Building Task-Oriented Dialogue Systems via Instruction Guidance without Annotated Data

Henry Gao and Jinho D. Choi, Emory University (Atlanta, GA)

Motivation

- Task-oriented dialogue (TOD) systems rely on costly **supervised fine-tuning** and annotated datasets, limiting scalability and domain transfer.
- LLMs have strong reasoning abilities, but their **effectiveness as TOD agents** without fine-tuning remains under-explored.
- **Automatically induced procedural instructions** from dialogue logs provide an interpretable and scalable alternative to fine-tuning.
- **Interactive evaluation** based on goal completion better captures flexible dialogue behavior than fixed trajectory matching.

Interactive User Simulator

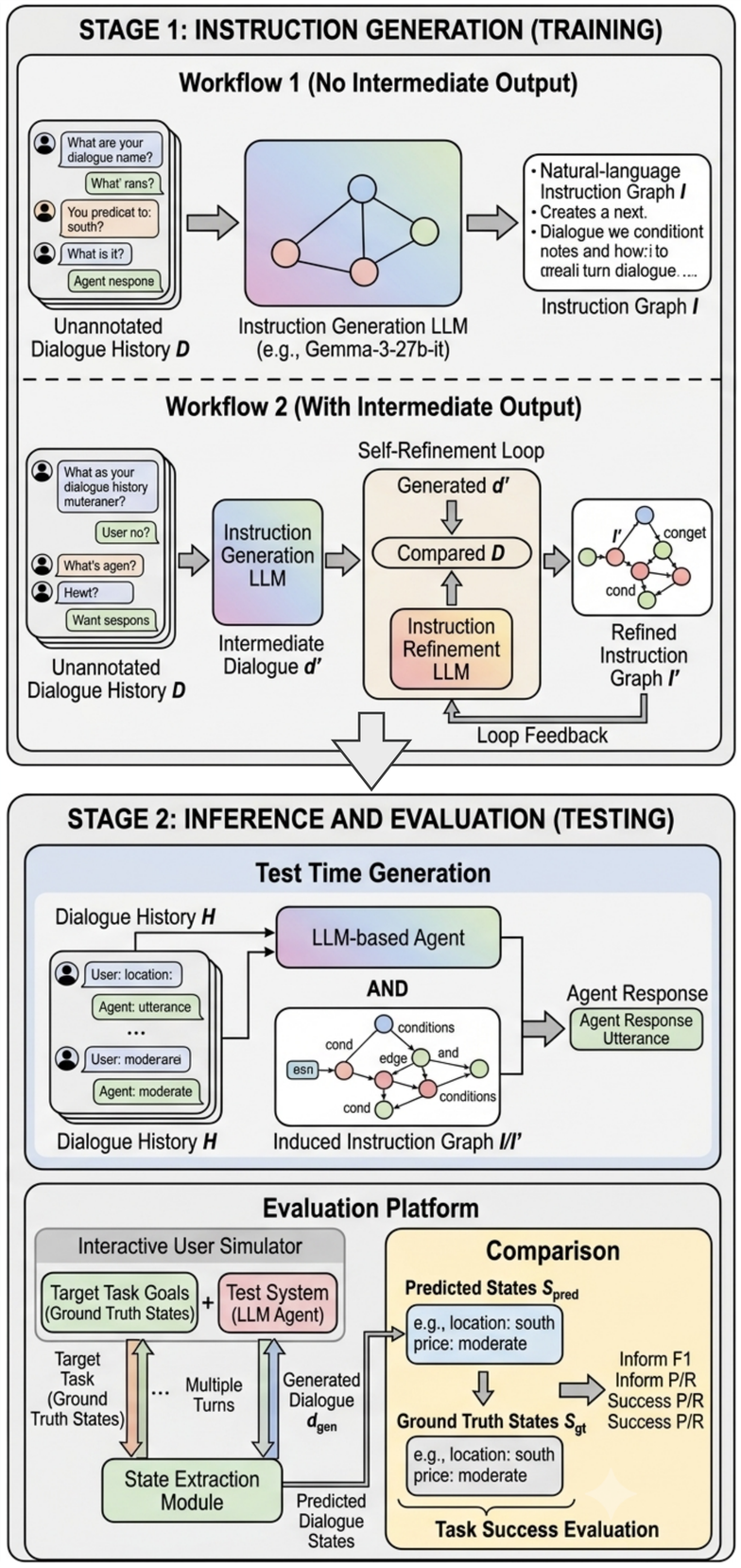
- Unlike traditional simulators that actively volunteer information, our simulator operates under strict **information-denial rules**: it withholds slots and answers with "I don't know" unless explicitly queried.
- By refusing to steer or maintain dialogue flow independently, the simulator forces the dialogue agent to actively select dialogue acts, **solicit missing parameters, and manage uncertainty**.
- Built using in-context prompting, the simulator **responds naturally to agent mistakes**, reminding the agent to correct errors and continuing multi-turn interactions until task goals are met or unresolvable.

Interactive Evaluation Framework

- Addresses a fundamental flaw in standard benchmarks by **assessing task success across dynamic**, non-canonical dialogue paths rather than penalizing valid, non-ground-truth action sequences.
- Evaluates generated system turns against the agent's own cumulative context rather than feeding ground-truth past turns, enabling **realistic measurement of error propagation and recovery**.
- Uses an independent **state-extraction** module to extract predicted slot-value pairs from **full generated conversations**, computing fine-grained Precision, Recall, and F1 metrics for both Inform and Success states.

Evaluation Metrics

- **Inform Precision/Recall**: % of slots in generated dialogue that appear in reference dialogue, and vice versa.
- **Inform Correctness Precision/Recall**: % of Inform Precision/Recall slots with correct value.
- **Success Precision/Recall**: % of requested slots fulfilled in generated dialogue that are also fulfilled in reference dialogue, and vice versa.



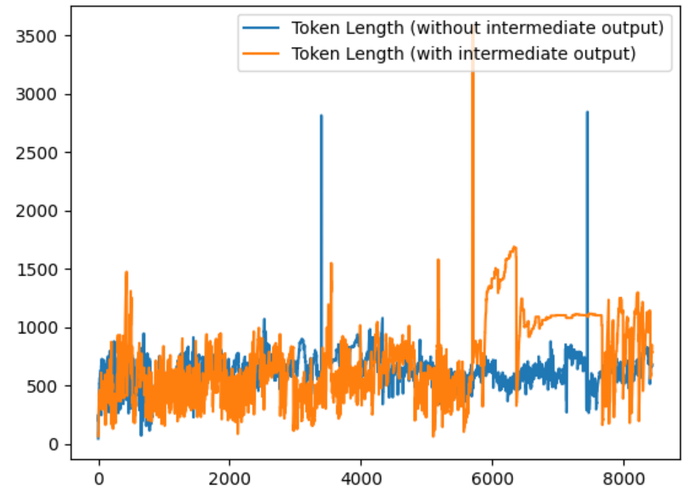
Experiments

- **Models**: Gemma-3-27b-it, Qwen3-30b-a3b, Gemma-4-26b-a4b-it
- **State Extraction**: Slot Schema Induction (Finch et al., 2025)
- **Dataset**: MultiWOZ dataset (Budzianowski et al., 2018)
- **Baseline**: GALAXY (He et al., 2021), MARS (Sun et al., 2023)

Model	IP	ICP	IR	ICR	I-F1	SP	SR
GALAXY	88.4	75.4	80.6	69.8	84.3	35.6	36.7
MARS	89.5	77.9	80.2	69.7	84.6	35.5	36.5
Qwen	71.1	64.4	83.9	76.1	70.0	67.7	72.2
Qwen+I	80.3	73.2	72.2	65.9	76.0	46.4	48.2
Gem3	81.1	74.6	92.1	85.1	86.3	65.1	69.9
Gem3+I	81.1	74.9	90.0	83.2	85.3	58.9	63.3
Gem4	77.3	70.1	84.2	76.7	80.7	71.9	76.4
Gem4+I	77.0	69.5	85.0	76.9	80.8	70.7	75.7

Analysis

- LLMs can reliably navigate **instruction graphs** to perform core dialogue acts without turn-level supervision.
- Pretrained LLMs have high dialogue recall, but explicit instruction graphs are necessary to enforce **domain-specific policies**.
- Label inconsistencies in MultiWOZ artificially lowers evaluation metrics, meaning **real-world model accuracy** is likely higher.



X: The number of dialogues processed
Y: The token length of instructions