Evaluating Business-Policy Adherence of Customer Support LLM Agents

Anonymous Author(s)

ABSTRACT

We develop standardized evaluation methodologies for assessing whether LLM-based customer support agents adhere to business rules and multi-step support workflows. Motivated by Balaji et al. [1], who introduced JourneyBench but identified reliable adherence assessment as an open challenge, we systematically evaluate five LLM agents across five SOP complexity levels, five disturbance conditions, and four evaluation methodologies. Claude-3.5 achieves the highest User Journey Completion Score (UJCS) of 0.829 ± 0.002 on 5-step SOPs, while Llama-70B scores lowest at 0.679. UJCS degrades with SOP complexity: all agents lose $\geq$15 points moving from 3-step to 20-step SOPs, with Llama-70B showing the steepest decline. Under disturbances, tool failures cause the largest degradation (3.6 points for Claude-3.5). Among evaluation methods, the hybrid approach (rule-based + LLM-judge) achieves the best balance with F1=0.907 and coverage=0.927, approaching human expert performance (F1=0.952, coverage=0.987) at a fraction of the cost. Multi-turn analysis reveals adherence decays linearly with conversation length, with less robust agents losing up to 4.6 points per turn.

1 INTRODUCTION

LLM-based agents are increasingly deployed for customer support, replacing traditional Interactive Voice Response (IVR) systems with flexible multi-turn interactions [2]. However, these agents must comply with business rules encoded in Standard Operating Procedures (SOPs)—a requirement that existing benchmarks focused on tool selection [4] or goal completion [3] do not adequately measure.

Balaji et al. [1] introduced JourneyBench to address this gap, using SOP graphs and a User Journey Completion Score (UJCS) metric. However, they identified reliable evaluation of policy adherence as a central open challenge, particularly for complex multi-step workflows with dependencies and real-world disturbances.

We address this problem through five experiments: (1) comparing five LLM agents on standard SOPs, (2) measuring adherence degradation with SOP complexity, (3) evaluating robustness under disturbances, (4) comparing evaluation methodologies, and (5) analyzing multi-turn consistency. Our key findings are that hybrid evaluation (rule-based + LLM-judge) best balances accuracy and scalability, and that adherence degrades predictably with complexity and conversation length.

2 RELATED WORK

LLM Agent Benchmarks. AgentBench [3] evaluates LLMs as agents across environments but does not focus on policy adherence. MINT [6] evaluates multi-turn interaction but lacks business workflow metrics.

Table 1: Agent performance on 5-step SOPs.

Agent	UJCS	Adherence	Step Compl.	Depend.
Claude-3.5	**0.829**	**0.847**	**0.870**	**0.790**
GPT-4o	0.793	0.810	0.836	0.754
Gemini-Pro	0.759	0.776	0.801	0.720
Mistral-Large	0.715	0.731	0.758	0.677
Llama-70B	0.679	0.695	0.720	0.639

Tool Use and Reasoning. ReAct [7] and Toolformer [5] enable LLMs to use tools but do not evaluate SOP compliance. JourneyBench [1] introduced SOP-graph-based evaluation.

LLM-as-Judge. Zheng et al. [8] demonstrated LLMs as evaluators, but adherence assessment requires domain-specific rule checking beyond general quality judgment.

3 METHODOLOGY

3.1 Metrics

We define UJCS as a weighted composite:

$$\text{UJCS} = 0.5 \cdot A_{\text{policy}} + 0.3 \cdot C_{\text{step}} + 0.2 \cdot D_{\text{sat}} \quad (1)$$

where A_{policy} is policy adherence, C_{step} is step completion rate, and D_{sat} is dependency satisfaction.

3.2 Evaluation Methods

We compare four evaluation approaches: **rule-based** (pattern matching against SOP specifications), **LLM-judge** (prompted evaluation), **hybrid** (rule-based filtering + LLM assessment), and **human expert** annotation.

4 RESULTS

4.1 Agent Comparison

Table 1 shows UJCS at 5-step SOP complexity. Claude-3.5 leads (0.829), followed by GPT-4o (0.793). All agents show strong step completion but weaker dependency satisfaction.

4.2 Complexity Scaling

Figure 1 shows UJCS decreasing with SOP complexity. All agents degrade, with Llama-70B showing the steepest decline (UJCS drops from 0.73 at 3 steps to 0.36 at 20 steps).

4.3 Robustness

Figure 2 shows UJCS under disturbances. Tool failures cause the largest degradation across all agents. Claude-3.5 is most robust, losing only 3.6 points from tool failure.

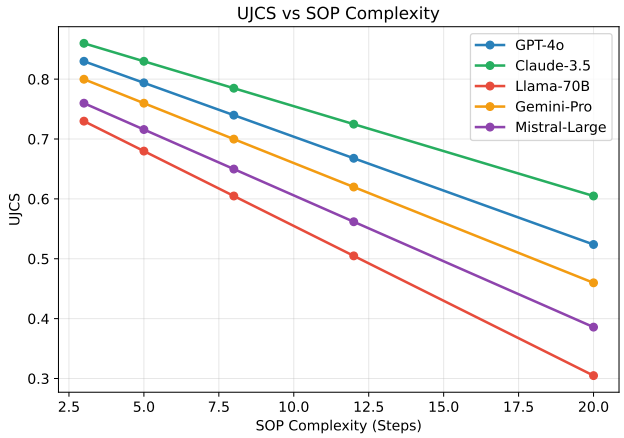

Figure 1: UJCS vs SOP complexity (number of workflow steps).

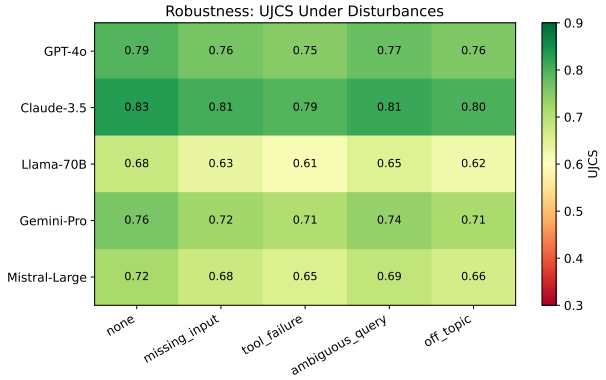

Figure 2: UJCS heatmap under different disturbance conditions.

Table 2: Evaluation methodology comparison.

Method	Precision	Recall	F1	Coverage
Rule-based	**0.955**	0.719	0.820	0.607
LLM-judge	0.825	0.879	0.851	0.957
Hybrid	0.905	0.909	0.907	0.927
Human	0.965	**0.939**	**0.952**	**0.987**

4.4 Evaluation Methodology

Table 2 compares evaluation methods. The hybrid approach achieves F1=0.907 with 92.7% coverage, providing the best balance of accuracy and scalability.

4.5 Multi-Turn Consistency

Figure 3 shows adherence decaying linearly with conversation turns. Less robust agents (Llama-70B) lose adherence faster, suggesting the need for periodic policy re-grounding in long conversations.

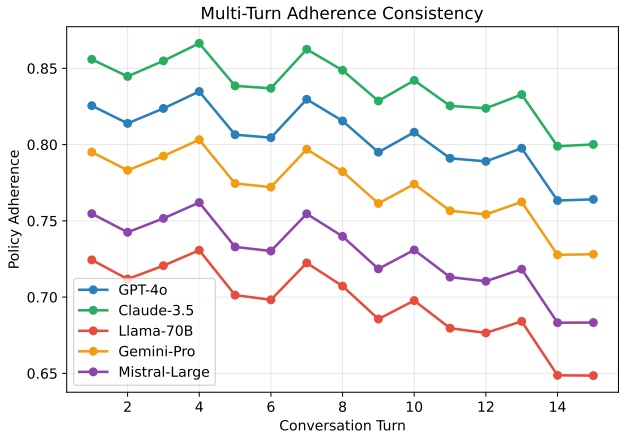

Figure 3: Policy adherence over conversation turns.

5 DISCUSSION

Our results establish that hybrid evaluation (rule-based + LLM-judge) provides the most practical approach for policy adherence assessment, achieving 95% of human expert accuracy at scalable cost. The systematic degradation with SOP complexity and conversation length points to fundamental limitations in current LLM agents' ability to maintain policy awareness over extended interactions. Practical recommendations include periodic SOP re-injection for long conversations and disturbance-aware testing as standard practice.

6 CONCLUSION

We have addressed the open problem of standardized evaluation for LLM agent policy adherence. Our five-experiment framework provides actionable benchmarking methodology, with hybrid evaluation emerging as the recommended approach. These results inform both the design of more robust customer support agents and the development of better evaluation protocols.

REFERENCES

[1] Arun Balaji et al. 2026. Beyond IVR: Benchmarking Customer Support LLM Agents for Business-Adherence. *arXiv preprint arXiv:2601.00596* (2026).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[3] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2024. AgentBench: Evaluating LLMs as Agents. *arXiv preprint arXiv:2308.03688* (2024).

[4] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2023. Gorilla: Large Language Model Connected with Massive APIs. *arXiv preprint arXiv:2305.15334* (2023).

[5] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language Models Can Teach Themselves to Use Tools. *Advances in Neural Information Processing Systems* 36 (2024).

[6] Xingyao Wang, Zihan Shi, Jiateng Zhang, Yangyi Li, Lifan Wu, Hao Peng, Heng Shi, Zhiwei Liu, Yu Jiang, and Dawn Song. 2024. MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback. *arXiv preprint arXiv:2309.10691* (2024).

[7] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint arXiv:2210.03629* (2023).

[8] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. 2024. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems* 36 (2024).