source: "User A Symptom Log".
- **Turn 23** — ambiguous_phrase: "one bottle on the trail"; resolved_entity: "Hydration bottle, 20 oz, Barton Creek Greenbelt"; resolution_source: "User A Object Log + GPS".
- **Turn 23** — ambiguous_phrase: "second in the parking lot"; resolved_entity: "Lost water bottle, Parking lot near trailhead"; resolution_source: "User A GPS".
- **Turn 25** — ambiguous_phrase: "new loop"; resolved_entity: "Barton Creek loop, creek crossing, railroad bridge"; resolution_source: "User A Calendar + GPS".
- **Turn 27** — ambiguous_phrase: "shady spot beneath the large oak"; resolved_entity: "Large oak tree rest area, Barton Creek Greenbelt"; resolution_source: "User A GPS".
- **Turn 28** — ambiguous_phrase: "your biking group"; resolved_entity: "Samantha's biking friends: Alex, Jamie, Lee"; resolution_source: "User A Calendar".
- **Turn 29** — ambiguous_phrase: "far trailhead"; resolved_entity: "South Entrance, Barton Creek Greenbelt"; resolution_source: "User A GPS".
- **Turn 34** — ambiguous_phrase: "office"; resolved_entity: "Design Studio, East 6th Street, Austin"; resolution_source: "User A GPS + Wifi".
- **Turn 35** — ambiguous_phrase: "my coworker Alex"; resolved_entity: "Alex N., Graphic Designer, Samantha's colleague"; resolution_source: "User A Calendar".
- **Turn 38** — ambiguous_phrase: "that folder"; resolved_entity: "Stress Tracking Folder, in Laptop Bag"; resolution_source: "User A Object Log".
- **Turn 39** — ambiguous_phrase: "my laptop bag"; resolved_entity: "Samantha's black laptop bag (main compartment)"; resolution_source: "User A Object Log".
- **Turn 40** — ambiguous_phrase: "lab today"; resolved_entity: "Texas Health Lab, 2:00 PM, June 12, 2024"; resolution_source: "User A Calendar".
- **Turn 42** — ambiguous_phrase: "appointment"; resolved_entity: "Blood Test, Texas Health Lab, 2:00 PM"; resolution_source: "User A Calendar".
- **Turn 44** — ambiguous_phrase: "that center I recommended"; resolved_entity: "Medical Imaging Center, West Campus, Austin"; resolution_source: "User B Calendar".
- **Turn 45** — ambiguous_phrase: "tomorrow, 4:30 PM"; resolved_entity: "June 13, 2024, 4:30 PM"; resolution_source: "User A Calendar".
- **Turn 46** — ambiguous_phrase: "here"; resolved_entity: "Dr. Sharma's Clinic, West Avenue, Austin"; resolution_source: "User B GPS + Wifi".
- **Turn 51** — ambiguous_phrase: "My teal Trek"; resolved_entity: "Trek bicycle, teal color, used for ride"; resolution_source: "User A Object Log".
- **Turn 52** — ambiguous_phrase: "it"; resolved_entity: "Trek bicycle"; resolution_source: "User A Object Log".
- **Turn 53** — ambiguous_phrase: "new hydration system"; resolved_entity: "Integrated hydration reservoir in Trek bike"; resolution_source: "User A Object Log".
- **Turn 56** — ambiguous_phrase: "there"; resolved_entity: "Barton Creek, creek

crossing area"; resolution_source: "User A GPS".

  - **Turn 57** — ambiguous_phrase: "that bridge"; resolved_entity: "Old railroad bridge, Barton Creek Greenbelt"; resolution_source: "User A GPS".

- **Queries**
  - **Turn 5** — query_quality_check: "HIGH_VALUE"; reason: "User A referenced the location ('benches') where dizziness was notable, but the specific location is missing for context."; protocol_payload: intent: "RESOLVE_MISSING_ENTITY", target_slot: "SYMPTOM_LOCATION", urgency: "IMMEDIATE", context_ref: "Turn 5"; natural_language_fallback: "Requesting precise location for dizziness episode mentioned in Turn 5 ('benches')."
  - **Turn 28** — query_quality_check: "HIGH_VALUE"; reason: "User B referenced User A's biking group ('your biking group') which includes individual identities relevant to history and risk assessment."; protocol_payload: intent: "RESOLVE_MISSING_ENTITY", target_slot: "PERSON_GROUP_LIST", urgency: "ROUTINE", context_ref: "Turn 28"; natural_language_fallback: "Requesting list of members present in User A's biking group from Turn 28."
  - **Turn 38** — query_quality_check: "HIGH_VALUE"; reason: "User B refers to 'that folder' with stress-related items, which could contain actionable health information."; protocol_payload: intent: "RESOLVE_MISSING_ENTITY", target_slot: "OBJECT_DOCUMENT", urgency: "ROUTINE", context_ref: "Turn 38"; natural_language_fallback: "Requesting details on 'stress tracking folder' referenced in Turn 38."
  - **Turn 44** — query_quality_check: "HIGH_VALUE"; reason: "User B references the MRI location ('that center I recommended'), but User A agent may miss the exact entity."; protocol_payload: intent: "RESOLVE_MISSING_ENTITY", target_slot: "LOCATION_DESTINATION", urgency: "IMMEDIATE", context_ref: "Turn 44"; natural_language_fallback: "Requesting MRI center name from Turn 44."
  - **Turn 45** — query_quality_check: "HIGH_VALUE"; reason: "User A mentions MRI time ('tomorrow, 4:30 PM') but User B agent may not have received exact time."; protocol_payload: intent: "RESOLVE_MISSING_ENTITY", target_slot: "APPOINTMENT_TIME", urgency: "IMMEDIATE", context_ref: "Turn 45"; natural_language_fallback: "Requesting confirmation of MRI appointment time from Turn 45."
  - **Turn 50** — query_quality_check: "HIGH_VALUE"; reason: "User B asks about which bike was used ('which bike'), relevant for health implications (hydration system, weight)."; protocol_payload: intent: "RESOLVE_MISSING_ENTITY", target_slot: "OBJECT_EQUIPMENT", urgency: "ROUTINE", context_ref: "Turn 50"; natural_language_fallback: "Requesting specifics about the bike used on recent ride from Turn 50."
  - **Turn 53** — query_quality_check: "HIGH_VALUE"; reason: "User A mentions a 'new hydration system,' relevant for hydration management."; protocol_payload: intent: "RESOLVE_MISSING_ENTITY", target_slot: "OBJECT_FEATURE", urgency: "ROUTINE", context_ref: "Turn 53"; natural_language_fallback: "Requesting details about the bike's hydration system mentioned in Turn 53."
  - **Turn 48** — query_quality_check: "HIGH_VALUE"; reason: "User B instructs User A to contact immediately if symptoms worsen; this is actionable, requiring symptom monitoring."; protocol_payload: intent: "RESOLVE_MISSING_ENTITY", target_slot: "SYMPTOM_ESCALATION_POLICY", urgency: "IMMEDIATE", context_ref: "Turn 48"; natural_language_fallback: "Requesting clarification on escalation

procedure for worsening symptoms from Turn 48."

- **Turn 19** — query_quality_check: "LOW_VALUE"; reason: "User A mentions sunlight being 'almost unbearable' as a casual observation related to light sensitivity. Not actionable itself."; protocol_payload: intent: "RESOLVE_MISSING_ENTITY", target_slot: "CASUAL_OBSERVATION", urgency: "NONE", context_ref: "Turn 19"; natural_language_fallback: "User A made a non-actionable remark about sunlight; filter should reject query."
- **Turn 36** — query_quality_check: "LOW_VALUE"; reason: "User B asks if tea helped; this is not critical, more conversational."; protocol_payload: intent: "RESOLVE_MISSING_ENTITY", target_slot: "CASUAL_COMMENT", urgency: "NONE", context_ref: "Turn 36"; natural_language_fallback: "Casual comment about coworker's tea; not actionable."
- **Turn 65** — query_quality_check: "LOW_VALUE"; reason: "User A jokes about hoping the sun isn't brutal. This is small talk."; protocol_payload: intent: "RESOLVE_MISSING_ENTITY", target_slot: "CASUAL_JOKE", urgency: "NONE", context_ref: "Turn 65"; natural_language_fallback: "User A made a weather joke in Turn 65. No follow-up query needed (testing filter)."
- **Turn 66** — query_quality_check: "LOW_VALUE"; reason: "User B jokes about Austin weather; this is not actionable."; protocol_payload: intent: "RESOLVE_MISSING_ENTITY", target_slot: "CASUAL_JOKE", urgency: "NONE", context_ref: "Turn 66"; natural_language_fallback: "Casual joke about unpredictable weather. Should be filtered out."

B Dataset statistics

Table 3 summarizes the different interactions included in the dataset across high-stakes, workplace, academic, and informal conversations. The generated interactions have median conversation lengths ranging from 14.6 to 22.7 minutes, and median turn counts between 48.5 and 75.5. This diversity enables evaluation of *Lexicon* under both structured task-oriented exchanges and open-ended social interactions.

C Reasoning based baselines

C.1 Chain of Thought (CoT)

In the Chain-of-Thought (CoT) prompting paradigm, the LLM is encouraged to generate intermediate reasoning steps before producing the final output. This approach improves performance on tasks requiring multi-step reasoning, contextual inference, and structured decision-making. By explicitly modelling intermediate reasoning, CoT helps the model decompose complex problems into smaller, logically connected sub-problems, often resulting in improved accuracy and interpretability.

■ **Sample prompt for CoT:** You are an expert at resolving ambiguous references in dialogue using provided context.

IMPORTANT:

- Think step-by-step internally, but **do not reveal chain-of-thought**.
- Output **only valid JSON** (no markdown and no commentary).
- The output must be a JSON object with a key named `ground_truth_resolutions`.
- The value of `ground_truth_resolutions` must be a list of objects, each having the following fields:
- `trigger_turn_id` : integer
- `ambiguous_phrase` : string
- `resolved_entity` : string
- `resolution_source` : string

C.2 Chain of Thought -Self-Reflexion (CoT-SR)

In the Chain-of-Thought with Self-Reflection (CoT-SR) paradigm, the LLM first generates intermediate reasoning steps and subsequently evaluates and revises its own reasoning before producing the final output. This iterative process encourages the model to identify potential errors, reconsider assumptions, and refine its solution.

■ **Sample prompt:** You are an expert at resolving ambiguous references in dialogue using the provided contextual information.

Critical Requirements:

Category	Interaction	Turns Med (Std)	Topics	Dur. (min) Med (Std)	Act. (%)
High-st.	Dr.-patient	51.0 (8.9)	Headaches, Diabetes, Surgery	15.3 (2.7)	69.3
	Law.-client	56.0 (8.7)	Contract, Payment, Property	16.8 (2.6)	78.3
Work	Colleague	55.0 (8.5)	Tasks, Exp., Manuscript	16.5 (2.5)	82.4
	Client-dev	52.5 (2.7)	Dashboard, Bugs, UAT	15.8 (0.8)	58.3
	Mgr.-emp.	48.5 (2.5)	Deliverables, Budget, Clients	14.6 (0.8)	60.0
Academic	Teach-stud	49.5 (17.3)	Density, Evolution, Circuits	14.9 (5.2)	52.6
	TA-student	50.0 (6.3)	Chem., Research, Prog.	15.0 (1.9)	66.7
Informal	Friends	75.5 (12.6)	Movies, Vacation, Games	22.7 (3.8)	36.4
	Housemate	56.5 (14.9)	Cleaning, Bills, Groceries	17.0 (4.5)	45.6

Table 3: Dataset statistics for the Lexicon framework. High-st.: High-stakes and Act.: Actionable Items.

- Do **not** reveal chain-of-thought or step-by-step reasoning.
- Output **only valid JSON** (no markdown formatting or additional commentary).
- The output must be a JSON object containing a key named `ground_truth_resolutions`.
- The value of `ground_truth_resolutions` must be a list of objects, where each object strictly follows the structure below:
- `trigger_turn_id`: Integer indicating the dialogue turn where ambiguity occurs.
- `ambiguous_phrase`: The ambiguous reference identified in the dialogue.
- `resolved_entity`: The explicit entity or concept to which the ambiguous phrase refers.
- `resolution_source`: The information source used to resolve the ambiguity.
- `rationale`: A concise justification supporting the resolution.
- `self_reflection`:
  - `confidence`: One of {low, medium, high}.
  - `possible_failure_modes`: Potential situations where the resolution may be incorrect.
  - `what_would_change_my_mind`: Additional evidence that could alter the resolution.

Guidelines for `rationale`:

- Limit the explanation to 1–2 sentences.
- Reference the most relevant supporting evidence, such as a nearby dialogue turn or contextual metadata.
- Do not include multi-step reasoning.

C.3 Tree of Thought (ToT)

Tree-of-Thought allows the model to explore multiple reasoning paths and select among them before producing a final output.

■ **Sample prompt: System Instructions:**

You are an expert at resolving ambiguous references in dialogue using the provided contextual information.

Critical Requirements:

- Do **not** reveal chain-of-thought or step-by-step reasoning.
- Output **only valid JSON** (no markdown formatting or additional commentary).
- The output must be a JSON object containing a key named `ground_truth_resolutions`.
- The value of `ground_truth_resolutions` must be a list of objects, where each object strictly follows the structure below:
- `trigger_turn_id`: Integer indicating the dialogue turn where ambiguity occurs.
- `ambiguous_phrase`: The ambiguous reference identified in the dialogue.
- `resolved_entity`: The explicit entity or concept to which the ambiguous phrase refers.

Framework Component	Evaluation Metric	Semi-Synthetic *Lexicon*	Human-Annotated *Lexicon*
Local Reference Resolution	True Positive Rate (TPR % ↑)	72.6 (−5.7)	70.8 (−1.8)
	Sentence Similarity (% ↑)	55.0 (−3.1)	50.3 (−2.5)
Info Gap Detection	True Positive Rate (TPR % ↑)	85.3 (−6.1)	83.9 (−4.6)
	False Positive Rate (FPR % ↓)	13.4 (+6.6)	15.6 (+5.3)
Relationship Assessment	Classification Accuracy (% ↑)	95.4 (−1.0)	97.0 (−1.0)
Hybrid Decision Gate	True Negative Rate (TNR % ↑)	98.5 (+1.5)	95.7 (+1.9)
	True Positive Rate (TPR % ↑)	81.5 (−4.5)	80.1 (−8.3)

Table 4: End-to-end performance of *Lexicon* under full pipeline evaluation. Values in parentheses denote absolute change compared to isolated component evaluation.

- `resolution_source`: The information source used to resolve the ambiguity.
- `rationale`: A concise justification supporting the resolution.

Tree-of-Thought Reasoning Procedure:

1. Generate multiple candidate interpretations of ambiguous references using diverse sampling configurations.
2. Evaluate each candidate using a structured critic model that assesses correctness, grounding, and schema compliance.
3. Select the highest-scoring candidate interpretation.
4. Optionally refine the selected candidate to improve clarity, correctness, and consistency while preserving schema requirements.

Guidelines for `rationale`:

- Limit the explanation to 1–2 sentences.
- Reference the most relevant supporting evidence, such as a nearby dialogue turn or contextual metadata.
- Do not include multi-step reasoning.

D Evaluation

***Lexicon* End-to-End Evaluation:** Table 4 shows the end-to-end evaluation of *Lexicon* across both semi-synthetic and human-annotated datasets. Overall, the framework maintains strong performance under full pipeline integration, with only moderate performance changes compared to isolated component evaluation.