TRACER: Early Failure Detection for Task-Oriented Dialogue

Erfan Nourbakhsh, Rocky Slavin, Ke Yang, and Anthony Rios

University of Texas at San Antonio
{erfan.nourbakhsh, anthony.rios}@utsa.edu

SIGdial 2026 Atlanta, Georgia, USA

Introduction & Problem Statement

- Task-oriented dialogue (TOD) systems often fail progressively before a breakdown becomes obvious, through **slot-value instability**, **unresolved contradictions**, **missing constraints**, and a **loss of task focus**.
- **The Evaluation Gap:** The dominant evaluation paradigm measures **outcome rather than trajectory**, such as whether the final state is correct or the overall task succeeded. By the time this final outcome is measured, the dialogue has often **drifted too far** for a lightweight intervention to help.
- Existing breakdown detection models only predict whether the next turn causes a breakdown, and do not forecast eventual task failure from **partial dialogue prefixes**.
- **Our Goal:** To formulate task-oriented dialogue failure prediction as a **partial-dialogue forecasting problem**. TRACER aims to detect rising failure risk early so the system can **decide whether to intervene** before the final breakdown.

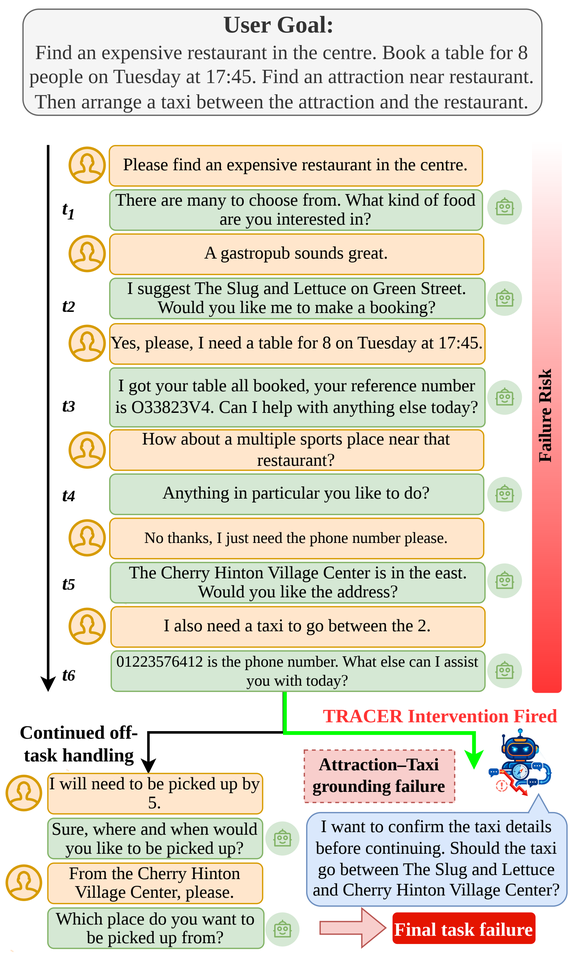

Fig 1. Task overview. The example shows a conversation that breaks down

Key Forecasting Results

Category	Model	Belief State	AUC-ROC	F1	EDS	Det. Turn
Heuristic Trigger	No intervention	Oracle	0.500	0.000	0.000	∞
	Fixed-schedule (every 3)	Oracle	0.499	0.236	0.690	2.00
	Slot-confidence (Osc>0)	Oracle	0.495	0.025	0.009	7.50
	Feature-threshold ensemble	Oracle	0.535	0.286	0.864	0.25
Classical Feature Model	Partial dialogue logreg (turn-level)	Oracle	0.574	0.344	0.849	0.35
	Partial dialogue logreg (engineered)	Oracle	0.600	0.351	0.728	1.03
Zero-ShotPrompting	Llama-3.1-8B	Oracle	0.519	0.268	0.474	2.59
	Mistral-7B	Oracle	0.525	0.278	0.558	2.08
	Qwen2.5-7B	Oracle	0.531	0.321	0.772	1.32
Zero-Shot CoT Prompting	Llama-3.1-8B	Oracle	0.516	0.290	0.566	2.49
	Mistral-7B	Oracle	0.526	0.275	0.568	1.95
	Qwen2.5-7B	Oracle	0.519	0.318	0.756	1.34
Few-ShotPrompting	Llama-3.1-8B	Oracle	0.492	0.337	0.993	0.06
	Mistral-7B	Oracle	0.524	0.312	0.648	1.89
	Qwen2.5-7B	Oracle	0.518	0.299	0.668	1.67
Our Approach	TRACER(Oracle)	Oracle	**0.617**	**0.370**	0.751	1.22
	TRACER(generated)	Generated	**0.741**	**0.463**	0.727	1.36

Table 1. Main failure-forecasting results on MultiWOZ

- TRACER achieves an **AUC-ROC of 0.741** and an **F1 of 0.463** (on oracle states), significantly outperforming classical models and the best zero-shot LLMs.
- **Early Failure Signals:** Meaningful forecasting occurs very early, achieving **0.666 AUC-ROC with only 25%** of the dialogue visible.
- By **75% visibility**, performance reaches **0.811 AUC-ROC**, which is nearly identical to its full-context discrimination quality (0.819).

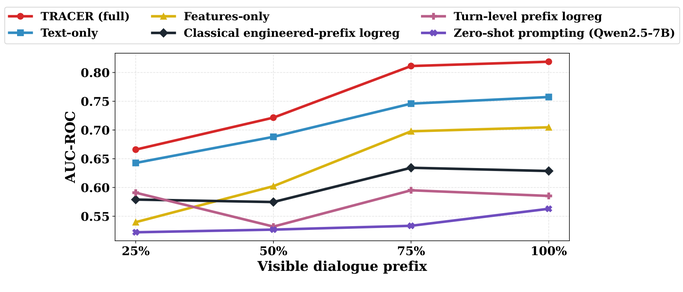

Fig 4. Fixed-context forecasting

- Simple heuristic baselines reach high Early Detection Scores (e.g., 0.864) by triggering almost immediately, causing **unusable false-positive rates**.
- TRACER solves the EDS trap by **balancing high early detection with selective triggering**.

Experimental Setup

- **Primary Dataset:** Evaluated on **MultiWOZ 2.4** (multi-domain dialogues).
- **Belief States:** Compared **oracle** (clean inputs) vs. **generated** (noisy inputs via Qwen3-30B).
- **Metrics:** AUC-ROC, F1, Early Detection Score (EDS), and Mean Detection Turn.

Methodology

We introduce a method with two sets of features to predict if an ongoing conversation will fail.

Temporal Trajectory Dynamics: At each turn, we track five simple signs of how the dialogue is progressing:

- **Oscillation score:** How often the user changes their mind about a specific detail (e.g., changing the hotel price repeatedly).
  - ➢ **Example:** A user asks for a cheap hotel, changes to expensive, then back to cheap.
- **Coverage rate:** The percentage of required information the system has successfully gathered.
  - ➢ **Example:** The user gives a hotel area and type, but forgets the price (2 out of 3 needed slots filled = 67% coverage).
- **Conflict count:** When a user's new input clashes with what the system previously understood.
  - ➢ **Example:** The system records "north," but the user suddenly says, "Actually, I would prefer the south side".
- **Fill velocity:** How quickly new details are being added over the last couple of turns.
  - ➢ **Example:** The system gathers 2 new slots (specific details like a hotel's price or location) over 2 turns (back-and-forth exchanges between the user and the system), showing steady progress.
- **Domain shifts:** How many different topics (like hotels, restaurants, or taxis) the user has jumped between.
  - ➢ **Example:** Discussing a hotel, switching to a restaurant, and then booking a train (3 domains).

Belief-State Text: Instead of using complex code, we turn the system's current knowledge into simple text sentences (e.g., "restaurant area is center"). This allows the model to read and understand exactly what is being tracked.

Training Objective: The model learns to predict success or failure, but it earns a **higher reward for spotting failures as early as possible** in the conversation.

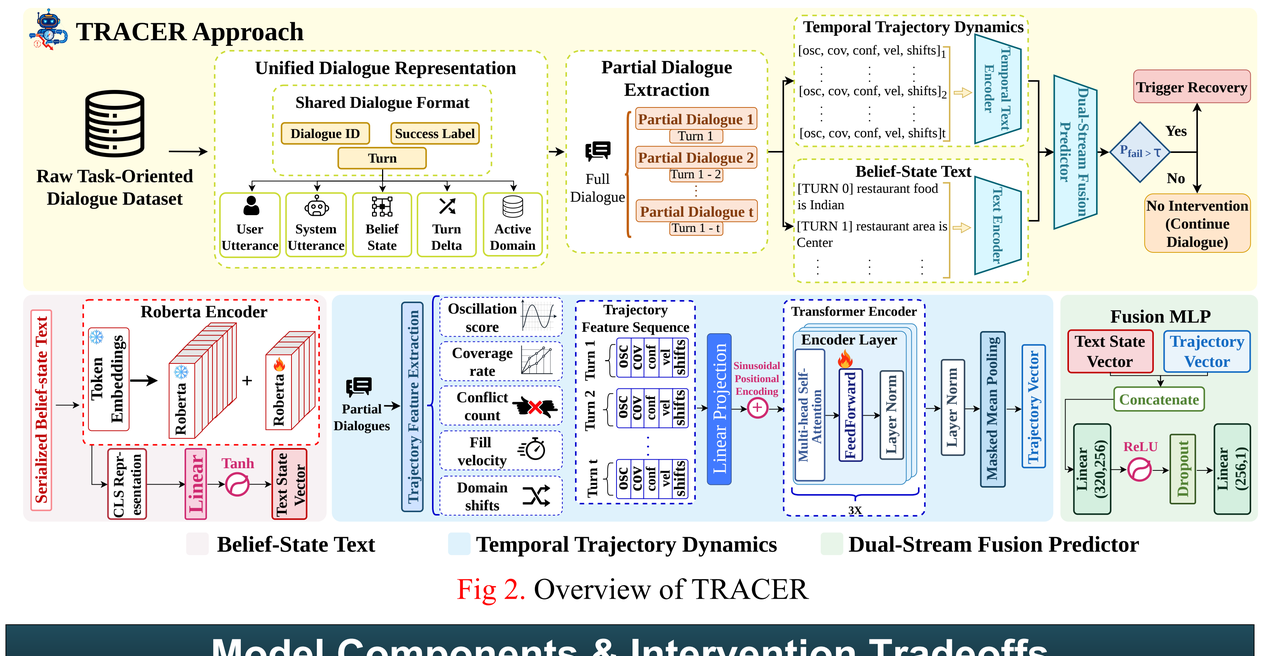

Fig 2. Overview of TRACER

Model Components & Intervention Tradeoffs

- Optimizing solely for early detection creates an **unusable 100% false-positive rate**, requiring a balance against unnecessary interruptions.
- TRACER applies a **conservative frozen threshold (τ = 0.7788)** tuned on the development set.
- This threshold ensures a **strict false-positive rate of ≤ 0.15** while successfully **detecting 54% of failures** at an average early detection turn of **3.70**.

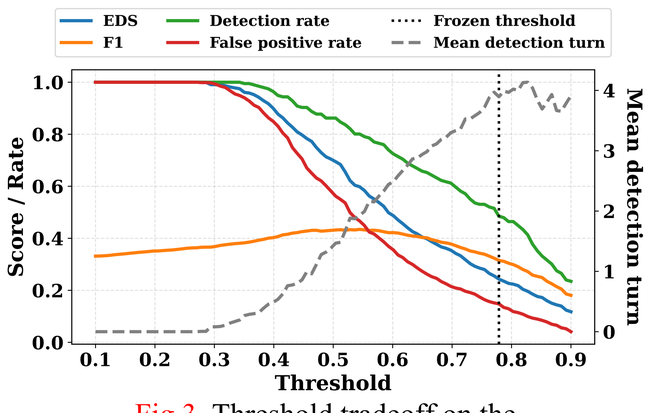

Fig 3. Threshold tradeoff on the development set

Model	Belief State Type	AUC-ROC	F1
Stream A only (features)	Generated	0.554	0.337
	Oracle	0.623	0.363
Stream B only (text)	Generated	0.583	0.332
	Oracle	0.692	0.432
TRACER	Generated	0.617	0.370
	Oracle	**0.741**	**0.463**

Table 2. Ablation on TRACER streams

- The fully fused model achieves **0.741 AUC-ROC**, significantly outperforming isolated features-only (0.623) and text-only (0.692) streams.
- Trajectory dynamics and semantic text provide **complementary signals that persist** even with noisier, generated belief states.

Conclusion

- TRACER proves that task-oriented dialogue failure can be reliably predicted **from partial conversations** by combining **trajectory dynamics** with **belief-state text**.
- Error analysis reveals that TOD failures fall into **four** main categories: **Information Incompleteness**, **Contradiction / Misalignment**, **State Instability**, and **Task Drift / Domain Confusion**.