When the Database Fails: Prompting LLM Dialogue Agents for Safe Recovery in Task-Oriented Dialogue

Mohammad Alijanpour Shalmani[1]* · Alale Rezvani Boroujeni[2] · Jiann Shiun Yuan[1]

[1]College of Engineering and Computer Science [2]Department of Marketing · University of Central Florida, Orlando, FL

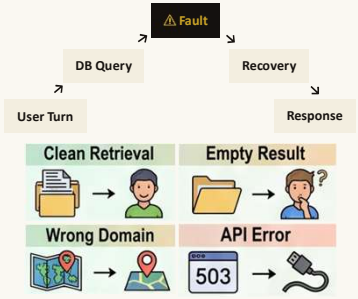
SIGdial 2026 — 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Atlanta

Introduction

Task-oriented dialogue (TOD) systems help users book hotels, find restaurants, and arrange transport.
LLMs in task-oriented dialogue often produce fluent but unsafe responses when backend database calls fail - inventing venues, confirmations, or booking details ungrounded in the database.
We test a lightweight, retraining-free **Guided-Retry** prompt across six open-weight model families and four DB conditions on MultiWOZ 2.2 and SGD, cutting hallucination **50%** (MultiWOZ) and **42%** (SGD) - though **6-37%** residual hallucination remains.

Motivation

TOD agents extract intent, query a database, and generate a response grounded in the result - a pipeline fragile at the database boundary:



Empty result - no record matches

Wrong domain - incorrect domain routed

API error - timeout or server failure

Contributions

1. A controlled fault-injection framework over MultiWOZ 2.2 and SGD

2. Evaluation of three prompting strategies across six model families

3. Automatic + human-validated metrics, incl. Commitment Safety Rate (CSR)

4. Structured prompting reduces but does not eliminate hallucination (6-37% residual)

Related Work

Prior fault-injection work targets general agent tasks (TRACE/SCOPE; ReliabilityBench) or trains recovery policies over injected malfunctions (PALADIN), not prompting-only recovery grounded in TOD. No prior study injects runtime DB faults on MultiWOZ and SGD while comparing prompting strategies without retraining.

Method

5	20	400	4
MultiWOZ dom.	SGD dom.	dlgs/dataset	conditions

MultiWOZ: hotel, restaurant, taxi, train, attraction.

SGD (20 domains): restaurants, flights, hotels, events, media, and more.

Same decoding (temp 0.2); strategies differ only in system prompt:

Strategy	Behavior
Naive	Raw DB response, no instruction
Inform	Told not to hallucinate; acknowledge failure
Guided-Retry	Structured procedure on DB status; no-fabrication rule

Stats: McNemar's exact test (HR, AAR), paired t-test (UFS); all comparisons significant at $p < 0.001$.

Guided-Retry excerpt: "status=wrong_domain: state the result appears to be from the wrong domain. Ask the user to confirm... NEVER fabricate a booking reference, venue name, or confirmation detail."

Guided-Retry is a single structured system prompt keyed on DB status - no retraining, no extra inference calls.

Metrics

HR - Hallucination Rate: fabricated results absent from the DB response

AAR - Appropriate Action Rate: matches ground-truth correct action

CSR_{auto} - automatic Commitment Safety Rate; narrower than HR, human-validated separately

UFS - User Friction Score, 0-3 composite (lower is better)

Results
Naive → Guided-Retry

Model	MWOZ HR↓	MWOZ AAR↑	SGD HR↓	SGD AAR↑
DeepSeek-R1	**37.8→5.8**	51.7→72.0	**33.2→5.2**	47.5→73.5
Gemma-2	12.5→6.8	65.8→75.0	7.2→5.8	55.8→81.0
Llama-3	28.5→16.2	**82.5→99.8**	17.2→9.5	**68.2→97.8**
Mistral	30.8→14.8	68.5→93.8	18.5→8.8	60.0→96.2
Phi-3*	44.5→35.2	68.2→81.0	32.2→37.0	62.3→76.2
Qwen-2.5	28.7→13.2	62.3→76.5	16.8→6.8	58.2→78.5

Exception - Phi-3: only model where Guided-Retry HR>Inform HR (MWOZ 31.0→35.2; SGD 25.2→37.0).

Cross-Dataset Average (6 models)

MultiWOZ 2.2 - HR↓ / AAR↑

30.5→20.6→15.3

66.5→71.1→83.0

Naive → Inform → Guided-Retry

SGD - HR↓ / AAR↑

20.9→14.0→12.2

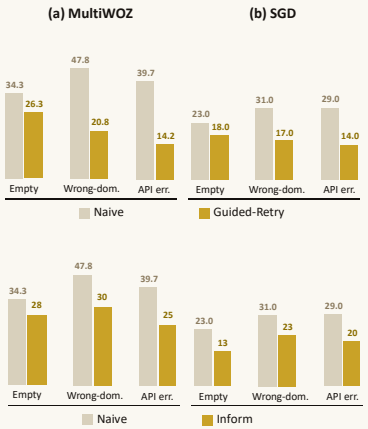
58.7→62.6→83.9

Naive → Inform → Guided-Retry

AAR rises from **66.5%→83.0%** (MultiWOZ) and **58.7%→83.9%** (SGD) under Guided-Retry - consistent across two structurally different benchmarks.

Wrong-domain is hardest: HR falls only **47.8%→20.8%** (MultiWOZ) and **31.0%→17.0%** (SGD) - far less than API-error or empty-result gains.

Hallucination Rate by Failure Type



Wrong-domain retrieval is hardest on both datasets; API errors show the clearest benefit from structured recovery.

Empty-result gains are more modest: MultiWOZ 34.3→26.3%, SGD 23.0→18.0%.

User Friction Score (mean, 6 models)

0-3 composite per response (lower = better); not a percentage.

$$UFS_i = \min(2H_i + F_i + S_i, 3)$$

Hallucination flag (+2), false Commitment flag (+1), and silent failure (no apology/acknowledgement, +1), capped at 3.

Strategy	MWOZ	SGD
Naive	0.67	0.54
Inform	0.45	0.34
Guided-Retry	0.43	0.41

Human Evaluation

9 annotators, 60 sampled responses, blind to strategy/condition.

κ=0.767	95.7%
Fleiss' agreement	observed agree.

Human CSR: **Guided-Retry 100.0%** , Inform 89.5%, Naive 81.0%.

Key Findings

Structured prompting reduces hallucination **42-50%** - no **50%** - no retraining, no extra inference calls.

Residual hallucination stays substantial: **5.8%** (best) to **35.2%** (worst).

Not every model benefits - Phi-3 does not reliably follow structured recovery instructions.

Findings hold across two structurally different benchmarks and all six model families.

Statistical Significance

DeepSeek-R1	Gemma-2	Llama-3
✓$p<.001$	✓$p<.001$	✓$p<.001$
Mistral	**Phi-3**	**Qwen-2.5**
✓$p<.001$	✓$p<.001$	✓$p<.001$

McNemar's exact test (HR, AAR) and paired t-test (UFS), Guided-Retry vs. Naive - significant at $p < 0.001$, both datasets.

Limitations & Conclusion

Synthetic fault injection; heuristic detectors; models limited to 7-9B scale; no multi-turn recovery dynamics studied.

Guided-Retry reduces hallucination 42-50% without retraining, but residual hallucination of 6-37% remains - wrong-domain retrieval is the hardest case for future work.

References

Hudeček & Dušek 2023 · Zang et al. 2020 (MultiWOZ 2.2) · Rastogi et al. 2020 (SGD) · Guo et al. 2025 (DeepSeek-R1)

Code & prompts
github.com/mohammad-AJP/llm-db-failure-recovery