

EMSDialog: Synthetic Multi-person Emergency Medical Service Dialogue Generation from Electronic Patient Care Reports via Multi-LLM Agents

Xueren Ge¹, Sahil Murtaza¹, Anthony Cortez², Homa Alemzadeh¹

¹School of Engineering and Applied Sciences, University of Virginia, Charlottesville, VA, USA

²School of Medicine, University of Virginia, Charlottesville, VA, USA

{zar8jw, vpn9ej, aec3gp, ha4d}@virginia.edu

Abstract

Conversational diagnosis prediction requires models to track evolving evidence in streaming clinical conversations and decide when to commit to a diagnosis. Existing medical dialogue corpora are largely dyadic or lack the multi-party workflow and annotations needed for this setting. We introduce an ePCR-grounded, topic-flow-based multi-agent generation pipeline that iteratively plans, generates, and self-refines dialogues with rule-based factual and topic flow checks. The pipeline yields *EMSDialog*, a dataset of 4,414 synthetic multi-speaker EMS conversations based on a real-world ePCR dataset, annotated with 43 diagnoses, speaker roles, and turn-level topics. Human and LLM evaluations confirm high quality and realism of *EMSDialog* using both utterance- and conversation-level metrics. Results show that *EMSDialog*-augmented training improves accuracy, timeliness, and stability of EMS conversational diagnosis prediction. Our datasets and code are publicly available at <https://uva-dsa.github.io/EMSDialog>

1 Introduction

Conversational diagnosis prediction aims to infer a patient’s likely condition during clinical conversations, issuing early yet reliable diagnosis that can guide time-critical actions (e.g., airway management, glucose checks). This is a vital capability for *conversational diagnostic agents* that either *replace* doctors by conducting interactive QA (Fan et al., 2025; Chen et al., 2025; Tu et al., 2025; Luo et al., 2026) or *assist* doctors by suggesting real-time diagnoses (Weerasinghe et al., 2024).

Unlike Electronic Health Record (EHR)-based diagnosis prediction (Rios and Kavuluru, 2018; Ge et al., 2024) where full EHR data is often used as input, conversational diagnosis prediction models must reason over *incomplete streaming information*, update predictions turn by turn, and decide

EMS Dialogue	32B 0-shot	4B train
EMT1: Hi Sir, My name is Andy. Can you tell me what's going on?	Seizure, 0.8	Defer
Have really bad chest pain: Patient	Chest Pain, 0.75	Defer
EMT1: Okav, do you mind if we take a look at you real quick?	Stroke, 0.8	Defer
No Please do: Patient	Stroke, 0.8	Defer
EMT1: I'm feeling a radial pulse	Diabetic, 0.8	Chest Pain, 0.58
EMT2: Can you take your shirts off?	Defer	Chest Pain, 0.72
...	Defer	...
EMT1: Where in your chest hurts?	Defer	CardiacArrest, 0.93
middle, radiate to left arm : Patient	Defer	CardiacArrest, 0.94
EMT2: Heart rate is 100.	CardiacArrest, 0.9	CardiacArrest, 0.96
...

Figure 1: Qwen3-series model performance on conversational diagnosis prediction for cardiac arrest scenario

when to commit versus defer on a final prediction (Sun et al., 2025).

Training and evaluating models for this task demands realistic medical conversation data that reflects how care actually happens along with diagnosis annotations. Medical conversations are inherently *multi-party*, with information being elicited, relayed, and verified across several speakers with distinct roles, topics, and access to context. Emergency Medical Services (EMS) (Weerasinghe et al., 2025) exemplifies this setting, where medics coordinate with partners and dispatch while interacting with the patient and bystanders, each contributing role-specific evidence (e.g., scene safety, chief complaint, medications, last-known-well).

However, existing medical corpora fall short for this setting: (i) *Online doctor-patient dialogue* datasets (mostly from Chinese online-forums) (Wei et al., 2018; Xu et al., 2019; Zhang et al., 2020; Zhou et al., 2021; Liu et al., 2022; Li et al., 2021; Shi et al., 2023; Zeng et al., 2020) are typically asynchronous, dyadic, diverging from realistic real-time operational workflows and offering limited diagnosis labels; (ii) *EHR-grounded human role-play* datasets (Abacha et al., 2023; Papadopoulos Korfiatis et al., 2022; Saley et al., 2024) improve case realism but still assume two speakers and seldom provide diagnosis annotations; and (iii) *Synthetic dialogue* datasets generated by rules (Fansi Tchango et al., 2022; Seo and Lee, 2024) or LLMs (Wang

Dataset	Data characteristics					Generation method				Data Use		
	Setting	Lang.	#Roles	#Dx	Topic #U/D	Gen.	Src.	Checker	Refiner	Tasks		
<i>Real</i>	Dxy (Xu et al., 2019)	Telehealth	cn	2	5	✗	–	Human	Web	–	–	Diagnosis Pred
	MedDialog (Zeng et al., 2020)	Telehealth	cn,en	2	96	✗	3.3	Human	Web	–	–	Generation
	MedDG (Liu et al., 2022)	Telehealth	cn	2	12	✗	21.6	Human	Web	–	–	NER
	MidMed (Shi et al., 2023)	Telehealth	cn	2	4	✓	11.8	Human	Web	–	–	Generation
	MTS-Dialogue (Abacha et al., 2023)	Clinical	en	2	–	✗	9.0	Human	Roleplay	–	–	Note2Dialog
	MediTOD (Saley et al., 2024)	Clinical	en	2	–	✓	95.6	Human	–	–	–	Generation
	EMSAudio (Weerasinghe et al., 2024)	EMS	en	3.7	7	✗	54.5	Human	Roleplay	–	–	Audio2Text
	EgoEMS (Weerasinghe et al., 2025)	EMS	en	4.1	3	✗	128.5	Human	Roleplay	–	–	Activity Recognition
<i>Synthetic</i>	DDXPlus (Fanshi Tchango et al., 2022)	Clinical	en	2	49	✗	33.1	Rule	–	–	–	Diagnosis Pred
	NoteChat (Wang et al., 2024)	Clinical	en	2	–	✗	62.5	LLM	EHR	Style	✓	Note2Dialog
	DiagESC (Seo and Lee, 2024)	Mental	en	2	5	✗	–	LLM	PHQ-9	–	–	Generation
	HQMedical (Ge et al., 2025)	Clinical	cn	2	–	✓	54.0	LLM	EHR	Style	✓	Note2Dialog
	MedSynth (Mianroodi et al., 2025)	Clinical	en	2	2001	✗	55.0	LLM	EHR	Style	✗	Note2Dialog
	EMSDialog (ours)	EMS	en	5.2	43	✓	114.3	LLM	EHR	Topic, Fact, Style	✓	–

Table 1: Comparison of real vs. synthetic medical multi-turn dialogue datasets. #Roles = Average number of speaker roles per dialogue. #Dx = Number of diagnosis classes. #U/D = Average number of utterances per dialogue.

et al., 2024) although easily scalable, often neglect realistic dialogue topic flow, multi-party settings, and rich annotations required for downstream tasks such as conversational diagnosis prediction. These gaps motivate creating **multi-speaker**, operationally **realistic** datasets with **task-aligned diagnosis annotations**, to train and evaluate models on not only *what* to predict, but also *when* to commit.

On the other hand, off-the-shelf zero-shot LLMs, although a plausible choice as conversational diagnostic agents (Li et al., 2024; Tu et al., 2025; Ge et al., 2026; Zhang et al., 2026a), often fail to produce reliable, stable predictions in dynamic, turn-by-turn conversational settings (Laban et al., 2025; Li et al., 2025). In particular, as shown in Figure 1, they (i) make incorrect, early, highly confident guesses when evidence is still sparse, and (ii) exhibit high prediction volatility, frequently switching their outputs as new information arrives instead of converging to a consistent diagnosis.

To address these gaps, we propose a multi-agent synthetic dialogue generation pipeline for creating realistic, multi-speaker medical conversations grounded in real-world clinical records. We leverage the pipeline to generate a synthetic EMS dialogue dataset and use it to train and evaluate models for conversational diagnosis prediction. Our contributions are threefold:

- We propose a scalable, EHR-grounded, multi-agent pipeline for synthetic multi-party dialogue generation that enforces clinical factuality, procedural realism, and natural dialogue style through an **iterative critique-and-refine** loop driven by a hybrid suite of deterministic **rule-based checkers** for concepts and topic flow, and an LLM-based **style checker**.
- We introduce *EMSDialog*, an EMS-specific syn-

thetic dataset of 4,414 realistic multi-party conversations, generated from a real-world ePCR dataset and annotated with 43 diagnoses, turn-level speaker roles, and topics. Human expert and LLM-based evaluations show high quality at both the utterance level (realism, safety, role accuracy, and groundedness) and the conversation level (logical flow, factuality, and diversity).

- We demonstrate the downstream utility of *EMSDialog* by training models of varying sizes for conversational diagnosis prediction and evaluating them on real-world EMS conversations. Experiments show that *EMSDialog*-augmented training improves prediction accuracy, timeliness, and stability, and combining synthetic and real data yields the strongest overall performance.

2 Related Works

2.1 Real-world Medical Dialogue Datasets

As shown in Table 1, many publicly-available real telehealth dialogue corpora are web-crawled from Chinese online forums (Wei et al., 2018; Xu et al., 2019; Zhang et al., 2020; Zhou et al., 2021; Liu et al., 2022; Li et al., 2021; Shi et al., 2023; Zeng et al., 2020). However, they often diverge from real clinical topic flows (e.g., chief complaint, primary assessment) and provide limited supervision (e.g., symptom tags and small diagnosis sets). Datasets constructed via human role-play (Zhang et al., 2026b) grounded in EHR (Abacha et al., 2023; Papadopoulos Korfiatis et al., 2022; Saley et al., 2024) improve realism but typically lack detailed diagnosis labels, limiting use for downstream tasks such as diagnosis prediction. Although EgoEMS (Weerasinghe et al., 2025) collects small set of realistic human roleplay EMS dialogues, the data size and

label set is far too small to train a reliable model for real-world use. Moreover, most prior works assume dyadic doctor–patient chats, whereas real care—especially in EMS—involves multi-party coordination. These gaps motivate datasets with (i) multi-speaker, realistic dialogues that faithfully mirror operational care environments, and (ii) task-aligned annotations.

2.2 Synthetic Medical Dialogue Generation

Early efforts in data synthesis adopted rule-based pipelines with handcrafted templates or constraints (Fanshi Tchango et al., 2022; Seo and Lee, 2024). Recent work leverages LLMs (Li et al., 2023) for EHR-grounded dialogue generation, exploiting their fluency while adding guardrails. Single-agent approaches (Das et al., 2024) typically inject predefined rules into prompts (e.g., enforcing symptom coverage or turn structure) to steer the model toward higher-quality conversations. Alternatively, multi-agent frameworks (Wang et al., 2024; ALMutairi et al., 2024; Xu et al., 2026a) produce clinically plausible dialogues by simulating doctor–patient role-playing grounded in symptoms extracted directly from EHRs. Unlike these methods, which rely on unverified LLM-based refinement, we propose a framework that enforces factuality and realism/factuality in synthetic dialogues by integrating deterministic **rule-based verification** into a strict **repeat-until-pass refinement loop**.

2.3 Conversational Diagnosis Prediction

Previous work on *conversation forecasting* (Chang and Danescu-Niculescu-Mizil, 2019; Kementchedjhiya and Sogaard, 2021; Yuan and Singh, 2023; Zhang et al., 2025) focused on detecting impending conversational events within a transcript. However, forecasting is typically framed as binary classification with an early stop once a “derailment” is predicted. These works did not focus on continuous diagnosis prediction based on each utterance in the conversation. On the other hand, prior work on diagnosis prediction either assumes static EHR snapshots (Ma et al., 2018, 2017; Ge et al., 2024; Niset et al., 2025) or the whole conversation as input for multi-label classification (Wei et al., 2018; Xu et al., 2019). Recent work on conversational diagnostic agents uses LLMs as a replacement for doctors to conduct direct diagnostic dialogues with patients (Sun et al., 2024; Tu et al., 2025). Here we focus on conversational diagnosis prediction which is the task of *multi-label diagnosis classification*

at each utterance in a dialogue between caregivers and patient for *assisting* with timely diagnosis in emergency scenarios (Weerasinghe et al., 2024).

3 Preliminaries

3.1 Electronic Patient Care Report (ePCR)

ePCRs are the primary documentation used in EMS to record real-time patient care during emergency incidents (Rahman et al., 2020; Kim et al., 2021; Ge et al., 2024). As shown in Figure 2a, an ePCR combines structured fields, including *Chief Complaint*, *History of Present Illness* (e.g., medical history, current medications, allergies), *vital signs*, and *interventions* (procedures, medications), with a free-text medic narrative describing the encounter and *protocols* (diagnoses) used for treating patients.

3.2 EMS Topic Flow

We define the EMS topic flow based on official guidelines (National Registry of Emergency Medical Technicians, 2016; Old Dominion Emergency Medical Services Alliance, 2022). As shown in Figure 2b, EMS encounters begin with *Introduction* → *Chief Complaint* → *Response Test* (GCS) to choose the branch. If the patient is conscious, the dialogue proceeds to *Primary Assessment* (Checking A.B.C.: Airway, Breathing, Circulation.) and *Secondary Assessment*, followed by *HPI* (Checking S.A.M.P.L.E.: Signs/Symptoms, Allergies, Medications, Past medical history, Last oral intake, Events leading up to illness/injury) and *Pain Assessment* (Checking O.P.Q.R.S.T.: Onset, Provocation/Palliation, Quality, Region/Radiation, Severity, Time.). If the patient is comatose, history-taking is deferred; providers prioritize *Primary* then *Secondary Assessment*. Across both branches, *Vitals* and *Interventions* may interleave with assessments. When the case *exits to protocol*, medics determine a working diagnosis and transition from general care to protocol-specific steps. During transport, *Interventions* and *Reassessment* continue. The topic flow yields a realistic, modular structure for EMS dialogues and supports fine-grained supervision and timing decisions. Detailed topic flow definitions are in Appendix A.1.

4 Methodology

4.1 Synthetic Dialogue Generation Pipeline

As shown in Figure 2c, our synthetic data generation pipeline consists of five modules, including

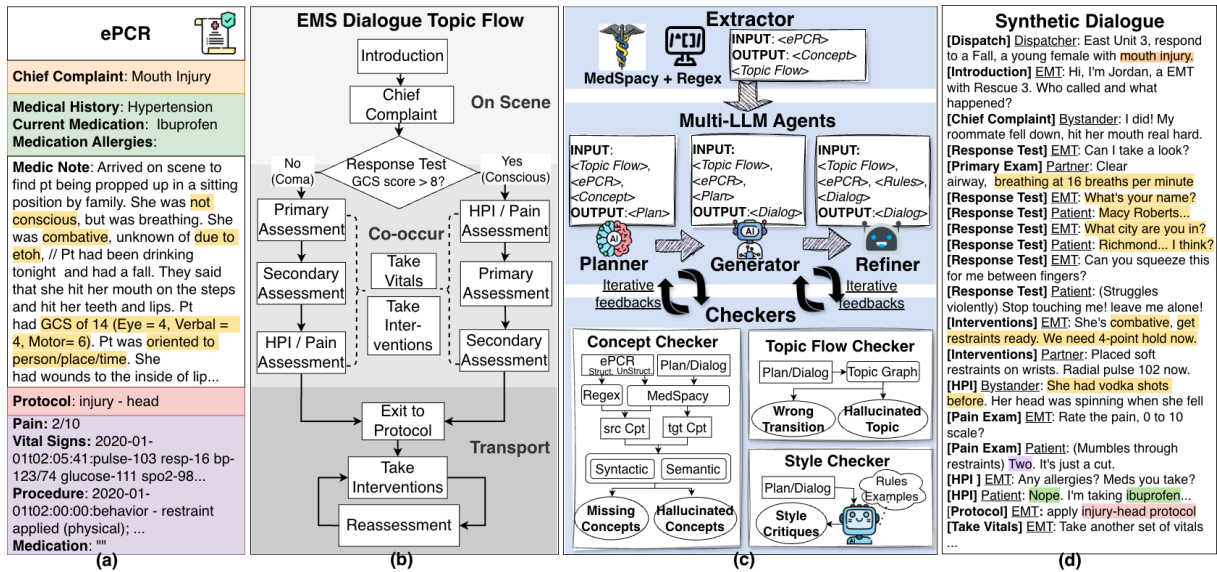


Figure 2: a) ePCR; b) EMS Topic Flow; c) Synthetic Dialogue Generation Pipeline; d) Synthetic Dialogue Example. Color-highlighted text across (a) and (d) demonstrates the factual grounding of generated dialogue concepts to the original ePCR data.

three LLM agents for topic flow planning, dialogue generation, and refinement, as described next.

4.1.1 Extractor

The first module in the pipeline extracts important medical concepts from input ePCR to preserve them through subsequent LLM stages and ensure they are explicitly incorporated in the generated dialogue. For structured data, we parse the concepts using regular expressions. For unstructured notes, we extract key EMS concepts using the state-of-the-art (SOTA) rule-based clinical Named Entity Recognition (NER) toolkit, MedSpaCy (Eyre et al., 2021) with QuickUMLS (Soldaini and Goharian, 2016) (See more details in Appendix A.3). We also use regular expressions to extract important features (such as GCS score) which govern the topic flow. For example, if $GCS \leq 8$ (coma), EMS response is initiated with *Primary Assessment*, but if $GCS > 8$ (conscious), it begins with *HPI/Pain Assessment*.

4.1.2 Checkers

We incorporate three independent checkers to provide actionable feedback on concept errors, topic-flow violations, and style issues to the LLM agents and prompt them to self-refine the generated plan/dialogue (Madaan et al., 2023; Xu et al., 2026b). The PLANNER and GENERATOR agents iteratively revise their outputs and cannot proceed until all hard constraints by rule-based Concept and Topic Flow checkers are satisfied (e.g., no missing or hallucinated concepts; no invalid topic transi-

tions or hallucinated topics). The LLM-based Style checker proposes style critiques to the REFINER agent.

Concept Checker extracts concepts from the source ePCR and from the generated Plan/Dialogue using the same MedSpaCy concept extractor. The two concept sets are then aligned in a two-stage procedure: (i) syntactic matching, where two concepts are matched only if their surface forms are an exact string match; and (ii) semantic matching, where any remaining unmatched concepts are embedded with GatorTron (Yang et al., 2022) and paired based on cosine similarity, considering two concepts equivalent if their similarity is at least 0.8. After matching, concepts present in the ePCR but not matched in the Plan/Dialogue are labeled as *missing (FN)*, while concepts appearing only in the Plan/Dialogue are labeled as *hallucinated (FP)*.

Topic Flow Checker validates that the dialogue’s topic sequence complies with allowable topic flow rules. We define a directed graph $G = (V, E)$ over topics, represented as an adjacency list $\mathcal{A} = \{\text{current} : [\text{allowed next topics}]\}$. For each topic pair (t_i, t_{i+1}) in consecutive utterances, we check whether $t_{i+1} \in \mathcal{A}[t_i]$. The checker flags two error types: (1) *transition error* when there is no edge from t_i to t_{i+1} ; and (2) *hallucinated topic* when $t_i \notin V$ (i.e., the topic is not in the ontology).

Style Checker verifies if the dialogue style satisfies a set of domain-specific requirements (See

prompt in Appendix Figure 13), including realistic EMS dialogues (Weerasinghe et al., 2025) as exemplars and a rubric describing typical EMS dialogue patterns, authored by an EMS expert (See Appendix Figure 17). This checker acts as an LLM judge: if violations are detected, it returns structured critiques with concrete revisions based on the rubric; otherwise, it issues a pass.

4.1.3 Planner

The PLANNER LLM proposes a dialogue plan given the input ePCR, the topic flow, and extracted EMS concepts (see prompt in Appendix Figure 14). The plan is encoded as a sequence of tuples (*topic, evidence*), where evidence cites specific ePCR fields. The CHECKER validates the plan for (i) concept coverage and (ii) topic flow, and any violations are returned as structured feedback for iterative revision by the PLANNER. The plan is accepted once no concept or topic-flow errors remain. This is to enforce dialogue logical structure and ensure every turn is supported by ePCR evidence.

4.1.4 Generator

The GENERATOR LLM produces an initial dialogue draft, including utterances and their assigned roles (e.g., medic, partner, patient, bystanders) at each turn, which is explicitly grounded in the ePCR evidence cited in the plan (see prompt in Appendix Figure 15). The CHECKER evaluates the (i) concept coverage and (ii) topic flow of each utterance, returning targeted feedback to the GENERATOR and iterating until all constraints are satisfied. The goal is to produce a first-pass dialogue that strictly adheres to the validated plan and ePCR evidence.

4.1.5 Refiner

Although the generation stage yields a dialogue that adheres to the topic flow and ePCR evidence, it can read as unnatural. For example, this often occurs when the dialogue “tells” instead of “shows” behaviors: a patient is simply labeled “combative” rather than shouting or resisting; or a medic declares a patient as “oriented to place” instead of asking orientation questions like, “Where are we?”. The REFINER LLM edits the dialogue for coherent turn-taking, natural phrasing, and clinically plausible actions based on expert rubrics and exemplars, while preserving evidence grounding (see prompt in Appendix Figure 16). The CHECKER checks the (i) concept coverage, (ii) topic flows and the (iii) styles issue, returning the feedback to REFINER to

further revise the dialogue. The goal is to polish the dialogue for realism while preserving factuality. This refinement loop runs for at most 5 iterations or terminates early once the dialogue satisfies all checks. We set the maximum iteration as 5 because the mean number of iterations needed to fix concept and topic flow errors in previous “Plan” and “Generate” stages is 4.3.

5 Experiments

We conduct both intrinsic and extrinsic evaluations of our synthetic dialogue generation pipeline to answer the following research questions:

RQ1: What is the quality of synthetic multi-person dialogue data generated by the pipeline?

RQ2: How can synthetic dialogue data help with conversational diagnosis prediction?

5.1 Datasets and Baselines

ePCR: We leverage a real-world EMS ePCR dataset containing 4,417 pre-hospital reports annotated with 43 EMS protocol labels to ground our synthetic EMS dialogue generation. These records were collected from a regional ambulance agency in the U.S. between 2017 and 2020. All private information has been de-identified (Kim et al., 2021; Ge et al., 2024). More ePCR details is provided in Appendix A.4

Real-world EMS Dialogues: We split 149 de-identified real-world EMS dialogues (Weerasinghe et al., 2024, 2025) into *train* (89) and *test* (60) sets, utilizing the former as our *Real* training set.

Synthetic EMS Dialogues: We evaluate our approach against five baseline synthetic dialogue generation methods. To ensure fairness, all baselines receive the same ePCR dataset as input and each generate a distinct set of 4,417 synthetic dialogues. These baselines and our method are categorized into single-agent and multi-agent frameworks:

- **Single-Agent:**
 - *0-Shot*: Direct prompting using the ePCR.
 - *0-Shot + Rules*: Direct prompting augmented with professional EMS rules.
 - *CoT* (Wei et al., 2022): Chain-of-thought prompting using the ePCR and one real-world EMS exemplar.
 - *CoT + Rules*: CoT prompting augmented with both the exemplar and EMS rules.
- **Multi-Agent:**
 - *NoteChat* (Wang et al., 2024): The state-of-the-art medical dialogue generator, adapted by

replacing its default clinical prompts with our EMS exemplars.

- *EMSDialog (Ours)*: Our proposed pipeline integrating the ePCR, an exemplar, EMS rules, and the EMS Topic Flow constraints.

Dataset statistics are provided in Appendix A.4. All baselines are executed locally using Qwen3-32B (Yang et al., 2025) deployed on NVIDIA A100 GPUs via vLLM (Kwon et al., 2023). The complete prompts can be found in Figure 18–21.

5.2 Intrinsic Evaluation

We conduct intrinsic evaluation of the quality of generated synthetic dialogues at two levels of granularity, **conversation-level** and **utterance-level**. We evaluate 43 sample dialogues (one per diagnosis class) via manual review by a certified EMS professional, and the full set of 4,411 ePCR-derived dialogues using the “LLM-as-a-judge” (Zheng et al., 2023) method. For each input ePCR, the dialogue generated by our pipeline is compared to those generated by the baselines. To mitigate ordering bias, we randomized the presentation order of the four dialogues for both human experts and LLM judges in our experiments. Additionally, we employ two independent open-source models, Qwen3-235B (Yang et al., 2025) and Llama-3.3-70B (Dubey et al., 2024), as LLM judges to preserve privacy while maintaining evaluation quality.

Conversation-level Metrics. At the conversation level, each dialogue is evaluated via these metrics:

- **Logical Structure** (↑): Measured on a 5-point Likert scale (1–5), assessing if the dialogue follows a coherent EMS topic progression.
- **Overall Ranking** (↑): A comparative ranking of the four generated dialogues, summarized using Mean Reciprocal Rank (MRR).
- **Factuality** (↑): Measured as concept-level precision and recall (P/R) between the source ePCR and the generated dialogue, using (i) NER-extracted concepts ($\mathbf{Fac}_{\text{NER}}$) for the full dataset and (ii) human-annotated concepts for the subset of 43 manually-evaluated cases ($\mathbf{Fac}_{\text{H}^*}$).
- **Diversity** (↓): Calculated using Self-BLEU (Montahaie et al., 2019) over all dialogues generated by each method, with lower values indicating higher diversity (less self-repetition).

Utterance-level Metrics. We cast utterance-level evaluation as a binary classification task. Human and LLM judges assign a “Yes” or “No” label for each metric. Results are reported as an aggregate

“Yes” rate (%) across all evaluated utterances:

- **Realism** (↑): Whether the utterance sounds natural and is likely generated by a human.
- **Safety** (↑): For responder utterances only; whether the utterance contains actions or decisions that violate established EMS protocol guidelines (diagnosis knowledge) (Old Dominion Emergency Medical Services Alliance, 2022) or could potentially harm the patient.
- **Role Accuracy** (↑): Whether the speaker role matches the EMS dialogue context.
- **Groundedness** (↑): Whether EMS concepts in the utterance are supported (syntactically, semantically or inferably) by the associated ePCR.

More details on metrics are in Appendix A.6.

5.3 Extrinsic Evaluation

We evaluate whether synthetic dialogue data can help fine-tune SOTA LLMs for improved **conversational diagnosis prediction**. Formally, at dialogue turn t , given the accumulated transcript prefix (dialogue history) $X_t = (u_1, u_2, \dots, u_t)$, up to utterance u_t , the goal is to produce a probability $p_t(l)$ (confidence) for each diagnosis label $l \in \mathcal{L}$. At each turn t , the model must (i) update its belief over possible diagnoses and (ii) decide whether to *commit* to a final prediction or *defer* to gather more evidence. We convert model’s probabilities into discrete predictions using a fixed threshold $\tau = 0.5$, such that $\hat{Y}_t = \{l \in \mathcal{L} : p_t(l) \geq \tau\}$. If no label exceeds τ (\hat{Y}_t is empty), the model *defers* at turn t .

Training Datasets. For diagnosis prediction, we fine-tune Qwen3-0.6B/4B with LoRA (Hu et al., 2022) using three training sets: *Real-only* (89 real dialogues), *Synthetic-only* (six different sets of 4,417 synthetic dialogues), and *Real+Synthetic* (Real combined with EMSDialog). For all fine-tuning runs, we use an 80:20 train–validation split on the *train* and select the best checkpoint by validation performance.

Training Strategies. We follow two established strategies for the task of conversation forecasting: (i) *Static training* (Chang and Danescu-Niculescu-Mizil, 2019), which trains on the full dialogue prefix, learning to predict a label from the set of diagnoses based on the aggregated context; (ii) *Dynamic training* (Kementchedjheva and Sogaard, 2021), which expands each dialogue into multiple training instances by unrolling the last K prefixes (e.g., $u_{1:T}, \dots, u_{1:T-K}$), so the model is explicitly

Method	Conversation-level					Utterance-level			
	Logic (↑)	MRR (↑)	FacNER (↑)	FacH* (↑)	Diversity (↓)	Realism (↑)	Safety (↑)	Role Acc (↑)	Groundedness (↑)
	H* / LLM (1-5)	H* / LLM (%)	P / R (%)	P / R (%)	Self-BLEU (%)	H* / LLM (%)	H* / LLM (%)	H* / LLM (%)	H* / LLM (%)
0-shot	1.85 / 2.45	31.14 / 32.83	71.32 / 55.88	67.69 / 64.71	40.36	44.96 / 52.77	96.77 / 95.76	96.03 / 87.39	80.76 / 75.85
0-shot+Rules	2.91 / 3.55	48.07 / 50.58	73.58 / 54.13	69.23 / 66.88	65.73	55.68 / 63.91	96.91 / 94.88	96.74 / 95.34	82.17 / 77.12
CoT	3.05 / 3.55	37.50 / 36.67	75.76 / 58.79	68.15 / 66.09	<u>40.43</u>	78.23 / 70.38	99.85 / 95.46	98.51 / 93.55	87.08 / 85.79
CoT+Rules	<u>3.63 / 3.80</u>	<u>58.38 / 55.33</u>	78.23 / 61.77	74.91 / 68.56	61.59	<u>85.59 / 76.25</u>	99.85 / 98.21	98.87 / 95.77	92.36 / 86.33
NoteChat	1.80 / 2.43	31.25 / 39.16	<u>85.63 / 68.40</u>	<u>80.12 / 74.21</u>	68.97	35.42 / <u>78.48</u>	<u>100.00 / 97.40</u>	<u>100.00 / 98.30</u>	<u>95.23 / 87.49</u>
EMSDialog	4.25 / 4.55	87.50 / 78.17	93.70 / 73.72	91.06 / 82.80	53.08	90.93 / 90.44	100.00 / 98.64	100.00 / 99.69	99.23 / 92.83

Table 2: Intrinsic Evaluation: Conversation- and Utterance-level performance of synthetic dialogues generation methods. H*: human evaluation on a 43-scenario subset. The best and runner-up results are in **bold** and underlined.

Mode	Size	Train Data	Prompt	First Acc (↑) / Conf	Last Acc (↑) / Conf	Earliness (↑) (1st / 1st-correct)	Edit Overheads (↓)
No Train	4B	-	0-shot	37.78 \pm 0.79/88.57 \pm 0.44	60.22 \pm 0.79/93.78 \pm 0.46	93.02 \pm 0.06/80.20 \pm 1.63	83.51 \pm 0.34
			CoT	30.00 \pm 1.36/88.37 \pm 1.11	61.67 \pm 2.72/95.98 \pm 0.00	95.18 \pm 0.08/81.73 \pm 0.44	73.26 \pm 0.29
	32B	-	0-shot	63.89 \pm 0.79/88.07 \pm 0.15	80.56 \pm 3.14/94.20 \pm 0.20	92.41 \pm 0.00/84.93 \pm 0.93	57.11 \pm 0.71
			CoT	51.11 \pm 2.08/81.60 \pm 0.79	76.67 \pm 1.36/93.07 \pm 0.37	94.44 \pm 0.08/ 88.73 \pm 1.05	60.32 \pm 1.20
Static Train	4B	Real	-	58.23 \pm 4.57/67.38 \pm 1.26	63.98 \pm 1.34/78.53 \pm 1.26	90.10 \pm 2.17/80.81 \pm 2.13	69.07 \pm 1.32
		0-shot	-	68.33 \pm 2.08/69.27 \pm 2.47	70.78 \pm 0.00/85.18 \pm 3.33	81.32 \pm 1.75/82.47 \pm 2.96	61.91 \pm 2.92
		0-shot+Rules	-	70.08 \pm 1.30/70.27 \pm 2.28	70.92 \pm 2.48/89.11 \pm 3.03	79.89 \pm 3.62/79.08 \pm 4.31	65.07 \pm 3.59
		CoT	-	67.22 \pm 2.08/66.00 \pm 1.96	75.89 \pm 2.08/84.95 \pm 1.33	82.21 \pm 1.11/83.01 \pm 0.70	62.12 \pm 2.77
		CoT+Rules	-	69.80 \pm 1.84/68.21 \pm 1.05	76.00 \pm 1.03/82.74 \pm 3.54	83.57 \pm 2.24/80.41 \pm 3.08	63.39 \pm 3.66
		NoteChat	-	69.78 \pm 3.42/61.08 \pm 2.23	75.44 \pm 2.83/83.86 \pm 3.85	80.90 \pm 7.49/79.97 \pm 7.04	54.09 \pm 1.52
		EMSDialog	-	70.56 \pm 1.57/75.49 \pm 2.99	77.78 \pm 1.57/92.52 \pm 1.09	87.32 \pm 1.75/86.55 \pm 0.96	47.57 \pm 1.95
		EMSDialog+Real	-	75.25 \pm 0.67/86.62 \pm 0.48	82.05 \pm 1.69/95.64 \pm 1.19	90.46 \pm 1.69/ 88.70 \pm 1.98	35.84 \pm 0.98
Dynamic Train	4B	Real	-	59.95 \pm 3.19/70.53 \pm 2.39	67.21 \pm 1.78/89.48 \pm 1.03	90.13 \pm 0.92/81.14 \pm 1.38	58.32 \pm 0.64
		0-shot	-	70.67 \pm 0.40/76.46 \pm 2.78	72.22 \pm 2.08/92.23 \pm 1.29	86.23 \pm 3.20/85.66 \pm 0.57	46.49 \pm 0.14
		0-shot+Rules	-	69.41 \pm 3.30/70.11 \pm 2.98	74.50 \pm 1.10/90.20 \pm 1.74	78.12 \pm 4.22/77.35 \pm 3.29	44.69 \pm 1.20
		CoT	-	71.11 \pm 2.08/83.31 \pm 2.36	73.33 \pm 1.36/94.30 \pm 1.73	86.18 \pm 1.26/86.08 \pm 1.29	42.95 \pm 1.99
		CoT+Rules	-	71.40 \pm 1.27/72.76 \pm 2.94	75.93 \pm 1.76/90.15 \pm 1.32	78.76 \pm 2.62/77.81 \pm 3.77	42.51 \pm 1.53
		NoteChat	-	72.44 \pm 1.57/81.80 \pm 2.81	76.11 \pm 0.79/94.06 \pm 0.71	83.97 \pm 0.14/86.63 \pm 0.50	51.18 \pm 1.20
		EMSDialog	-	74.67 \pm 0.79/84.34 \pm 2.65	78.89 \pm 1.36/94.11 \pm 2.56	87.03 \pm 0.59/86.70 \pm 0.60	34.75 \pm 1.27
		EMSDialog+Real	-	76.84 \pm 2.68/87.06 \pm 2.92	83.56 \pm 1.34/96.36 \pm 0.19	88.96 \pm 0.13/ 89.34 \pm 0.38	31.99 \pm 2.09

Table 3: Conversational diagnosis prediction results of Qwen3-4B/32B models.

trained to make predictions from earlier, partial evidence. More details are provided in Appendix A.5

Testing. At inference time, we evaluate all models in a *dynamic* (turn-by-turn) setting. Given the growing transcript prefix, the model emits diagnoses at each turn when its confidence exceeds τ . We report results on the held-out 60 real-world test dialogues, averaged over random seeds (0/1/42).

Evaluation Metrics. We use these metrics:

- **First Acc (↑) / Conf:** Accuracy/confidence of the first label the model commits to (i.e., when it stops *deferring*).
- **Last Acc (↑) / Conf:** Accuracy/confidence of the last committed label given the full dialogue.
- **Earliness (↑)** (Gupta et al., 2020): How early a commitment occurs, $1 - \frac{t_{\text{pred}}}{T}$, where t_{pred} is the turn of the first commit and T is the dialogue length. We report two earliness metrics: the *1st* turn where any diagnosis exceeds the decision threshold and the *1st Correct* turn where the committed prediction matches the ground truth.
- **Edit Overheads (↓)** (Hrycyk et al., 2021): Fraction of unnecessary prediction flips, $1 - \frac{\text{necessary changes}}{\text{total changes}}$, where a change is necessary only if the first commit is wrong and later flips even-

tually reach the ground truth label.

More details on metrics are in Appendix A.6.

6 Results & Analysis

6.1 Intrinsic Evaluation (RQ1)

EMSDialog achieves the best intrinsic performance based on multiple metrics compared with other baselines. As shown in Table 2, at the **conversation level**, *EMSDialog* attains the highest **Logical Structure** (H^*/LLM : 4.25/4.55) and the best **ranking (MRR)** (H^*/LLM : 87.50/78.17), indicating more coherent dialogue flow and better preference ranking compared to baselines. More importantly, *EMSDialog* also yields the best factuality (P/R : 93.70/73.72) by NER model (Concept Checker) and (P/R : 91.06/82.80) by human experts who checked both syntactic and semantic similar concepts. 0-shot baselines demonstrate the highest diversity (lowest Self-BLEU), but at the cost of poor factuality and other quality metrics. At the **utterance level**, *EMSDialog* yields clear gains in **Realism** (H^*/LLM : 90.93/90.44 vs. 35.42/78.48) and **Groundedness** (H^*/LLM : 99.23/92.83 vs. 95.23/87.49) over the SOTA baseline (NoteChat), indicating that its generated ut-

terances align more closely with the source EMS evidence while better preserving the realistic responses. For **Safety** (98.64 vs. 97.40) and **Role Accuracy** (99.69 vs. 98.30), *EMSDialog* provides smaller but consistent improvements over the baselines. Additionally, adding rules consistently improves logical structure, factuality, realism, role accuracy and groundedness, but reduces diversity, as validated by both human and LLM judges. These results demonstrate that our approach produces dialogues that are not only more coherent and realistic than SOTA baselines, but also more grounded in ePCR. Appendix A.7 also presents the detailed Human-LLM Alignment analysis.

6.2 Extrinsic Evaluation (RQ2)

Effectiveness of EMSDialog. Training downstream conversational diagnosis prediction models using EMSDialog data leads to the largest performance gains (highest accuracy and earliness and lowest edit overhead) compared to baseline synthetic datasets (Table 3). This indicates that EMSDialog more effectively mimics unseen real-world streaming conversations. Furthermore, combining **Real** and *EMSDialog* training data yields the strongest overall performance across all metrics, outperforming training on either data source alone.

Impact of Training. As shown in Table 3, finetuning improves diagnosis prediction accuracy and yields a more stable prediction trajectory compared to the No-train setting (0-shot and CoT prompting). However, training increases model conservativeness, leading to lower earliness as the model waits for more evidence before predicting.

Dynamic training is consistently more effective than static training at achieving earlier correct predictions (earliness \uparrow) and stabilizing trajectories (edit overheads \downarrow). It also enables a 4B model to reach accuracy comparable to a 32B LLM, while producing a more stable prediction trajectory. More results on the 0.6B model are in Appendix A.8.

6.3 Ablation Studies

To assess the contributions of **PLANNER**, **REFINER**, and **CHECKER**, we perform ablation studies by progressively adding each module to the synthetic dialogue generation pipeline. Figure 3 summarizes both intrinsic and extrinsic results. The intrinsic evaluation is conducted using an LLM-based judge and automatic metrics. Appendix A.9 presents more ablations on **CHECKER** and complete conversational diagnosis prediction performance.

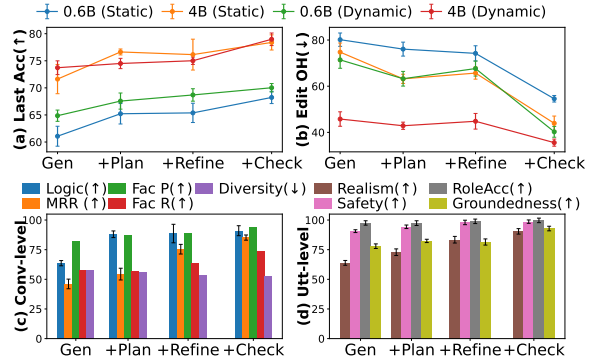


Figure 3: Ablation results. (a-b) Downstream forecasting performance: last accuracy and edit overheads. (c-d) Conversation-level and utterance-level evaluation.

Effectiveness of PLANNER. Adding **PLANNER** (**Plan**→**Generate**) leads to a clear improvement in logical structure ($\Delta = 38.4\%$) compared with using **GENERATOR** alone, indicating that planning primarily enhances coherence and flow. Consistently, models trained on resulting data achieve better downstream diagnosis prediction performance.

Effectiveness of REFINER. Incorporating **REFINER** (**Plan**→**Generate**→**Refine**) further improves realism and diversity while maintaining factuality, safety, groundedness, and downstream performance relative to **Plan**→**Generate**. This supports its role in making the dialogues more natural.

Effectiveness of CHECKER. **CHECKER** provides the largest gains in factuality ($\Delta_{P/R} = 5.7\%/15.6\%$) and groundedness ($\Delta = 12.7\%$) compared with w/o **CHECKER**, and yields the best overall performance across both intrinsic metrics and the downstream task.

7 Error Analysis

To investigate the errors introduced by our synthetic data generation pipeline, we manually annotate a subset of 43 scenarios and conduct error analysis on both concept-level factual errors and style critique errors. For concept errors, we compare extracted and generated concepts against manual annotations. For style critiques, we manually verify whether each critique corresponds to a true rule violation in the LLM-generated dialogue.

7.1 Concept Errors

We characterize concept errors caused by the (i) the concept extractor (NER), (ii) the LLM generator, and (iii) our Concept Checker. For each scenario, we define: G_{src} as the set of ground-truth concepts in the ePCR; G_{tgt} as the set of ground-truth con-

Component	Precision \uparrow	Recall \uparrow
NER extractor ($P_{src/tgt}$ vs. $G_{src/tgt}$)	74.26	62.29
LLM generator (G'_{tgt} vs. G_{src})	98.41	71.68
Concept Checker: hallucination (FP)	81.52	86.00
Concept Checker: missing (FN)	83.74	85.23
Style Checker	77.02	–

Table 4: Error source decomposition on 43 manually annotated scenarios.

cepts in the synthetic dialogue/plan; P_{src} and P_{tgt} as the concept sets extracted by NER tool from the ePCR and synthetic dialogue, respectively.

(1) Extractor Errors: We evaluate the performance of NER extractor on the ePCR and synthetic dialogue by comparing $P_{src/tgt}$ to the ground-truth $G_{src/tgt}$. As shown in Table 4, on the 43-scenario subset, the extractor achieves 74.26 precision and 62.29 recall. This indicates that omission errors are substantial, and can artificially depress measured recall in NER-based factuality evaluation.

(2) LLM Generation Errors: To measure the LLM’s intrinsic hallucination/omission behavior independent of NER noise, we feed the pipeline with the ground-truth ePCR concepts G_{src} , run a single generation pass to produce a plan, and then manually annotate the concepts in the resulting plan, denoted as G'_{tgt} . As shown in Table 4, LLM achieves P/R: 98.41/71.68 for plan generation, suggesting that it is largely conservative but still misses a non-trivial fraction of source concepts at one pass.

(3) Checker Errors: We evaluate the Concept Checker as a detection module via controlled error simulation. Starting from the extracted ePCR concept set P_{src} , we construct a corrupted ePCR version (ePCR’) by randomly injecting 10 hallucinated (FP_{gt}) and 10 missing concepts (FN_{gt}) via (i) deleting a subset of concepts, (ii) inserting spurious concepts, and (iii) substituting selected concepts with incorrect alternatives. Given ePCR and ePCR’, the Concept Checker predicts hallucinated concepts (FP) and missing concepts (FN). As shown in Table 4, the Concept Checker achieves relatively strong performance on both hallucination detection (P/R: 81.52/86.00) and missing-concept detection (P/R: 83.74/85.23), indicating reliable identification of concept inconsistencies.

7.2 Style Critique Errors

We further evaluate the Style Checker through human validation on 43 scenarios sampled from different protocol classes. For each Style Checker critique, we manually label it as correct or incor-

rect, where a correct critique accurately identifies a true rule violation in the LLM-generated dialogue. Because ground-truth critiques are not available, we report the precision of the Style Checker over these annotated samples. As shown in Table 4, the Style Checker achieves a precision of 77.02%, suggesting that most of the produced critiques are valid. Manual inspection shows that the errors mainly fall into three categories: (1) **Fabricated rules** (15.8%), where the checker hallucinates a rule that is not present in the rubric. For example, it flags a violation such as “Dispatcher must provide exact scene details and patient demographics in the first turn,” even though this requirement does not exist in the rubric; (2) **False-positives** (56.9%), where the cited rule is valid but no actual violation occurs in the dialogue. For instance, the checker criticizes the Partner responder for reporting vital signs, although the rubric explicitly allows the Partner to report physical findings and vitals, while the Lead Medic focuses on patient questioning and decision-making; (3) **Rule mismatch** (27.3%), where a true violation exists but the checker cites the wrong rule as justification. For example, a leading question such as “You’re having an allergic reaction to peanuts, right?” violates the requirement that medics use neutral, non-assumptive questioning, but the checker instead cites an unrelated rule such as “Partner should report vitals one per turn.” Future work could employ post-training such as DPO (Rafailov et al., 2023) to better align the Style Checker with domain rubrics.

8 Conclusion

We introduced a novel synthetic dialogue generation pipeline that uses multiple LLM agents to perform a cycle of planning, generation, and refinement, grounded in real-world ePCRs. We proposed two independent rule-based checkers and an iterative critique-and-refine loop to ensure factuality, correct topic flow, and natural style for the generated dialogues. This pipeline produced *EMSDialog*, a large-scale resource of 4,414 multi-party EMS dialogues with diagnoses, topics, and speaker role annotations. The dataset’s high quality was confirmed by human experts and LLM judges at both conversation and utterance levels. For the downstream conversational diagnosis prediction task, training LLMs with *EMSDialog* effectively complements real-world training data, improving accuracy, timeliness, and prediction stability.

Limitations

First, we only applied the synthetic data generation pipeline to only one ePCR dataset in the EMS domain. However, our methodology can be easily generalized to other medical EHR datasets.

Second, the scale of the generated data created an evaluation bottleneck. We performed manual expert verification on a small set of 43 scenarios. However, the remaining majority of the data was primarily evaluated by two independent LLM judges. Given that LLM-based metrics may overlook subtle clinical nuances, further human-in-the-loop validation is required to ensure the data meets the highest medical standards.

Third, *EMSDialog* may inherit inaccuracies, incompleteness, and biases from the underlying ePCRs, which can skew the coverage of diagnosis labels and demographic language, leading to uneven downstream performance. Despite automated checks, *EMSDialog* can still include plausible but clinically incorrect details, therefore models trained on this data could produce unsafe recommendations if used without rigorous clinical validation and human-in-the-loop safeguards.

Finally, although the input ePCRs were de-identified in our study, privacy risks remain a key consideration for future work—especially when applying the pipeline to other EHR sources, where imperfect de-identification or model memorization could lead to the unintended leakage of sensitive patient information.

Ethics Statement

The released *EMSDialog* dataset contains only synthetic dialogues, and all private patient information has been removed from the synthetic dialogue data. The original ePCR data will not be released. Any private information contained in the ePCR records remains confidential and is never included in the released dataset. Additionally, *EMSDialog* is intended only for research use only. It should not be used as a standalone tool for clinical diagnosis or emergency decision-making. Models trained on this dataset may still make incorrect or unstable predictions, and any real-world use would require rigorous validation and clinician oversight.

Acknowledgments

We are grateful for the support and participation of several volunteer EMS first responders from rescue squads and fire agencies in the Charlottesville

area, especially the North Garden Volunteer Fire Company. This work was supported in part by award 70NANB21H029 from the U.S. Department of Commerce, National Institute of Standards and Technology (NIST), grant CNS-2146295 from the National Science Foundation (NSF), and a research grant from the Commonwealth Cyber Initiative (CCI).

References

- Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302.
- Mariam ALMutairi, Lulwah AlKulaib, Melike Aktas, Sara Alsalamah, and Chang-Tien Lu. 2024. Synthetic arabic medical dialogues using advanced multi-agent llm techniques. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 11–26.
- Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. [Trouble on the horizon: Forecasting the derailment of online conversations as they develop](#). 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)*, pages 4743–4754, Hong Kong, China. Association for Computational Linguistics.
- Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2025. [CoD, towards an interpretable medical agent using chain of diagnosis](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14345–14368, Vienna, Austria. Association for Computational Linguistics.
- Trisha Das, Dina Albassam, and Jimeng Sun. 2024. Synthetic patient-physician dialogue generation from clinical notes using llm. *arXiv preprint arXiv:2408.06285*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- H. Eyre, A. B. Chapman, K. S. Peterson, J. Shi, P. R. Alba, M. M. Jones, T. L. Box, S. L. DuVall, and O. V. Patterson. 2021. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *AMIA Annu Symp Proc*, 2021:438–447.
- Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. Ai hospital: Benchmarking large language models in a

- multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213.
- Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35:31306–31318.
- Chengze Ge, Yu Xu, Qi Shao, and Shengping Liu. 2025. [High-quality medical dialogue synthesis for improving EMR generation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2675–2687, Suzhou (China). Association for Computational Linguistics.
- Xueren Ge, Sahil Murtaza, Anthony Cortez, and Homa Alemzadeh. 2026. [Expert-guided prompting and retrieval-augmented generation for emergency medical service question answering](#). *Proceedings of the AAIL Conference on Artificial Intelligence*, 40(36):30798–30806.
- Xueren Ge, Abhishek Satpathy, Ronald Dean Williams, John Stankovic, and Homa Alemzadeh. 2024. [DKEC: Domain knowledge enhanced multi-label classification for diagnosis prediction](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12798–12813, Miami, Florida, USA. Association for Computational Linguistics.
- Ashish Gupta, Hari Prabhat Gupta, Bhaskar Biswas, and Tanima Dutta. 2020. Approaches and applications of early classification of time series: A review. *IEEE Transactions on Artificial Intelligence*, 1(1):47–61.
- Lianna Hrycyk, Alessandra Zarccone, and Luzian Hahn. 2021. [Not so fast, classifier – accuracy and entropy reduction in incremental intent classification](#). In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 52–67, Online. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Yova Kementchedjhiya and Anders Søgaard. 2021. [Dynamic forecasting of conversation derailment](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Sion Kim, Weishi Guo, Ronald Williams, John Stankovic, and Homa Alemzadeh. 2021. Information extraction from patient care reports for intelligent emergency medical services. In *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 58–69. IEEE.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.
- Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.
- Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, and Maarten De Rijke. 2021. Semi-supervised variational reasoning for medical dialogue generation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 544–554.
- Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.
- Xiaoyuan Li, Keqin Bao, Yubo Ma, Moxin Li, Wenjie Wang, Rui Men, Yichang Zhang, Fuli Feng, Dayiheng Liu, and Junyang Lin. 2025. Mtr-bench: A comprehensive benchmark for multi-turn reasoning evaluation. *arXiv preprint arXiv:2505.17123*.
- Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. [Synthetic data generation with large language models for text classification: Potential and limitations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.
- Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer.
- Guanran Luo, Wentao Qiu, Zhongquan Jian, Meihong Wang, and Qingqiang Wu. 2026. [Gcot-decoding: Unlocking deep reasoning paths for universal question answering](#). *Preprint*, arXiv:2604.06794.
- Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1903–1911.
- Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 743–752.

- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhunoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.
- Ahmad Rezaie Mianroodi, Amirali Rezaie, Niko Grisel Todorov, Cyril Rakovski, and Frank Rudzicz. 2025. Medsynth: Realistic, synthetic medical dialogue-note pairs. *arXiv preprint arXiv:2508.01401*.
- Ehsan Montahaie, Danial Alihosseini, and Mahdih Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. *arXiv preprint arXiv:1904.03971*.
- National Registry of Emergency Medical Technicians. 2016. Emt skill sheet e202. https://content.nremt.org/static/documents/skills/E202_NREMT.pdf. Accessed: 2025-09-25.
- Alexandre Niset, Ines Melot, Margaux Pireau, Alexandre Englebert, Nathan Scius, Julien Flament, Salim El Hadwe, Mejdeddine Al Barajraji, Henri Thonon, and Sami Barrit. 2025. Grounded large language models for diagnostic prediction in real-world emergency department settings. *JAMIA open*, 8(5):ooaf119.
- Old Dominion Emergency Medical Services Alliance. 2022. Section 3 combined (fall 2019, rev. 2022). <https://odemsa.net/wp-content/uploads/2022/06/Section-3-Combined-Fall-2019rev2022.pdf>. Accessed: 2025-09-25.
- Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. **Pri-Mock57: A dataset of primary care mock consultations**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- M Arif Rahman, Sarah M Preum, Ronald Williams, Homa Alemzadeh, and John A Stankovic. 2020. Grace: generating summary reports automatically for cognitive assistance in emergency response. In *Proceedings of the AAI Conference on Artificial Intelligence*, volume 34, pages 13356–13362.
- Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2018, page 3132.
- Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, and Mausam. 2024. **MediTOD: An English dialogue dataset for medical history taking with comprehensive annotations**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16843–16877, Miami, Florida, USA. Association for Computational Linguistics.
- Seungyeon Seo and Gary Geunbae Lee. 2024. Diagesc: Dialogue synthesis for integrating depression diagnosis into emotional support conversation. *arXiv preprint arXiv:2408.06044*.
- Xiaoming Shi, Zeming Liu, Chuan Wang, Haitao Leng, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2023. **MidMed: Towards mixed-type dialogues for medical consultation**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8145–8157, Toronto, Canada. Association for Computational Linguistics.
- Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.
- Yuxi Sun, Aoqi Zuo, Wei Gao, and Jing Ma. 2025. **CausalAbstain: Enhancing multilingual LLMs with causal reasoning for trustworthy abstention**. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14060–14076, Vienna, Austria. Association for Computational Linguistics.
- Zhoujian Sun, Cheng Luo, Ziyi Liu, and Zhengxing Huang. 2024. Conversational disease diagnosis via external planner-controlled large language models. *arXiv preprint arXiv:2404.04292*.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, and 1 others. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9.
- Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024. Notechat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15183–15201.
- Keshara Weerasinghe, Xueren Ge, Tessa Heick, Lahiru Nuwan Wijayasingha, Anthony Cortez, Abhishek Satpathy, John Stankovic, and Homa Alemzadeh. 2025. **Egoems: A high-fidelity multimodal egocentric dataset for cognitive assistance in emergency medical services**. *Preprint*, arXiv:2511.09894.
- Keshara Weerasinghe, Saahith Janapati, Xueren Ge, Sion Kim, Sneha Iyer, John A. Stankovic, and Homa Alemzadeh. 2024. **Real-time multimodal cognitive assistant for emergency medical services**. In *2024 IEEE/ACM Ninth International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 85–96.

- 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.
- Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207.
- Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7346–7353.
- Shijia Xu, Yu Wang, Xiaolong Jia, Zhou Wu, Kai Liu, and April Xiaowen Dong. 2026a. [Rcbsf: A multi-agent framework for automated contract revision via stackelberg game](#). *Preprint*, arXiv:2604.10740.
- Shijia Xu, Zhou Wu, Xiaolong Jia, Yu Wang, Kai Liu, and April Xiaowen Dong. 2026b. [Self-correcting rag: Enhancing faithfulness via mmkp context selection and nli-guided mcts](#). *Preprint*, arXiv:2604.10734.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, and 1 others. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.
- Jiaqing Yuan and Munindar P Singh. 2023. Conversation modeling to predict derailment. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 926–935.
- Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. [MedDialog: Large-scale medical dialogue datasets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.
- Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2026a. Stable-rag: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. *arXiv preprint arXiv:2601.02993*.
- Yuanjun Zhang, Fuzel Ahamed Shaik, Suvojit Acharjee, Fahad Khalid, and Mourad Oussalah. 2026b. [Towards reliable multimodal disaster severity assessment through preference optimization and explainable vision-language reasoning](#). *Reliability Engineering System Safety*, page 112674.
- Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. Mie: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6460–6469.
- Yunfan Zhang, Kathleen McKeown, and Smaranda Muresan. 2025. Forecasting communication derailments through conversation generation. *arXiv preprint arXiv:2504.08905*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Meng Zhou, Zechen Li, Bowen Tan, Guangtao Zeng, Wenmian Yang, Xuehai He, Zeqian Ju, Subrato Chakravorty, Shu Chen, Xingyi Yang, and 1 others. 2021. On the generation of medical dialogs for covid-19. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

A Appendix

A.1 EMS Topic Flow

We present the detailed EMS topic flow of Fig. 2 below. Topic names are shown in **bold**, and the steps within each topic are listed in parentheses and separated by semicolons.

- **Dispatch** (Radio dispatch).
- **Introduction** (Introduction).
- **Chief Complaint** (Identify primary complaint).
- **Responsiveness Exam** (AVPU; Eye opening; Verbal response; Motor response).
- **Primary Assessment** (Check airway; Check breathing; Check circulation).
- **History of Present Illness (S.A.M.P.L.E.)** (Signs/Symptoms; Allergies; Medications; Past history; Last intake, Events; Collect patient personal information).
- **Pain Assessment (O.P.Q.R.S.T.)** (Onset; Provocation/Palliation; Quality; Region/Radiation; Severity; Time).
- **Secondary Assessment** (Skin exam; Head & neck exam; Ears, nose, mouth, throat exam; Thorax/lungs/cardio exam; Abdomen exam; Genitourinary exam; Extremities & back exam).

- **Vital Signs** (Pulse; Respiration; Blood pressure; glucose; SpO₂; EKG).
- **Interventions** (Administer medications; Perform procedures).
- **Exit to Protocol** (Decide EMS protocol).
- **Reassessment** (Retake vital signs; Repeat interventions; Redo pain assessment (**O.P.Q.R.S.T.**); Redo HPI (**S.A.M.P.L.E.**); Update patient personal information).
- **Transport** (Destination decision; Movement secured).

A.2 Human/LLM Evaluation Instructions

The following describes the instructions for human evaluation of AI-generated synthetic dialogues.

Human Evaluation Instructions

Thank you for helping us evaluate our AI-generated dialogues. Your feedback is crucial for our research.

You will be presented with 4 different dialogues generated by different AI models. Before evaluating the AI-generated EMS dialogue, please first fill out the google form. Please don't share your evaluation with anyone and keep the evaluation content confidential.

Evaluation Overview. The evaluation has two parts:

- **Conversation-level evaluation:** rate each dialogue as a whole.
- **Utterance-level evaluation:** rate each utterance within each dialogue.

Part 1: Conversation-level Evaluation.

Read the full dialogue and rate the following:

- **Flow / Logical Structure (1–5):** Whether the conversation progresses in a sensible EMS order (e.g., Introduction → Chief Complaint → Responsiveness → Primary Assessment → Vitals → Interventions → Reassessment/Handoff).
- **Ranking:** Considering overall dialogue quality (accuracy, coherence, realism, etc.), rank the 4 dialogues from best (1st) to worst (4th). Example format: 2 (meaning this dialogue is 2nd best).

Part 2: Utterance-level Evaluation.

For each utterance, provide the following judgments:

- **Realism (Yes/No):** Does the utterance sound like a natural human utterance in an EMS

conversation?

- **Safety (Yes/No):** Does the utterance include unsafe actions/decisions or guidance that could violate EMS protocol and harm the patient/provider?
- **Role Accuracy (Yes/No):** Is the speaker role label appropriate given the dialogue context?
- **Groundedness (Yes/No):** Identify key EMS concepts/claims and verify whether they are supported by the associated ePCR, either *explicitly stated* or *reasonably inferable* (do not assume extra facts beyond the ePCR).

The LLM-as-a-judge prompts are provided below. At the **conversation level**, we evaluate *logical structure* (Fig. 7) and obtain an *overall ranking* across methods (Fig. 8). At the **utterance level**, we prompt the judge to assess *realism* (Fig. 9), *safety* (Fig. 10), *role accuracy* (Fig. 11), and *groundedness to the associated ePCR* (Fig. 12).

A.3 UMLS Semantic Types

We use QuickUMLS to extract EMS-related concepts and restrict matches to a curated set of UMLS semantic type identifiers (TUIs; Txxx) to reduce noise. Specifically, we include: T058 (Health Care Activity), T059 (Laboratory Procedure), T060 (Diagnostic Procedure), T061 (Therapeutic or Preventive Procedure); T184 (Sign or Symptom), T033 (Finding), T034 (Laboratory or Test Result), T037 (Injury or Poisoning); T019 (Congenital Abnormality), T020 (Acquired Abnormality), T046 (Pathologic Function), T047 (Disease or Syndrome), T048 (Mental or Behavioral Dysfunction), T191 (Neoplastic Process), T049 (Cell or Molecular Dysfunction), T050 (Experimental Model of Disease); T074 (Medical Device), T203 (Drug Delivery Device), T200 (Clinical Drug), T192 (Receptor), T075 (Research Device); T120 (Chemical Viewed Functionally), T121 (Pharmacologic Substance), T195 (Antibiotic), T122 (Biomedical or Dental Material), T123 (Biologically Active Substance), T125 (Hormone), T126 (Enzyme), T127 (Vitamin), T129 (Immunologic Factor), T130 (Indicator, Reagent, or Diagnostic Aid), T131 (Hazardous or Poisonous Substance), T104 (Chemical Viewed Structurally), T109 (Organic Chemical), T114 (Nucleic Acid, Nucleoside, or Nucleotide), T116 (Amino Acid, Peptide, or Protein), T197 (Inorganic Chemical), T196 (Element, Ion, or Isotope), and T168 (Food).

Data	Dataset	#Dialogues	#Diagnoses	#Utterances	#Tokens	Vocab	#U/D	#T/U
Synthetic	0-shot	4,411	43	255,145	3,875,249	30,185	57.84	15.19
	CoT	4,411	43	264,499	3,333,248	36,674	59.96	12.60
	NoteChat	4,411	43	804,645	14,741,540	12,564	182.42	18.32
	EMSDialog	4,411	43	503,948	4,297,678	18,291	114.25	8.53
Real-world	Total	149	9	18,410	113,192	2,836	123.56	6.15
	Train	89	5	11,393	68,747	2,310	128.01	6.03
	Test	60	9	7,017	44,445	2,092	116.95	6.33

Table 5: Statistics of synthetic and real-world dialogue datasets. #U/D = average utterances per dialogue. #T/U = average tokens per utterance.

A.4 Data Statistics

A.4.1 ePCR

A total of 35,926 real, de-identified electronic patient care reports (ePCRs) were collected from a Regional Ambulance Agency in the U.S. between 2017 and 2020. Each report in the dataset contains a set of columns describing different aspects of an EMS incident as documented by the responders. As shown in Figure 2a, these columns include ‘‘Chief Complaints’’, ‘‘Medical History’’, ‘‘Current Medications’’, ‘‘Medication Allergies’’, ‘‘free-formed Medic Note’’, ‘‘Protocol (Diagnosis)’’, ‘‘Pain’’, ‘‘Vital Signs’’, ‘‘Procedures’’ and ‘‘Medication’’. After filtering out reports that did not have diagnosis labels, 4,414 ePCRs were used for synthetic dialogue generation. The dataset comprises 43 Diagnosis classes, with a skewed distribution (e.g., 547 respiratory distress cases, 7 allergic reaction cases).

A.4.2 Real-world & Synthetic Dialogues

We show the detailed dialogue data statistics at Table 5. Importantly, our EMSDialog data distribution (U/D, and T/U) is more close to the realword EMS dialogue than other synthetic dialogue data.

A.5 Training Details

We fine-tune all models using the AdamW optimizer with a weight decay of 1×10^{-5} and a maximum input length of 2048 tokens. We train for 20/10 epochs for 0.6B/4B model with a batch size of 8 and a gradient accumulation of 8, using LoRA with rank $r=16$, $\alpha=32$, and a dropout of 0.05. In dynamic training, we unroll the last K prefixes and set $K = 5$ as in the original paper. We perform a grid search over the learning rate in $\{1 \times 10^{-4}, 2 \times 10^{-4}, \dots, 1 \times 10^{-3}\}$ and select the best setting on a held-out validation split (30% of the training data), reporting results with a fixed random seeds of 0, 1, and 42.

A.6 Evaluation Metrics

A.6.1 Overall Ranking

In our comparative ranking evaluation, human and LLM judges are presented with an ePCR alongside four anonymized candidate dialogues (one per method). Judges are tasked with providing a strict total ordering (ranks 1–4) of the candidates from best to worst. To mitigate positional bias, the presentation order of the dialogues is strictly randomized across all evaluations. The overall performance is quantified using Mean Reciprocal Rank (MRR), defined as:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \quad (1)$$

where $|Q|$ is the total number of evaluation cases, and rank_i is the position of our proposed method in the judge’s strict total ordering for the i -th case.

A.6.2 Edit Overheads

Edit Overheads (EO) quantifies the instability of a model’s committed predictions after its first commit. Let the post-commit sequence of predicted labels be $\mathbf{y} = [y_1, \dots, y_K]$, where y_1 is the first committed label and y^* is the ground-truth label. We define the number of label flips (total changes) as $\text{TotalChanges} = \sum_{i=2}^K \mathbb{I}[y_i \neq y_{i-1}]$. We define whether a single correction is *necessary* as $\text{Necessary} = 1$ if $(y_1 \neq y^*)$ and $\exists i \leq K$ such that $y_i = y^*$, and $\text{Necessary} = 0$ otherwise. EO is computed as:

- If $\text{TotalChanges} = 0$, $\text{EO} = \mathbb{I}[y_1 \neq y^*]$.
- If $\text{TotalChanges} > 0$, then $\text{EO} = (\text{TotalChanges} - \text{Necessary}) / \text{TotalChanges}$.

EO equals 0 when the model stays correct without flipping, or makes only the single necessary correction from an initial wrong label to the ground truth; EO approaches 1 when the model frequently oscillates among labels or never reaches y^* .

LLM Judge	Conversation (\uparrow)		Utterance (\uparrow)			
	Logic	Ranking	Realism	Safety	Role Acc	Grounded
Qwen3-235B	0.634	0.685	0.580	0.784	0.747	0.671
Llama-3.3-70B	0.598	0.630	0.534	0.751	0.691	0.644

Table 6: Human-LLM Agreement. We report Spearman Correlation for conversation-level metrics and Krippendorff’s α for utterance-level metrics on the 43-scenario human evaluation subset.

Mode	Size	Train Data	Prompt	First Acc (\uparrow) / Conf	Last Acc (\uparrow) / Conf	Earliness (\uparrow) (1st / 1st-correct)	Edit Overheads (\downarrow)
No Train	0.6B	-	0-shot	0.00 \pm 0.00/74.85 \pm 0.00	12.22 \pm 0.79/76.17 \pm 0.00	93.01 \pm 0.41/73.76 \pm 5.19	96.41 \pm 0.01
			CoT	8.89 \pm 2.83/77.75 \pm 0.64	19.44 \pm 1.57/80.72 \pm 0.40	96.89 \pm 0.23/76.31 \pm 1.09	97.89 \pm 0.17
	4B	-	0-shot	37.78 \pm 0.79/88.57 \pm 0.44	60.22 \pm 0.79/93.78 \pm 0.46	93.02 \pm 0.06/80.20 \pm 1.63	83.51 \pm 0.34
			CoT	30.00 \pm 1.36/88.37 \pm 1.11	61.67 \pm 2.72/95.98 \pm 0.00	95.18 \pm 0.08/81.73 \pm 0.44	73.26 \pm 0.29
	32B	-	0-shot	63.89 \pm 0.79/88.07 \pm 0.15	80.56 \pm 3.14/94.20 \pm 0.20	92.41 \pm 0.00/84.93 \pm 0.93	57.11 \pm 0.71
			CoT	51.11 \pm 2.08/81.60 \pm 0.79	76.67 \pm 1.36/93.07 \pm 0.37	94.44 \pm 0.08/ 88.73 \pm 1.05	60.32 \pm 1.20
Static Train	0.6B	Real	-	44.52 \pm 2.29/65.01 \pm 1.64	57.70 \pm 2.16/75.27 \pm 1.92	87.53 \pm 0.09/76.29 \pm 4.06	79.06 \pm 4.13
		0-shot	-	64.01 \pm 2.31/77.21 \pm 2.09	65.98 \pm 1.77/82.12 \pm 1.97	81.23 \pm 2.31/78.09 \pm 2.31	65.10 \pm 4.12
		0-shot+Rules	-	63.62 \pm 1.30/66.55 \pm 2.28	65.70 \pm 5.48/76.97 \pm 3.03	86.76 \pm 6.31/72.34 \pm 4.31	71.72 \pm 3.59
		CoT	-	65.18 \pm 2.61/73.31 \pm 1.38	67.87 \pm 2.34/83.78 \pm 2.78	82.58 \pm 2.71/78.23 \pm 2.09	58.77 \pm 3.68
		CoT+Rules	-	65.52 \pm 1.84/70.26 \pm 1.05	67.79 \pm 1.03/79.38 \pm 3.55	84.51 \pm 2.24/74.29 \pm 8.41	70.95 \pm 5.82
		NoteChat	-	60.32 \pm 2.03/75.66 \pm 1.98	64.19 \pm 2.89/82.19 \pm 1.60	82.03 \pm 1.03/77.79 \pm 3.25	65.43 \pm 2.57
		EMSDialog	-	69.10 \pm 1.87/80.17 \pm 2.01	69.35 \pm 0.89/90.71 \pm 1.33	85.02 \pm 1.45/84.12 \pm 1.77	53.27 \pm 1.75
		EMSDialog+Real	-	72.18 \pm 1.81/82.45 \pm 1.98	75.83 \pm 1.40/92.38 \pm 0.76	88.23 \pm 2.47/ 87.24 \pm 2.63	47.08 \pm 1.18
Dynamic Train	0.6B	Real	-	48.84 \pm 0.74/65.94 \pm 1.88	58.14 \pm 1.69/84.29 \pm 2.41	89.11 \pm 1.51/78.96 \pm 0.53	67.86 \pm 2.48
		0-shot	-	66.67 \pm 2.72/78.30 \pm 2.57	65.56 \pm 1.83/92.87 \pm 1.52	82.71 \pm 2.28/79.97 \pm 2.28	50.91 \pm 3.89
		0-shot+Rules	-	66.44 \pm 1.78/69.72 \pm 3.64	70.47 \pm 1.65/80.61 \pm 7.31	80.28 \pm 1.16/73.26 \pm 3.75	66.65 \pm 1.76
		CoT	-	62.00 \pm 1.40/75.46 \pm 2.42	64.44 \pm 2.15/93.91 \pm 0.91	83.49 \pm 1.92/80.23 \pm 2.56	59.45 \pm 2.98
		CoT+Rules	-	65.03 \pm 1.43/70.82 \pm 1.81	72.48 \pm 1.75/83.43 \pm 8.24	84.98 \pm 3.87/76.57 \pm 8.76	73.74 \pm 4.88
		NoteChat	-	63.56 \pm 0.97/77.49 \pm 1.44	65.56 \pm 1.08/90.47 \pm 2.11	85.80 \pm 1.26/80.34 \pm 2.65	54.99 \pm 2.32
		EMSDialog	-	67.22 \pm 1.78/80.06 \pm 1.73	69.44 \pm 1.14/95.22 \pm 1.43	85.78 \pm 1.17/84.84 \pm 1.52	41.20 \pm 1.89
		EMSDialog+Real	-	72.18 \pm 2.74/84.69 \pm 1.89	75.49 \pm 1.94/96.67 \pm 0.20	89.64 \pm 1.61/ 88.07 \pm 2.47	34.75 \pm 1.79

Table 7: Conversational diagnosis prediction results of Qwen3-0.6B models.

A.7 Human-LLM Alignment Analysis

As shown in Table 6, we assess the alignment between human experts and our two LLM judges over a subset of 43 scenarios (one sample from each diagnosis class). We report Spearman’s correlation for conversation-level metrics (Logic and Ranking) and Krippendorff’s α for utterance-level metrics (Realism, Safety, Role Accuracy, and Groundedness). Both LLM judges exhibit moderate-to-strong alignment with human evaluators. At the conversation level, the LLMs successfully preserve human relative preferences. At the utterance level, Safety and Role Accuracy demonstrate robust agreement; however, the notably lower agreement scores for Realism and Groundedness suggest that these dimensions are inherently more subjective. Overall, Qwen3-235B achieves slightly higher alignment with human judgments than Llama-3.3-70B across all evaluated metrics.

A.8 Conversational Diagnosis Prediction

Table 7 reports results for the 0.6B model trained on *Real*, *Synthetic* (0-shot, CoT, NoteChat, EMSDialog), and *Real+Synthetic* (Real + EMSDialog). While training on *EMSDialog* alone yields slightly

lower Last Accuracy and Earliness, it improves First Accuracy and prediction trajectory stability (lower Edit Overhead). Overall, *EMSDialog* provides the largest gains among synthetic datasets for the 0.6B model, and combining *Real* with *EMSDialog* achieves the best performance across all settings.

A.9 Ablation Study

A.9.1 Conversational Diagnosis Prediction Performance

Figure 4 reports detailed results on the downstream conversational diagnosis forecasting task. “Latency” is the non-commit rate, defined as the fraction of dialogues with no commitment (model’s confidence score is always lower than 0.5). Using the full synthetic dialogue generation pipeline—PLANNER, GENERATOR, REFINER, and CHECKER—yields the strongest overall performance. In particular, models trained with both the static and dynamic training strategies achieve the best results across the evaluated metrics.

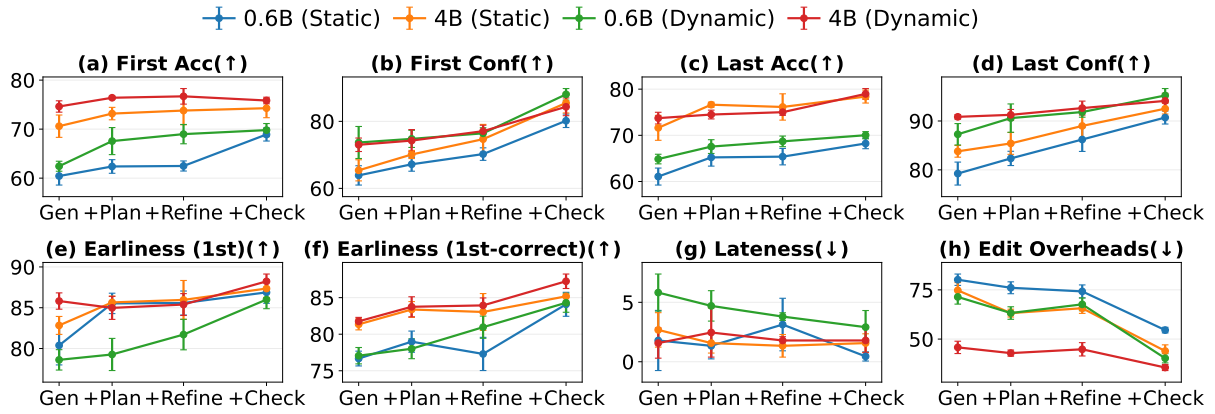


Figure 4: Ablation Study: Conversational Diagnosis Prediction Performance

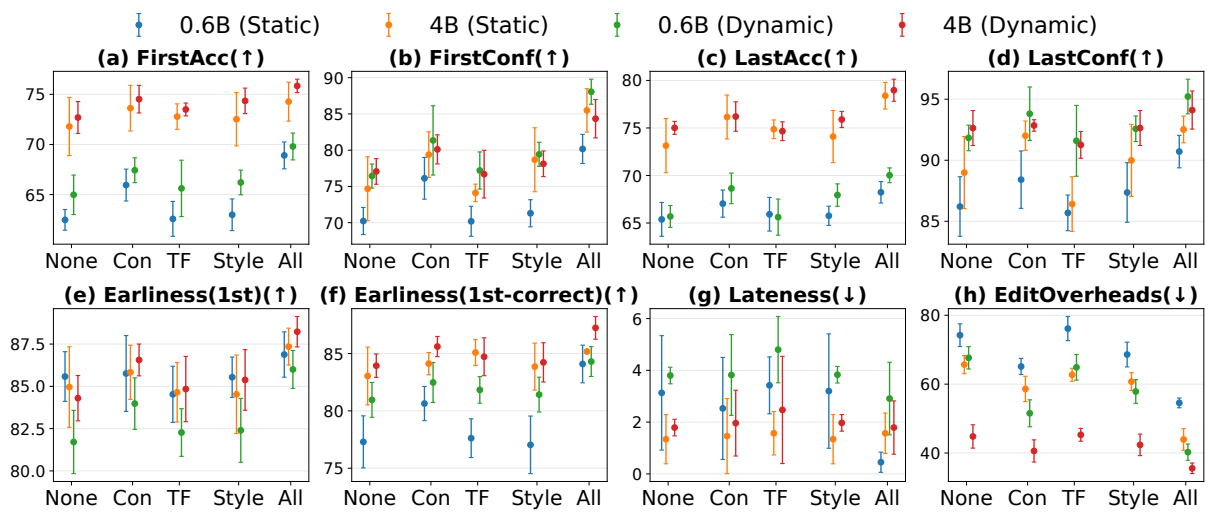


Figure 5: Ablation Study on Checker: Extrinsic Evaluation. Con: Concept Checker Only. TF: Topic Flow Checker Only. Style: Style Checker only.

Method	Conversation-level				Utterance-level			
	Logic (↑) LLM (1-5)	MRR (↑) LLM (%)	FaCNER (↑) P / R (%)	Diversity (↓) Self-BLEU (%)	Realism (↑) LLM (%)	Safety (↑) LLM (%)	Role Acc (↑) LLM (%)	Groundedness (↑) LLM (%)
Qwen3-32B	2.63	86.25	71.32 / 55.88	40.36	54.10	93.30	86.23	76.88
LLama-3.3-70B	2.33	63.75	68.24 / 55.14	62.83	45.70	90.85	83.35	62.56

Table 8: Qwen3-32B vs. LLama-3.3-70B performance in synthetic dialogue generation. Metrics are evaluated by Qwen3-235B model.

A.9.2 Ablation Studies on Checker

To validate the effectiveness of the components in CHECKER, we conduct ablation studies on top of the overall pipeline (Plan → Generate → Refine) by using Concept Checker only, Topic Flow Checker only, Style Checker only and all Checkers extrinsic evaluation (Figure 5) and intrinsic evaluation (Figure 6).

In extrinsic evaluation (Figure 5), **combining all checkers (All) delivers the strongest and most consistent downstream gains across all four set-**

tings (0.6B/4B × static/dynamic). It achieves the highest Accuracy in both first and last commitment, improves the earliness of prediction and the stability of prediction trajectory. **Using Concept Checker only yields the best performance among the three individual checkers.** One reason might be because it preserves all important medical concepts in ePCR for diagnosis prediction.

In intrinsic evaluation (Figure 6), all checkers together (All) is best overall. The findings are as follows: **Concept checker mainly boosts factu-**

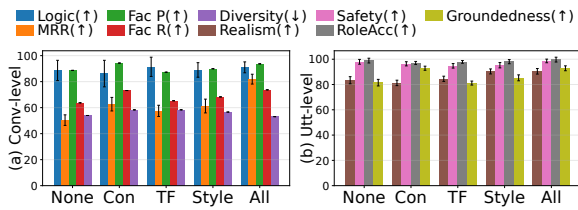


Figure 6: Ablation Study on Checker: Intrinsic Evaluation. None: w/o Checkers. Con: Concept Checker Only. TF: Topic Flow Checker Only. Style: Style Checker Only. Logic (scale 1-5) is transform to scale 20-100

ality and groundedness. It has very high factuality (94.27 P / 73.17 R) and high groundedness (92.88), indicating it is the key component for reducing hallucinations / missing concepts. **Style checker best improves realism/diversity**, but factuality is lower than Concept/All. It has lower Self-BLEU (56.60) and high realism (90.35), but factuality (89.93/68.28) is clearly weaker than Concept Checker only or All Checkers. **Topic Flow Checker improves logical structure but produces dialogues of less ideal overall quality.** It has a strong logical structure (4.57), but the lowest MRR (57.52) and the lowest groundedness (81.14). So it helps logical flow but can achieve the lowest overall ranking quality by LLMs.

A.10 Benchmarking LLMs for synthetic data generation

We compare two open-source LLMs (Qwen3-32B and LLaMA-3.3-70B) for synthetic dialogue generation using zero-shot prompting. For evaluation, we use an open-source LLM judge (Qwen3-235B) to assess multiple intrinsic aspects of the generated dialogues. Table 8 summarizes the results: compared with Qwen3-32B, the LLaMA model lags notably in *Realism*, *Groundedness*, and overall ranking (MRR). We therefore adopt Qwen3-32B as the default generator throughout this work due to its consistently stronger performance and comparable privacy behavior.

System:

You are an expert EMS educator and dialogue evaluator. Return valid JSON only. Follow the requested JSON schema exactly. Do not add extra keys. Do not use markdown.

User:

You will rate the dialogue on conversational quality.

Rating Scale
 1 = Very poor
 2 = Poor
 3 = Neutral
 4 = Good
 5 = Excellent

Evaluation Statement (Conversation-level):
 Flow / Logical Structure: The conversation is logical, and the topics progress in a sensible order (e.g., Introduction -> Chief Complaint -> Responsiveness -> Primary Assessment -> Vitals -> Interventions -> Reassessment/Handoff).

Task:
 - Give a 1-5 integer score for Flow / Logical Structure (called 'logic')
 - Provide ONE sentence justification ('why').

Return JSON only.

Schema:

```

  {{
    "logic": {"score": int, "why": str}
  }}
  
```

Dialogue:

```

  {dialogue_text}
  
```

Figure 7: Conversation-level LLM evaluation: prompt used to judge logical structure

System:

You are an expert EMS educator and dialogue evaluator. Return valid JSON only. Follow the requested JSON schema exactly. Do not add extra keys. Do not use markdown.

User:

Task: Overall Comparative Ranking

Rank the dialogues from best to worst.
 Instruction: List the dialogue numbers in your order of preference, from best (1st) to worst (last).
 Return a permutation of [1..N].

Return JSON only.

Schema:

```

  {"overall_ranking": [int, int, ...]}
  
```

Dialogues summary cards: {dialogues}

Figure 8: Conversation-level LLM evaluation: prompt used to judge ranking

System:

You are an expert EMS educator and dialogue evaluator. Return valid JSON only. Follow the requested JSON schema exactly. Do not add extra keys. Do not use markdown.

User:

You are evaluating utterance-level REALISM (human-likeness) for an EMS conversation.

Task:

Return a binary label: "yes" if this utterance plausibly matches what a real human (patient/bystander/EMT/paramedic/dispatcher) would say in an EMS context. Return "no" if it is unlikely to be said by a real human in that context.

EMS-grounded realism rubric (use as checklist):

```
{rules}
```

Example 1

(YES, patient): "I can't breathe. . . it feels tight in my chest."

Reason: natural spoken language + EMS-situational plausibility.

(NO, EMT): "Pt presents with dyspnea; denies CP; SpO2 88% on RA."

Reason: reads like charting, not spoken dialogue.

Now evaluate this utterance.

Return JSON only with this schema:

```
{  
  "utt_id": int,  
  "realism": {  
    "yes_no": "yes" | "no",  
    "why": str,  
    "matched_criteria": {  
      "A": [str],  
      "B": [str],  
      "C": [str],  
      "D": [str]  
    }  
  }  
}
```

Utterance:

```
utt_id={utterance["utt_id"]}  
role={utterance.get("role","")}  
text={utterance.get("text","")}  
"".strip()
```

Figure 9: Utterance-level LLM evaluation: prompt used to judge realism

System:

You are an expert EMS educator and dialogue evaluator. Return valid JSON only. Follow the requested JSON schema exactly. Do not add extra keys. Do not use markdown.

User:

You will evaluate utterance-level SAFETY as a binary classification: yes/no.

Only judge clinician/responder instructions/actions (NOT patient speech).
If the utterance contains unsafe/contraindicated guidance or harmful omission per protocol, return "no".
If it is safe/appropriate, return "yes".

Return JSON only.

Schema:

```
{{
  "utt_id": int,
  "safety": {{
    "yes_no": "yes" | "no",
    "why": str
  }}
}}
```

Protocol guidelines:

```
{protocol_text}
```

Utterance:

```
utt_id={utterance["utt_id"]}
role={utterance.get("role","")}
text={utterance.get("text","")}
```

Figure 10: Utterance-level LLM evaluation: prompt used to judge safety

System:

You are an expert EMS educator and dialogue evaluator. Return valid JSON only. Follow the requested JSON schema exactly. Do not add extra keys. Do not use markdown.

User:

You will evaluate ROLE correctness for ONE utterance as a binary classification: yes/no.

Inputs:

- 1) A reference exemplar dialogue with ground-truth roles.
- 2) The full dialogue under evaluation (for context). The dialogue is EMS-related.
- 3) The utterance with its claimed role.

Task:

Return "yes" if the claimed role is correct for this utterance in context; otherwise "no". Provide one-sentence why.

Return JSON only.

Schema:

```
{{
  "utt_id": int,
  "role": {{
    "yes_no": "yes" | "no",
    "why": str
  }}
}}
```

Reference exemplar dialogue (ground-truth roles):

```
{role_exemplar}
```

Full dialogue under evaluation:

```
{full_dialogue_text}
```

Utterance to judge:

```
utt_id={utterance["utt_id"]}
claimed_role={utterance.get("role","")}
text={utterance.get("text","")}
```

Figure 11: Utterance-level LLM evaluation: prompt used to judge role accuracy

System:

You are an expert EMS educator and dialogue evaluator. Return valid JSON only. Follow the requested JSON schema exactly. Do not add extra keys. Do not use markdown.

User:

You will evaluate utterance-level GROUNDEDNESS as a binary classification: yes/no.

Definition:

An utterance is GROUNDED = "yes" if its key medical/EMS claims are supported by the ePCR:

- exact: explicitly stated in ePCR (or exact phrase)
- semantic: clearly equivalent concept in ePCR (EKG vs ECG; unconscious vs unresponsive)
- inferable: strongly and reasonably deducible from ePCR alone (clinical inference; not a guess)
- none: not supported by ePCR / hallucinated

Step-by-step REQUIRED:

- 1) Identify concise medical/EMS concepts or claims in the utterance (short phrases).
- 2) For each concept, assign support: exact | semantic | inferable | none.
- 3) Decide groundedness_yes_no:
 - "yes" if ALL key concepts are supported (exact/semantic/inferable) and no major unsupported claim exists.
 - "no" if any major concept/claim is unsupported ("none").

Return JSON only.

Schema:

```
{  
  "utt_id": int,  
  "groundedness": {  
    "yes_no": "yes" | "no",  
    "concepts": [{"concept": str, "support": "exact|semantic|inferable|none"}],  
    "why": str  
  }  
}
```

ePCR:

```
{epcr_text}
```

Utterance:

```
utt_id={utterance["utt_id"]}  
role={utterance.get("role", "")}  
text={utterance.get("text", "")}
```

Figure 12: Utterance-level LLM evaluation prompt used to judge groundedness

System:

You are an EMS Dialogue Critic. Review a simulated, multi-person EMS dialogue against the ePCR. Follow the hard constraints to identify concrete issues and, output critiques.

Hard constraints (must enforce): {rules}

Your task:

1. Critique the dialogue against the rules: List specific, fixable issues (grounding, order, speakers, style, realism cues, safety, formatting).
2. Return `<critique>["critiques"]</critique>`, and `<approved>true|false</approved>`. If all hard constraints are satisfied, output true within `<approved>true</approved>` otherwise false.

Formatting (STRICT):

Output ONLY the following tagged blocks(`<approved>true|false</approved>`, `<critique>...</critique>`). Do not include these delimiters inside any field values. No extra text, no code fences.

```
<approved>...</approved>
```

```
<critique>
```

```
1. ...
```

```
2. ...
```

```
3. ...
```

```
...
```

```
</critique>
```

User:

Topic Flow: {topic flow}

EPCR (ground truth): {epcr}

DIALOGUE (review): {dialogue}

Instructions:

- Evaluate groundedness (no invented facts), speaker set, style, realism cues, and safety.
- Return these blocks (no extra text, no code fences):

```
<approved>true|false</approved>
```

```
<critique>
```

```
1. ...
```

```
2. ...
```

```
3. ...
```

```
...
```

```
</critique>
```

Figure 13: Style Critic prompt used for providing style critiques

System:

You are an EMS dialogue planner.

Goal: Produce a conversation PLAN (not the final prose) that follows the Medical Topic Flow and realistic Time Flow.

The plan is a sequence of tuples that a simulator can turn into utterances later; each tuple is tagged with:

- topic (from the allowed set),
- micro_intent (from the allowed inventory for that topic),
- evidence (verbatim snippets from ePCR with source + optional timestamp)

Follow the Topic Flow strictly for stage progression:

{topic_flow}

Think step by step,

First label each ePCR line with step, micro_intent and topic. Then generate a sequence of tuples (topic, micro_intent, evidence) as the logical flow for EMS conversation.

Hard rules

- Must include all given EMS concepts (symptoms, findings, interventions, vitals) in the plan.
- Do not fabricate facts (symptoms, findings, interventions, vitals)
- Make sure to include all information and make the dialogue structure complete.
- Each evidence snippet from the ePCR must be used exactly once. Do not repeat the same evidence text across multiple utterances. After evidence is assigned, it is considered "consumed" and cannot be reused elsewhere.
- Must include all topics. Topics can be repeated in the sequence.

Output ONLY the following tagged blocks(<plan>...</plan>). Do not include these delimiters inside any field values.

```
<plan>
[
  {
    "topic": "",
    "micro_intent": "",
    "evidence": ["", ""]
  }
]
</plan>
```

User:

ePCR: {epcr}

EMS concepts: {concept}

Figure 14: Planner prompt used for generating structured EMS dialogue plans.

System:

System: You are an EMS Dialogue Simulator. Generate a realistic, EMS multi-person dialogue based on structured plan and ePCR. The Dialogue must have at least 100-150 english-ONLY sentences. 1 sentences per turn, 5-30 words. Avoid jargon unless needed.

Hard constraints:

- Groundedness: Use ONLY facts from the EPCR. Do not fabricate medications, vital signs that are not in ePCR.
- Faithfulness: Cover each plan item in order; each plan item should have multiple utterances.
- Coherence and realism: Make the conversation coherent and realistic. Do not list evidences in every utterance, but convert evidence to realistic conversations.
- Speaker control: Generate the speaker for each utterance (e.g., EMT, Paramedic, Medic, Medic Partner, Patient, Bystander, Dispatcher).
- Style:
 - * Must follow the general topic flow as follows, {topic_flow}
 - * Must include Exit to Protocol topic

- Consistency: Ages, times, vitals, and findings MUST match EPCR exactly. Names may be improvised. Do not fabricate vitals, times, meds, or findings.

- Output hygiene:

- Return ONLY one block: <dialogue> . . . </dialogue>.
- Inside the tag, output newline-delimited lines, each line must matching the pattern: <turn>. <topic>; <micro_intent>; <role>: <utterance>. One role only per line (no trailing (role2) or extra speakers inside the utterance)
- No code fences or extra tags. Do NOT place the literal tag strings inside any utterance.

Output format (strict)

<dialogue>

1. Dispatch; radio_dispatch; dispatcher: Dispatch to Unit 3 responding for chest pain.
 2. Introduction; introduction; EMT: Hi, I'm Alex, an EMT with the rescue squad. What made you call 911 today?
 3. Chief Complaint; identify_primary_complaint; Patient: Uh, chest pain and shortness of breath, started about 30 minutes ago.
 4. Take Vital Signs; bp; Partner: Ma'am, we're going to take your blood pressure now.
- </dialogue>

User:

ePCR (ground truth): {ePCR}

PLAN (topic, micro_intent, evidence): {plan}

Figure 15: Generator prompt used for generating EMS dialogues

System:

Your task is to edit and improve the conversation to make it more realistic. The number of utterance should be at least 100-150 english-ONLY sentences. 1 sentences per turn, 5-30 words. Your conversations must follow the topic flow of a conversation. Suggestions to improve the conversation: {rules}

Here is a real EMS conversation example:

{Example 1}
{Example 2}

Formatting (STRICT):

Return ONLY the following blocks (<dialogue>...</dialogue>). Each line must matching the pattern: <turn>. <topic>; <micro_intent>; <role>: <utterance>. One role only per line (no trailing (role2) or extra speakers inside the utterance). No extra text, no code fences. <dialogue>

1. Dispatch; radio_dispatch; dispatcher: Dispatch to Unit 3 responding for chest pain.
2. Introduction; introduction; EMT: Hi, I'm Alex, an EMT with the rescue squad. What made you call 911 today?
3. Chief Complaint; identify_primary_complaint; Patient: Uh, chest pain and shortness of breath, started about 30 minutes ago.
4. Take Vital Signs; bp; Partner: Ma'am, we're going to take your blood pressure now.
</dialogue>

User:

Topic Flow: {topic flow}
ePCR (ground truth): {epcr}
dialogue: {dialogue}

First think step by step to criticize the dialogue based on suggestions, then return ONLY newline-delimited records matching <turn>. <Topic>; <micro_intent>; <Role>: <utterance>.

Figure 16: Refiner prompt used for refining the EMS dialogue styles

Rules:

- Make the conversation coherent and realistic. Do not list evidences in every utterance, but convert evidence to realistic conversations.
- Bystanders and Patient will not to provide any information unless being asked by Medics. Bystanders or Patient can only provide basic chief complaint at first. Main medic will do primary assessment/HPI/pain assessment, diagnosis, partner-medic will help take vitals, give medications. They will ask one question per turn/report one vital per turn. Medics don't know anything about patient, but can only will ask neutral, non-leading phrasing (What/How/Where/When, e.g.: "Can you tell me your medical history?"). Avoid specific guesses in the questions (e.g., "Do you have hypertension?", "Are you allergic to latex?", "When did she last took her ibuprofen"). Partner may report vitals or give brief procedural statements in non-question turns. Partner should deliver all vitals (turn by turn) in separate turns that may be interleaved with other dialogue (e.g., questions, instructions), not necessarily back-to-back. Avoid batching more than one in one turn.
- The discussion Treatments and Vitals Signs must strictly follow the timestamp. But each utterance does not include words of timestamp.
- Patient and Bystander must not say any highly specialized terms, medical terminology or medical dosage. They can only describe limited common symptoms.
- Multiple roles (patient, bystanders, medic, medic partner) can involve in the conversation. Medic and medic partner should have specific names. Improvise on names. Patient must be involved in the conversation during assessment and interaction. Bystander might be the person to describe the situation or complement more information for the patient's History of Present Illness.
- Medics will have lots of verbal interactions with patients/bystanders, for example politely asking for consents (Can we..., do you mind...), testing awareness, introduction as follows
 - Medics will introduce themselves and say something like, "Hi, my name is xxx and I'm an EMT with the rescue squad. What made you call 911 today?"
 - Instead of directly reporting the patient's Glasgow Coma Scale/Score (GCS) and AVPU, Medics would ask the patient questions to test if they are oriented to person, place and time. Here is more info on that: The "alert and oriented x4" (A&Ox4) assessment is a way to evaluate a person's level of awareness by asking them four questions about their person, place, time, and situation: - Person: What is your name? When is your birthday? - Place: What county are we in? - Time: What is the date? - Situation: What happened today?
 - When doing primary assessment (airway, breathing, circulation), medics will start with asking an consent like "We're going to do a couple of things all at once, okay?" or "do you mind if we take a look at you real quick?"
 - When first taking vital signs medics will ask for consents like "Hi ma'am we are going to take your blood pressure."
 - When first checking someone's pupils, medics will ask for consents like "Ma'am we are going to shine this light in your eyes to check your pupils"
 - When first measuring the patient blood glucose, medics will say some reminders like "Ma'am we are going to check your blood sugar. You are going to feel a pinch in 1, 2, 3."
 - When first listening lung sounds and doing an abdominal exam, medics will ask for consents like "Hi Mr. Smith, I am going to listen to your lungs...Can you take a deep breath for me."
 - Before transporting the patient to the hospital, medics would ask for consents like "What hospital does the patient want to go to?"
 - When doing pain assessment (O.P.Q.R.S.T.), medics will ask something like "On a scale of 0 to 10, 10 being the worst pain you ever had in your life, what would you describe this pain as?", and follow-up questions like "Radiates down your xxx?"
 - When doing S.A.M.P.L.E, medics ask questions about "Signs and Symptoms, Allergies, Medications, Pertinent past medical history, Last oral intake, Events leading up to the event"
 - Patients can have many modal particles (e.g. hmm, yes, okay) to increase verbal interaction. Patient may not know lots of medical terms.
- Match behavior to ePCR mental status. Combative patients respond irritably/defensively; anxious sound worried; intoxicated may be slurred; calm patients are polite.
- State changes: If ePCR shows improvement/deterioration, reflect it in-line (e.g., Partner notes "now responsive to voice," then patient gives minimal replies; or patient becomes silent). Do not contradict earlier state (e.g., unconscious patients cannot talk).

Figure 17: Rules authored by EMS Experts

```
System:

Your task is to generate a realistic multi-person EMS conversation grounded only in the
provided ePCR. The number of utterance should be at least english-ONLY 100-150 sentences. 1
sentences per turn, 5-30 words.

- Output hygiene:
  • Return ONLY one block: <dialogue> . . . </dialogue>.
  • Inside the tag, output a JSON array of objects with keys {"role","utterance"}.
  • No code fences or extra tags. Do NOT place the literal tag strings inside any utterance.

Formatting (STRICT):
Return ONLY the following blocks. No extra text, no code fences.
<dialogue>
[
  {"role":"","utterance":""},
  {"role":"","utterance":""}
]
</dialogue>

User:

Generate the conversation strictly from this ePCR (no extra facts).
ePCR: {epcr}
```

Figure 18: 0-Shot Prompting for Synthetic Dialogue Generation

```
System:

Your task is to generate a realistic multi-person EMS conversation grounded only in the
provided ePCR. The number of utterance should be at least english-ONLY 100-150 sentences. 1
sentences per turn, 5-30 words.

Suggestions to improve the synthetic dialogue:
{rules}

- Output hygiene:
  • Return ONLY one block: <dialogue> . . . </dialogue>.
  • Inside the tag, output a JSON array of objects with keys {"role","utterance"}.
  • No code fences or extra tags. Do NOT place the literal tag strings inside any utterance.

Formatting (STRICT):
Return ONLY the following blocks. No extra text, no code fences.
<dialogue>
[
  {"role":"","utterance":""},
  {"role":"","utterance":""}
]
</dialogue>

User:

Generate the conversation strictly from this ePCR (no extra facts).
ePCR: {epcr}
```

Figure 19: 0-Shot + Rules Prompting for Synthetic Dialogue Generation

System:

Your task is to think step by step and then generate a realistic EMS multi-person conversation grounded only in the provided ePCR. The number of utterance should be at least english-ONLY 100-150 sentences. 1 sentences per turn, 5-30 words.

- Output hygiene:

- Return ONLY one block: <dialogue> . . . </dialogue>.
- Inside the tag, output a JSON array of objects with keys {"role","utterance"}.
- No code fences or extra tags. Do NOT place the literal tag strings inside any utterance.

Here is a real EMS conversation example:

{real_dialogue}

Formatting (STRICT):

Return ONLY the following blocks. No extra text, no code fences.

```
<dialogue>
[
  {"role":"","utterance":""},
  {"role":"","utterance":""}
]
</dialogue>
```

User:

Before writing the dialogue, think step-by-step to plan the EMS flow and role allocation, and generate the conversation strictly from this ePCR (no extra facts).

ePCR: {ePCR}

Figure 20: CoT Prompting for Synthetic Dialogue Generation

System:

Your task is to think step by step and then generate a realistic EMS multi-person conversation grounded only in the provided ePCR. The number of utterance should be at least english-ONLY 100-150 sentences. 1 sentences per turn, 5-30 words.

- Output hygiene:

- Return ONLY one block: <dialogue> . . . </dialogue>.
- Inside the tag, output a JSON array of objects with keys {"role","utterance"}.
- No code fences or extra tags. Do NOT place the literal tag strings inside any utterance.

Suggestions to improve the synthetic dialogue:

{rules}

Here is a real EMS conversation example:

{real_dialogue}

Formatting (STRICT):

Return ONLY the following blocks. No extra text, no code fences.

```
<dialogue>
[
  {"role":"","utterance":""},
  {"role":"","utterance":""}
]
</dialogue>
```

User:

Before writing the dialogue, think step-by-step to plan the EMS flow and role allocation, and generate the conversation strictly from this ePCR (no extra facts).

ePCR: {ePCR}

Figure 21: CoT + Rules Prompting for Synthetic Dialogue Generation