CLOUDINTENT-BENCH: INTENT MULTIPLICITY CHALLENGES LLM AGENTS IN REAL-WORLD CLOUD SERVICES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) agents are increasingly deployed in interactive scenarios, such as cloud services, where conversations often involve multiple user requests, challenging the underlying LLMs to reason over evolving dialogue histories. Although existing analyses often attribute lower performance to longer dialogue histories, interaction turn counts alone provide limited insight into which aspects of a conversation drive this degradation. We therefore investigate intent multiplicity, defined as the number of distinct user intents identified in a dialogue history by intent annotations. It provides an intermediate characterization between raw conversational context and agent performance, enabling us to analyze agent performance in real-world applications at a more fundamental level. Furthermore, we introduce CloudIntent-Bench, a benchmark of decision tasks derived from authentic cloud service tickets, to systematically examine the LLM-agent performance in different intent multiplicity scenarios. Experimental evaluations show that performance declines as intent multiplicity increases within a transition region, and controlled studies that remove or provide intent information offer complementary evidence for attributing this decline to intent multiplicity. Our work identifies intent multiplicity as a more fundamental perspective for analyzing agent performance and provides a practical benchmark for developing reliable LLM-agent applications in the cloud services domain.

1 INTRODUCTION

Large language model (LLM) agents are increasingly used in service applications, where they must interpret multi-turn conversations and decide how to respond to user requests. In cloud services, these decisions require agents to identify the current user intent, recover relevant information from earlier messages, and determine what is still needed to move forward. A central challenge is that dialogue histories vary not only in length but also in the user intents they contain: a long conversation may concern a single intent, whereas a shorter one may involve several distinct intents with different objects and constraints. This raises a central research question: how does agent performance vary with the number of distinct user intents expressed in the dialogue history?

Research on multi-turn interaction has documented reduced reliability when task information unfolds across a conversation (Laban et al., 2026). Recent studies further highlight difficulties in interpreting user intent and adapting to changing user goals (Liu et al., 2026; Tack et al., 2026). Dialogue length alone, however, provides limited insight into these difficulties. A long conversation may elaborate a single user intent, whereas an equally long conversation may involve several distinct intents with different objects and constraints. Figure 1 illustrates this distinction through two support histories in cloud services with the same message count but different numbers of user intents. This motivates examining the intents expressed in a dialogue history to better understand performance differences that message count alone cannot characterize.

We investigate *intent multiplicity*, denoted by U, as the number of distinct user intents in a visible dialogue history. Building on reviewed human annotations of intent labels, we train a language model to annotate complete source tickets, then compute U from the labels within each visible

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

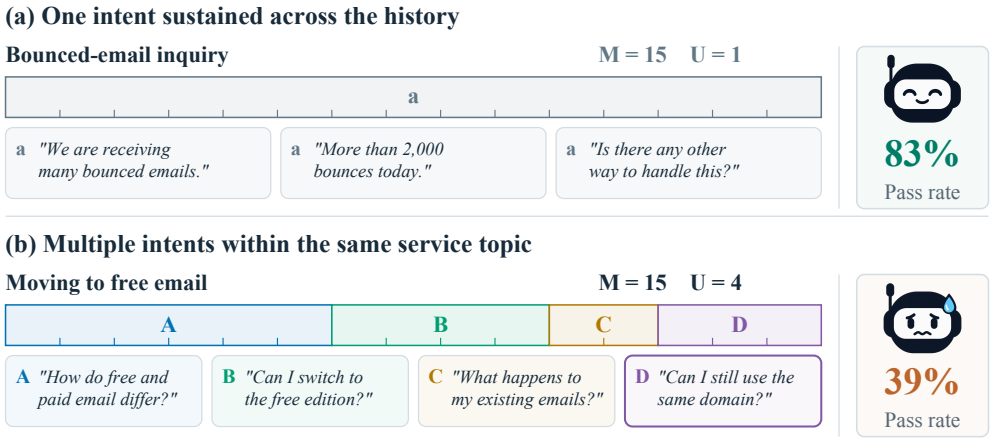


Figure 1: **Equal-length histories can contain different numbers of user intents.** Two authentic email-support visible dialogue histories contain $M = 15$ messages but (a) $U = 1$ and (b) $U = 4$ predicted intents. Letters denote distinct intent groups within each history. Excerpts are translated and shortened.

dialogue history. This measure complements message count M: equally long histories can contain different numbers of intents.

To study how agent performance varies with intent multiplicity and provide a platform for evaluating agents in cloud services, we introduce CloudIntent-Bench, comprising 1,973 decision tasks derived from authentic, anonymized support conversations in cloud services. Each task presents a dialogue history ending at an intermediate customer request, with subsequent messages and the final resolution hidden from the agent. The agent must identify the active intent, select a next action, and write a customer-facing reply based on the available evidence. These fixed checkpoints allow us to compare agent performance across histories with different intent counts while evaluating all agents on the same decisions and visible information.

Our experiments show that agent performance declines as intent multiplicity increases within a transition region. Our controlled studies further suggest that removing distracting historical intents or explicitly stating the current intent can improve agent performance. These findings support intent multiplicity as a complementary descriptor of task difficulty. We further apply our intent annotation method to two external dialogue collections and observe poorer average outcomes for multi-intent conversations than for single-intent conversations.

Our contributions are twofold. First, we develop and evaluate an intent annotation method for measuring intent multiplicity and examine its relationship with agent performance. This provides a semantic perspective on conversational difficulty and helps characterize the challenges agents face when handling multiple user intents. Second, we introduce CloudIntent-Bench, a benchmark of tasks derived from authentic support conversations in cloud services. It provides a common setting for studying agent performance across different levels of intent multiplicity, supporting the evaluation and development of reliable service agents.

2 Related Work

History use and multi-turn performance. Long-context and multi-turn benchmarks examine information position, cross-turn reasoning, and instruction following (Liu et al., 2024; Deshpande et al., 2025; Li et al., 2025), alongside sensitivity to intervening dialogue and earlier model replies (Kwan et al., 2024; Bai et al., 2024). Laban et al. (2026) observe lower reliability when task specifications unfold across simulated turns than when supplied at once; early assumptions and errors can persist through later exchanges. In a related fragmented-instruction setting, Liu et al. (2026) separate intent inference from execution and report gains from explicit intent reconstruction. Tack et al. (2026) find lower GPT-5.1 accuracy when added turns introduce intent transitions rather than

repeat information in a turn-matched GSM8K comparison. CloudIntent-Bench complements these studies by examining whether the number of distinct user intents in a dialogue history is associated with agent performance after accounting for message count.

Intent representation and multiplicity. MSDialog annotates utterance-level information-seeking intents (Qu et al., 2018), while conversational intent induction and human-aligned clustering recover intent categories (Gung et al., 2023; Hong et al., 2025). BlendX studies within-utterance multiplicity, and TIAGE examines cross-turn topic shifts (Yoon et al., 2024; Xie et al., 2021). These perspectives differ in their unit of analysis: utterance-level labels, multiple intents within an utterance, and topic changes across turns do not directly measure the number of distinct user intents in a dialogue history. CloudIntent-Bench uses dialogue-local semantic annotations to count distinct user intents within the visible history, counting repeated occurrences of the same intent only once.

Support dialogue datasets and service-agent evaluation. Ubuntu and JDDC provide technical-support and customer-service conversations (Lowe et al., 2015; Chen et al., 2020). For agent evaluation, API-Bank assesses API planning, retrieval, and invocation, while ToolSandbox adds stateful tool execution and interactive evaluation (Li et al., 2023; Lu et al., 2025). CRMArena studies professional customer-relationship-management tasks (Huang et al., 2025), while τ-bench and τ^2-bench study tool–agent–user interaction and dual-control environments (Yao et al., 2025; Barres et al., 2026). Other service benchmarks address unstructured knowledge (τ-Knowledge), service utility and compensation cost (FoodDeliveryService), and service quality and resolution efficiency on authentic support tickets in cloud services (CirrusBench) (Shi et al., 2026; Gao et al., 2026; Yu et al., 2026). Work on clarification (Saeidi et al., 2018; Aliannejadi et al., 2021), response faithfulness and attribution (Dziri et al., 2022a;b), and LLM-based evaluation (Liu et al., 2023) informs the assessment of information needs and evidence-supported replies. CloudIntent-Bench evaluates current-intent identification, next-action selection, and evidence-supported replies at fixed decision checkpoints in authentic support histories in cloud services.

3 Intent Multiplicity

User intents are not explicitly delineated in dialogue histories, as a single user intent may be expressed across multiple messages. We therefore describe an intent annotation approach that makes these intents explicit, providing an intermediate representation for analyzing how intent multiplicity relates to agent performance.

3.1 Intent Definition and Representation

We represent each distinct user intent within an observation window using a intent label. An intent concerns the goal being pursued, interpreted with the relevant object and contextual constraints. Messages expressing or addressing that intent share a group identity. Consecutive messages may form an intent chunk. Intent annotation is guided primarily by user messages. To preserve the completeness and coherence of each chunk, support-agent, assistant, and tool messages are assigned to the same chunk as the nearest relevant user messages.

The intent representation captures the user intents expressed in the observed dialogue, including those associated with requests that have already been resolved. The relevance of information associated with these intents depends on the current goal and context.

3.2 Intent Annotation and Multiplicity Measurement

The annotation pipeline draws on existing annotations from customer-support operations in cloud services, based on authentic ticket conversations between customers and support staff. Business annotators adjusted the boundaries of clustered text segments and assigned semantic intent labels using practical judgment. Chunk boundaries reflect changes in the user intent being addressed, while consecutive messages concerning the same intent are kept together. The annotated boundaries and labels undergo review for structural consistency. To apply this annotation approach to a larger collection of dialogues, we train a 32B-parameter language model on the double-reviewed human annotations to predict intent-chunk boundaries and semantic intent labels.

The annotator predicts chunk boundaries and a semantic intent label for each chunk. Assigning each message the label of its containing chunk produces a intent label sequence $\mathbf{z} = (z_1, \ldots, z_M)$, where M is the number of messages in the observation window. Appendix A details this representation and its conversion to message-level labels. We measure intent multiplicity as the number of distinct labels in this sequence:

$$U = |\{z_1, \ldots, z_M\}| \,. \tag{1}$$

For each support task in cloud services, we compute U from the labels within the dialogue history visible at the decision checkpoint, using annotations obtained from the complete source ticket.

3.3 Measurement Validation

The historical annotation corpus contains 15,985 dialogues, split by source ticket before model training into 11,197 training dialogues, 1,588 validation dialogues, and 3,200 held-out test dialogues. We evaluate the trained annotator along three dimensions: agreement with human intent annotations, accuracy of the resulting intent counts, and repeatability across inference runs. Bracketed values below denote 95% confidence intervals; Appendix A details the evaluation and bootstrap procedures.

The reference comparison uses 3,196 complete valid prediction–reference pairs. We assess intent structure through same-intent message-pair F1, partition agreement, and boundary recovery. Same-intent message-pair F1 is 0.870 [0.856, 0.884], and macro adjusted Rand index is 0.761 [0.747, 0.774]. Boundary F1 is 0.397 under exact matching and 0.521 with a one-message tolerance; the corresponding intervals are reported in Appendix A.

To evaluate intent multiplicity directly, we compare predicted and human-reference U on the same 3,196 dialogues. Exact-count agreement is 76.4% [74.9, 77.9], with a mean absolute error of 0.273 [0.255, 0.292] intents. These results quantify count accuracy on complete held-out dialogues; they do not directly establish accuracy on the dialogue histories visible at the benchmark's decision checkpoints. Appendix A provides the count-confusion matrix.

For repeatability, we apply the same frozen annotator twice to the held-out set and compare chunk counts, boundary sequences, and semantic-group sequences. Among the 3,193 dialogues with valid outputs in both runs, 94.5% have identical annotations across all evaluated fields. This comparison measures consistency across inference runs, separately from agreement with human references: repeated predictions can agree while sharing an error relative to the reference.

4 CloudIntent-Bench

4.1 Support Dialogues in Cloud Services and Dataset Profile

CloudIntent-Bench is built from anonymized conversations between customers and support staff in cloud services. Each task is constructed by selecting an intermediate customer request as a decision checkpoint and retaining the dialogue up to that point as the agent's input. Subsequent messages and the final resolution are hidden from the agent and used to inform the task's Golden reference. The benchmark comprises 1,973 tasks, with 1,330 multi-intent tasks complemented by 643 single-intent tasks. Visible histories contain a median of 16 messages, with an interquartile range of 10–25. The primary Golden actions comprise 805 direct replies, 257 clarification requests, 533 knowledge queries, and 378 tool calls. Figure 2 summarizes the task distribution across intent-multiplicity groups and the distribution of primary Golden actions within each group.

4.2 Task Formulation

Given a visible dialogue history, the agent identifies the active intent, selects one next action, and writes a customer-facing reply. It also reports the current active intent for diagnostic evaluation. The active intent states the current customer goal and its relevant constraints; the reply communicates an answer or the next step needed to advance that goal.

The four actions distinguish the information source needed next. `direct_reply` applies when the visible dialogue supports an answer. `ask_clarification` requests a necessary fact or choice from the customer. `knowledge_query` identifies missing product or procedural knowledge, while `tool_call` identifies missing private or dynamic operational state.

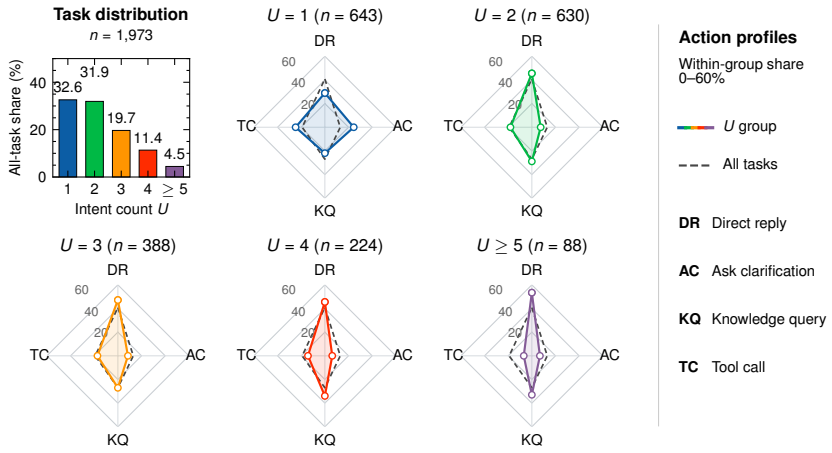


Figure 2: Dataset composition of CloudIntent-Bench. The bar chart shows the percentage of all 1,973 tasks in each intent-multiplicity group. Radar charts show the within-group distribution of primary Golden actions, with task counts above each panel. Colored outlines denote the displayed group and dashed gray outlines denote the all-task reference. All radars use the same 0–60% scale. $U \geq 5$ pools $U = 5$–9. DR: direct reply; AC: ask clarification; KQ: knowledge query; TC: tool call.

The evidence and missing-information fields make the decision auditable. Evidence references identify supporting messages in the visible history. A direct reply uses this evidence and leaves the missing-information list empty. An information-seeking action names a consequential missing item and indicates how it should be obtained. Its reply should explain that next step without assuming the result of a future query or operation.

4.3 Golden Package Construction and Quality Control

We construct a hidden Golden package for each checkpoint from the recorded service process. Model-assisted drafting organizes the current intent, the Golden action, available evidence, missing information, and reply requirements into a structured reference. Construction of the Golden package draws on the actions and replies of human support staff in the original tickets to define the reference action and reply requirements at each checkpoint. Reference responses allow alternative phrasings that satisfy the same requirements.

Quality control combines human and LLM review of the current intent, Golden action, missing information, and reply requirements. The review focuses on whether the requirements specified in the Golden package are justified by the dialogue visible at the checkpoint. A structural audit additionally confirms that visible-message references match the checkpoint boundaries recorded in the Golden package across all 1,973 tasks.

4.4 Evaluation Protocol

Scoring requirements. Action correctness requires the predicted action $\hat{a}_i$ to match the Golden action a_i^* for task i: $S_{\text{action},i} = \mathbb{I}[\hat{a}_i = a_i^*]$. Intent correctness concerns the current customer goal and its relevant constraints. Reply correctness requires addressing that goal with supported claims and an appropriate next step. For information-seeking actions, the reply should identify what is missing and how to obtain it.

Action-balanced pass rate. The four action categories have unequal frequencies in the benchmark. An overall pass rate therefore gives greater weight to common actions, potentially obscuring weaker performance on less frequent actions. We report an action-balanced pass rate by computing the pass rate within each primary Golden action category and averaging the four rates with equal weights. This measure gives each action category equal importance regardless of its frequency in the dataset.

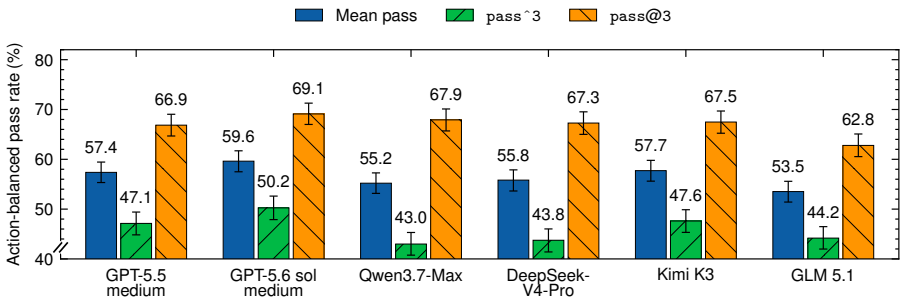


Figure 3: Action-balanced verifier performance on 1,973 tasks over three registered runs per configuration. Error bars denote pointwise 95% source-ticket bootstrap confidence intervals.

5 Experiments

We study the model performance, and how agent performance varies with intent multiplicity. Our main analysis covers six model configurations, each evaluated in three registered runs on the same 1,973 tasks. We use action-balanced `pass^3` as the primary metric, requiring a task to pass the verifier in all three runs. Mean pass summarizes average performance across runs, while `pass@3` records success in at least one run. Appendix B provides details of the execution setups.

For the overall and U-stratified pass rates in the main evaluation, we report pointwise 95% confidence intervals based on 5,000 source-ticket bootstrap resamples, keeping all observations from each ticket together. The paired-input and external studies use separate samples and outcome definitions, described in their respective subsections.

5.1 Overall Task Performance

Figure 3 summarizes action-balanced verifier performance across six configurations. The primary metric, `pass^3`, ranged from 42.99% to 50.25%. GPT-5.6 sol medium had the highest point estimates across all three metrics: 59.63% for mean pass, 50.25% for `pass^3`, and 69.11% for `pass@3`.

The gap between `pass@3` and `pass^3` was 18.61–24.93 percentage points, representing the action-balanced share of tasks that passed in only one or two runs. These results characterize variability across registered runs without evaluating feedback-guided recovery. Appendix B provides additional results and action-specific breakdowns.

5.2 Intent Multiplicity and Task Difficulty

Performance across intent counts. Figure 4 shows action-balanced `pass^3` across five intent-count strata. All six configurations show lower point estimates as U increases from 2 to 3 and from 3 to 4. Rates range from 46.91–52.60% at $U = 2$ to 19.86–33.62% at $U = 4$, indicating a shared decline within this transition region. Exploratory adjacent comparisons on exact $U = 1, \ldots, 5$ provide statistical support for parts of this pattern: after Holm correction over 24 comparisons, declines from $U = 2$ to $U = 3$ are significant for Qwen, DeepSeek, Kimi, and GLM, and those from $U = 3$ to $U = 4$ are significant for DeepSeek and GLM.

Interpretation and scope. All six configurations exhibit a shared decline in performance within the $U = 2$–4 transition region. Their trends diverge in the pooled $U \geq 5$ group: GPT-5.6 sol medium, Kimi K3, and GLM 5.1 have higher point estimates than at $U = 4$, whereas the other three configurations decline further. The small number of tasks in this tail makes the estimates more sensitive to individual task outcomes, so these fluctuations should be interpreted cautiously. Appendix C provides the complete contrasts, pooled-tail sensitivity analysis, and action-cell counts. Despite the variation in the tail, the shared decline across all six configurations within the transition region indicates that higher intent multiplicity is associated with poorer agent performance under the registered evaluation protocol.

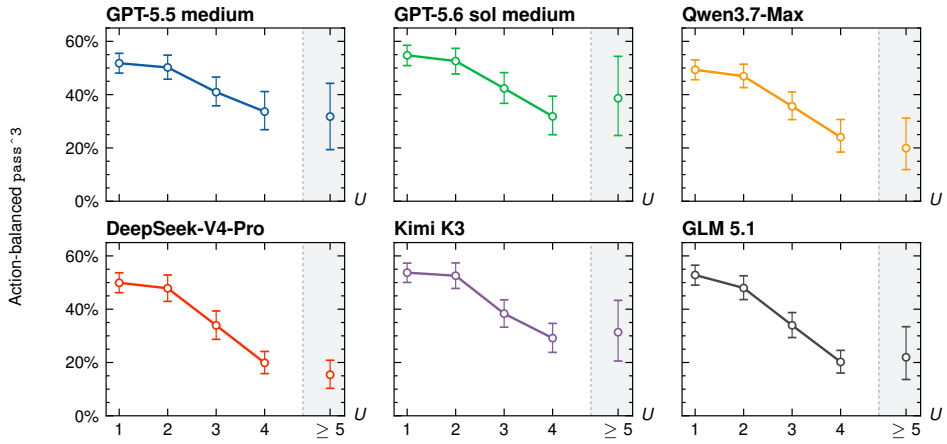


Figure 4: Action-balanced `pass^3` by intent multiplicity U across six configurations. Error bars: pointwise 95% source-ticket bootstrap CIs (5,000 resamples). Shading pools $U = 5$–9.

Association beyond message count. Even after accounting for message count M, higher intent multiplicity U remains negatively associated with agent performance, supporting its role as a complementary descriptor of task difficulty beyond dialogue length. This exploratory analysis uses equal task weights; Appendix B presents the model specifications and detailed results.

5.3 Controlled Study: History Filtering and Intent Guidance

To complement the association between intent multiplicity and task performance, we examine how agents respond to changes in historical context and explicit information about the current intent. This diagnostic study assesses whether filtering history or stating the current user intent changes intent matching and action selection. We compare three input conditions using GPT-5.5 medium on 100 diagnostic tasks from 100 distinct source tickets.

Full History provides all messages visible at the decision checkpoint. Focused History removes designated historical distractors while retaining evidence relevant to the current request, without rewriting or reordering the retained messages. Intent Hint preserves Full History and adds a description of the current user intent without specifying the action or answer. The three conditions share the same primary reference action and structured inline-ticket setup.

We report action correctness, the proportion of predictions matching the primary reference action, and Intent Match, a continuous Judge-based score assessing alignment between the agent's expressed intent and the reference intent and its constraints (Figure 5). Observed action correctness increases from 55.0% under Full History to 65.0% under both Focused History and Intent Hint. Mean Intent Match rises from 0.5545 to 0.6289 and 0.7589, respectively. Intent Hint thus achieves a higher intent-match score than Focused History while yielding the same aggregate action correctness.

On these 100 diagnostic tasks, removing historical intent distractors and providing current-intent hints are both associated with higher action correctness. These findings provide complementary evidence that intent-related information affects agent performance, suggesting that reducing historical intent interference and clarifying the current request can help agents understand the user's goal and select an appropriate next action.

Both input modifications yield higher action-correctness point estimates on these tasks, while explicit intent guidance produces the larger increase in Intent Match. Equal action correctness therefore coexists with different intent-match scores, showing why the two outcomes should be examined separately.

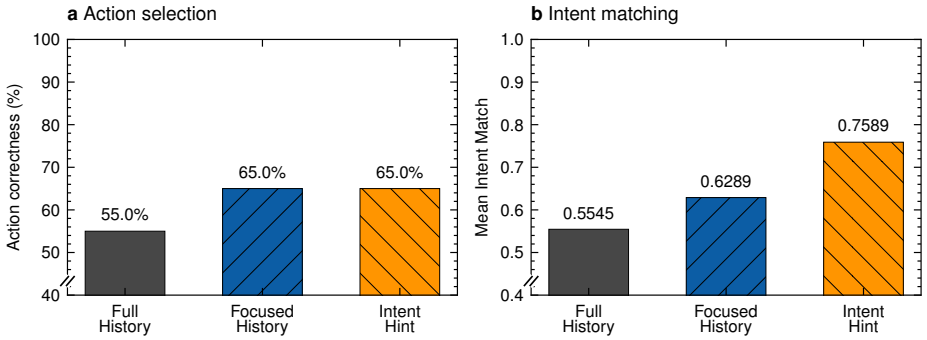


Figure 5: GPT-5.5 medium on 100 paired diagnostic tasks. Action correctness is the percentage of predictions matching the primary reference action; Intent Match is the mean Intent Judge score. Bars show point estimates describing this diagnostic set.

Table 1: Unadjusted outcomes for single-intent ($U = 1$) and multi-intent ($U > 1$) completed conversations: task success (%) for τ and mean simulator-generated satisfaction (1–5) for FDS. Overall rows pool conversations within each source.

Source / domain	Single-intent n	Multi-intent n	Single-intent outcome	Multi-intent outcome
τ: Overall	**212**	**163**	**79.72%**	**61.96%**
Airline	28	22	85.71%	81.82%
Retail	50	64	90.00%	79.69%
Telecom	91	23	98.90%	100.00%
Banking Knowledge	43	54	23.26%	16.67%
FDS: Overall	**393**	**119**	**3.160**	**2.168**
Easy	200	56	3.355	2.286
Hard	193	63	2.959	2.063

5.4 Cross-Source Analysis of Intent Multiplicity

We compare single-intent ($U = 1$) and multi-intent ($U > 1$) outcomes in two external collections: 375 completed τ trajectories across four domains, evaluated by source-defined binary task rewards (Barres et al., 2026; Shi et al., 2026), and 512 generated FoodDeliveryService (FDS) conversations from 363 source conversations, evaluated by simulator-generated satisfaction on a one-to-five scale (Gao et al., 2026). Single-intent conversations predominate in both collections (56.5% and 76.8%, respectively). Because FDS contains only four conversations with $U \geq 3$, we focus on the single-/multi-intent comparison; full distributions appear in Appendix D.

Multi-intent conversations have lower pooled task success in τ (61.96% vs. 79.72%) and lower mean simulated satisfaction in FDS (2.168 vs. 3.160). This direction holds in three of four τ domains and both FDS settings, with Telecom as the exception (Table 1). These unadjusted comparisons show that multiple intents occur beyond CloudIntent-Bench and accompany poorer average outcomes in several domains and settings.

6 Discussion and Limitations

Intent Multiplicity and Dialogue Difficulty. Intent multiplicity complements message count by characterizing distinct user intents in a dialogue history. Its granularity depends on the intent distinctions established during initial human annotation, which may vary across application settings and should be considered when comparing U. Within a given annotation scheme, the observed transition region suggests that U provides a useful perspective on task difficulty alongside history length. It motivates examining how agents distinguish and relate user intents to identify the current goal, although histories with the same U may present different decision challenges.

Context Management and Agent Decisions. On the 100 diagnostic tasks in Section 5.3, Focused History and Intent Hint both yield higher action-correctness point estimates than Full History. Intent Hint achieves higher Intent Match than Focused History despite equal action correctness. These findings motivate studying how context management can help agents identify the current goal, retain relevant evidence, and recognize missing information. One design direction is to link retained evidence to the user intents and constraints to which it applies, updating these associations as the dialogue evolves. This could help agents distinguish earlier requests from the current goal while preserving information that remains relevant. Further controlled studies could clarify the role of intent multiplicity by separating it from changes in history content and length or the addition of goal information.

Evaluating Agent Decisions. Evaluation should assess whether an agent's decision is appropriate given the information available at the checkpoint. When essential information is missing, a suitable response should identify the gap and explain how to address it through clarification, knowledge retrieval, or tool use. The diagnostic results also show that equal action correctness can coexist with different levels of intent alignment, motivating separate assessment of intent understanding and reply quality. Current verifier results cover action correctness and deterministic requirements. Complementing these checks with calibrated semantic assessment would provide a fuller account of whether agents understand the current goal and communicate an appropriate next step.

Scope and Future Validation. Intent counts are derived from model-predicted annotations of complete tickets. Annotation using only the visible dialogue history and sensitivity analyses of count errors would help assess their use at decision time. Sparse high-U strata and variation in task composition also constrain interpretation: action balancing adjusts category proportions but does not control message count or other task attributes. The observed transition region therefore remains specific to the evaluated tasks and protocol. External comparisons use completed interactions and different outcome measures, so their descriptive associations do not establish U as a pre-decision difficulty predictor. Larger samples in high-U strata and replication across application settings would help assess the stability of these findings.

7 Conclusion

We introduced intent multiplicity to characterize the distinct user intents represented in dialogue histories and developed CloudIntent-Bench, comprising 1,973 decision tasks from authentic support conversations in cloud services.

Across six agent configurations, action-balanced verifier `pass^3` point estimates decline as intent multiplicity increases within a transition region. A complementary pooled analysis shows that higher intent multiplicity remains associated with lower agent performance success after accounting for message count. On the diagnostic sample, history filtering and explicit intent guidance both yield higher action-correctness point estimates, with intent guidance producing the larger increase in Intent Match. These findings support intent multiplicity as a complementary descriptor of task difficulty under the evaluated protocol. CloudIntent-Bench provides a common setting for evaluating agent performance across different levels of intent multiplicity and supporting the development of reliable agents for cloud services.

AI Use Statement

In this work, we used generative AI tools for polishing the writing. We have not used generative AI tools for any tasks requiring disclosure under the ICLR AI-use policy. Additionally, for tasks with recommended disclosure, our use was limited to editing the manuscript to improve readability. We have reviewed all AI-assisted work. Specifically, the authors reviewed all AI-assisted revisions paragraph by paragraph to ensure that they preserved the intended meaning and did not introduce inaccurate facts, research claims, or citations. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics Statement

CloudIntent-Bench is derived from customer-support conversations in cloud services used with company authorization for research purposes. The materials were de-identified by replacing keys, real names, and other identifying information. The benchmark is intended for research evaluation, and generated responses should not be treated as operational advice for live customer accounts. Any public release of task artifacts remains subject to privacy review and applicable access permissions. Satisfaction scores reported for external simulators are generated by those simulators and do not constitute feedback from human participants.

Reproducibility Statement

Our main analysis uses a fixed cohort of 1,973 tasks, with each of six configurations evaluated in three registered runs. Appendix A describes intent-count construction and measurement validation, while Appendix B specifies outcome aggregation, uncertainty estimation, and conditional-association analyses. We retain task and source-ticket identifiers, registered-run mappings, per-task outputs, and input hashes to support traceability. The benchmark dataset and code have been provided to reviewers and will be made publicly available upon acceptance of the paper.

References

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4473–4484, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.367. URL https://aclanthology.org/2021.emnlp-main.367/.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-Bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.401. URL https://aclanthology.org/2024.acl-long.401/.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64377.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. The JDDC corpus: A large-scale multi-turn Chinese dialogue dataset for E-commerce customer service. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 459–466, Marseille, France, May 2020. European Language Resources Association. URL https://aclanthology.org/2020.lrec-1.58/.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.958. URL https://aclanthology.org/2025.findings-acl.958/.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490, 2022a. doi: 10.1162/tacl_a_00529. URL https://aclanthology.org/2022.tacl-1.84/.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*,

10:1066–1083, 2022b. doi: 10.1162/tacl_a_00506. URL https://aclanthology.org/2022.tacl-1.62/.

Ning Gao, Wei Zhang, Yuqin Dai, Ling Shi, Ziyin Wang, Yujie Wang, Wei He, Jinpeng Wang, and Chaozheng Wang. Reinforcing real-world service agents: Balancing utility and cost in task-oriented dialogue. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65380.

James Gung, Raphael Shu, Emily Moeng, Wesley Rose, Salvatore Romeo, Arshit Gupta, Yassine Benajiba, Saab Mansour, and Yi Zhang. Intent induction from conversations for task-oriented dialogue track at DSTC 11. In *Proceedings of the Eleventh Dialog System Technology Challenge*, pp. 242–259, Prague, Czech Republic, September 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.dstc-1.27/.

Mengze Hong, Wailing Ng, Chen Jason Zhang, Yuanfeng Song, and Di Jiang. Dial-In LLM: Human-aligned LLM-in-the-loop intent clustering for customer service dialogues. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5885–5900, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.300. URL https://aclanthology.org/2025.emnlp-main.300/.

Kung-Hsiang Huang, Akshara Prabhakar, Sidharth Dhawan, Yixin Mao, Huan Wang, Silvio Savarese, Caiming Xiong, Philippe Laban, and Chien-Sheng Wu. CRMArena: Understanding the capacity of LLM agents to perform professional CRM tasks in realistic environments. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3830–3850, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.194. URL https://aclanthology.org/2025.naacl-long.194/.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. MT-Eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20153–20177, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1124. URL https://aclanthology.org/2024.emnlp-main.1124/.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pp. 54738–54778, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/59f6421e64707225fdf5b28840679a07-Paper-Conference.pdf.

Jinnan Li, Jinzhe Li, Yue Wang, Yi Chang, and Yuan Wu. StructFlowBench: A structured flow benchmark for multi-turn instruction following. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9322–9341, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.486. URL https://aclanthology.org/2025.findings-acl.486/.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-Bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3102–3116, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.187. URL https://aclanthology.org/2023.emnlp-main.187/.

Geng Liu, Fei Zhu, Rong Feng, Changyi Ma, Shiqi Wang, and Gaofeng Meng. Intent mismatch causes LLMs to get lost in multi-turn conversation. arXiv preprint arXiv:2602.07338, 2026. URL https://arxiv.org/abs/2602.07338.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 285–294, Prague, Czech Republic, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4640. URL https://aclanthology.org/W15-4640/.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.65. URL https://aclanthology.org/2025.findings-naacl.65/.

Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. Analyzing and characterizing user intent in information-seeking conversations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pp. 989–992. ACM, June 2018. doi: 10.1145/3209978.3210124. URL https://doi.org/10.1145/3209978.3210124.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2087–2097, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1233. URL https://aclanthology.org/D18-1233/.

Quan Shi, Alexandra Zytek, Pedram Razavi, Karthik Narasimhan, and Victor Barres. τ-Knowledge: Evaluating conversational agents over unstructured knowledge. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63427.

Jihoon Tack, Philippe Laban, and Jennifer Neville. LLMs get lost in evolving user intent. arXiv preprint arXiv:2607.20734, 2026. URL https://arxiv.org/abs/2607.20734.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. TIAGE: A benchmark for topic-shift aware dialog modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1684–1690, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.145. URL https://aclanthology.org/2021.findings-emnlp.145/.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, volume 2025, pp. 9965–10017, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1b126cc38b8638e07bef37e7b2bb72bf-Paper-Conference.pdf.

Yejin Yoon, Jungyeon Lee, Kangsan Kim, Chanhee Park, and Taeuk Kim. BlendX: Complex multi-intent detection with blended patterns. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2428–2439, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.218/.

Yi Yu, Guangquan Hu, Chenghuang Shen, Xingyan Liu, Jing Gu, Hangyi Sun, Junzhuo Ma, Weiting Liu, Jianfeng Liu, Mingyue Pu, Yu Wang, Zhengdong Xiao, Rui Xie, Longjiu Luo, Qianrong Wang, Gurong Cui, Honglin Qiao, and Wenlian Lu. CirrusBench: Evaluating LLM-based agents beyond correctness in real-world cloud service environments. arXiv preprint arXiv:2603.28569, 2026. URL https://arxiv.org/abs/2603.28569.

APPENDIX

A INTENT ANNOTATION AND MEASUREMENT DETAILS

Count construction. We fine-tuned a 32-billion-parameter model on the reviewed human annotations and retained the fine-tuned weights to support reproduction of the annotation procedure. The annotator predicts chunk boundaries and semantic-group identities on complete source tickets. We expand these predictions to message-level labels, restrict the sequence to the dialogue history visible at each checkpoint, and count distinct group identities as defined in Section 3.2. Recurring groups are counted once.

Reference comparison. The 3,196-pair held-out measurement comparison excludes four test records without complete valid pairs. Reference labels are historical message-level Tag-4 annotations. Same-intent pairwise precision and recall compare which message pairs share a group, without requiring equality of arbitrary group identifiers. Macro adjusted Rand index summarizes partition agreement per ticket. Boundary F1 instead tests change locations. The 95% confidence intervals use ticket-level bootstrap resampling: 2,000 resamples for partition and boundary metrics (seed 20260726), and 5,000 for count metrics (seed 20260924).

Table 2: Held-out intent-partition agreement on 3,196 paired dialogues. These are measurement results, not agent intent-understanding scores.

Measure	Estimate	95% CI
Same-intent pairwise F1 (micro)	0.8701	[0.8559,0.8835]
Adjusted Rand index (macro)	0.7606	[0.7471,0.7741]
Aligned message accuracy (macro)	0.9033	[0.8977,0.9087]
Exact boundary F1	0.3967	[0.3740,0.4189]
Boundary F1, one-message tolerance	0.5209	[0.4993,0.5435]

A separate repeated-inference comparison has 3,193 jointly valid outputs and 94.49% exact agreement in chunk count, boundary sequence, and semantic-group sequence. This comparison retains its own denominator and measures repeatability. Repeated predictions can agree while sharing an error relative to the reference. None of these numbers is reported as exact U accuracy on the visible dialogue histories of the 1,973 tasks.

Direct count validation. The count comparison in Section 3.3 uses distinct historical human Tag-4 labels and distinct predicted semantic groups on the same 3,196 complete tickets. Mean signed error (prediction minus reference) is $+0.019$ [-0.002, 0.040]. The reference-$U \geq 5$ group has 45 tickets, with 23 undercounts and five overcounts. Figure 6 displays the distribution of count errors. These are neither fresh human annotations nor a validation using only the visible dialogue histories of the 1,973 decision tasks.

B TASK OUTCOMES AND STATISTICAL PROTOCOL

Main cohort and registered scores. The main evaluation fixes 1,973 task identities from 1,766 source tickets and six configurations with three registered runs each, giving 35,514 task–run scores. Recoveries replace records within their logical run by task identity; they never become additional trials. The registered results include selective output-format recovery and differing execution environments. Cross-configuration comparisons therefore include harness and recovery differences. Under the fixed paper-score policy, 482 remaining anomalous records receive zero: 480 already had raw zero rewards, and two had missing raw rewards. Raw missingness and anomaly provenance are retained separately from these paper scores.

Repeated outcomes and weighting. For each task and configuration, mean pass averages its three binary paper scores, `pass^3` is one only if all three scores equal one, and `pass@3` is one if at least one score equals one. Task-weighted rates average these task outcomes directly. Action-balanced

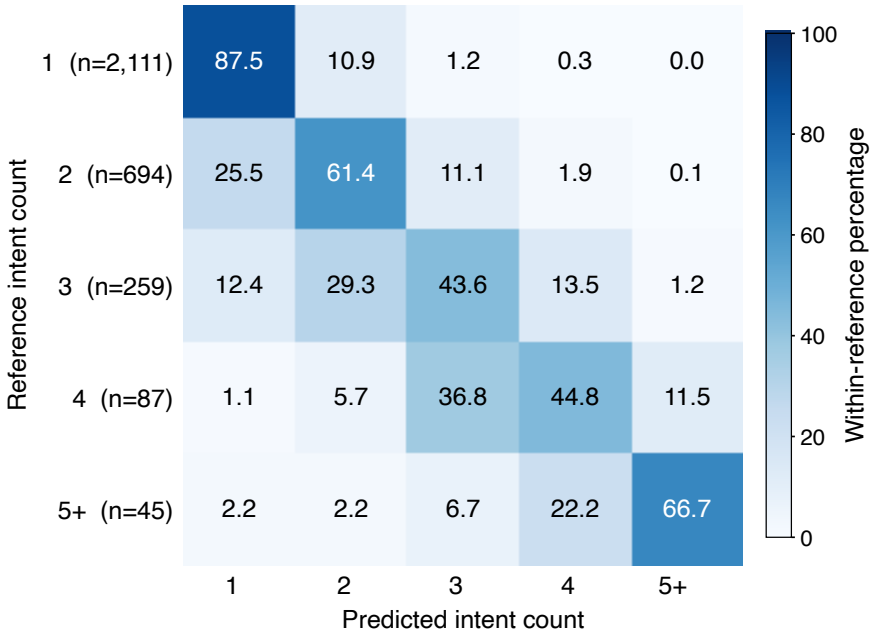


Figure 6: Row-normalized predicted versus reference intent counts on the 3,196 held-out complete tickets. Cells show percentages within each reference group; the final row and column pool counts of five or more. The bottom-right cell shows 66.7% agreement in this pooled category, whereas exact-count agreement among the same 45 tickets is 37.78%. Row sample sizes are shown beside the reference groups.

rates first average within each primary Golden action category and then give each of the four categories 25% weight. The full cohort contains 805 direct-reply, 257 clarification, 533 knowledge-query, and 378 tool-call tasks. Tables 3 and 4 retain the aggregate and action-specific three-run results.

Table 3: Task-weighted `pass^3` on the same 1,973 tasks and registered runs as Figure 3. Intervals are pointwise 95% source-ticket bootstrap intervals.

Configuration	All-three passes	`pass^3` (%) [95% CI]
GPT-5.5 medium	965	48.91 [46.76, 51.10]
GPT-5.6 sol medium	950	48.15 [46.00, 50.28]
Qwen3.7-Max	887	44.96 [42.78, 47.12]
DeepSeek-V4-Pro	895	45.36 [43.18, 47.52]
Kimi K3	997	50.53 [48.34, 52.65]
GLM 5.1	907	45.97 [43.87, 48.16]

Table 4: Action-specific `pass^3` (%) with pointwise 95% source-ticket bootstrap intervals. Equal averaging of these four rates gives the main metric.

Configuration	Direct reply	Clarification	Knowledge query	Tool call
GPT-5.5 medium	49.4 [46.0, 52.9]	28.4 [22.8, 33.9]	48.8 [44.6, 53.0]	61.9 [57.1, 66.8]
GPT-5.6 sol medium	38.8 [35.3, 42.0]	47.1 [40.8, 53.1]	52.7 [48.6, 57.1]	62.4 [57.5, 67.3]
Qwen3.7-Max	46.8 [43.4, 50.2]	26.8 [21.5, 32.4]	44.8 [40.6, 49.2]	53.4 [48.4, 58.4]
DeepSeek-V4-Pro	56.0 [52.4, 59.4]	51.4 [45.2, 57.5]	36.4 [32.5, 40.5]	31.2 [26.4, 35.8]
Kimi K3	51.7 [48.1, 55.2]	26.5 [21.1, 32.1]	57.0 [52.9, 61.3]	55.3 [50.4, 60.4]
GLM 5.1	41.6 [38.2, 45.0]	27.2 [21.9, 32.8]	60.8 [56.8, 64.9]	47.1 [41.9, 52.2]

Uncertainty for the main cohort. Rate intervals draw 1,766 source tickets with replacement 5,000 times (seed 20260922), preserving all tasks, models, and runs from each sampled ticket. All overall draws are valid. In the pooled $U \geq 5$ stratum, 32 draws empty an action category and cannot define a four-action average; its interval uses the remaining 4,968 draws without reweighting or filling the empty category. The intervals reflect task-sampling variability conditional on these runs.

Length-adjusted conditional models. Let Y_{icr} denote the registered binary verifier outcome for task i, configuration c, and run r, and let $Z_{ic} = \prod_{r=1}^{3} Y_{icr}$. The pooled three-run model is

$$\operatorname{logit} P(Z_{ic} = 1) = \alpha_c + \beta_U U_i + \beta_M \log_2 M_i.$$

It contains 11,838 task–configuration observations from 1,973 tasks and 1,766 source tickets. A complementary model uses all 35,514 individual run outcomes and replaces α_c with configuration–run intercepts α_{cr}. Both use equal task weights, rather than action-balanced weights. Standard errors use source-ticket cluster-robust covariance with a finite-sample correction; 95% intervals use a normal reference on the log-odds scale and are exponentiated to odds ratios. Within each pooled outcome analysis, Holm correction covers the two tests for U and $\log_2 M$ (Table 5). The individual-run analysis is a consistency check using the same registered outcomes, rather than an independent replication.

Table 5: Pooled conditional associations on the same 1,973 tasks and registered outcomes. The U odds ratio is per additional intent; the M odds ratio is per doubling. The three-run model has configuration intercepts (11,838 task–configuration outcomes); the individual-run model has configuration–run intercepts (35,514 outcomes). Intervals are pointwise 95% source-ticket cluster-robust intervals. Holm correction covers two tests separately for each outcome.

Outcome	U OR [95% CI]	Holm p	M OR [95% CI]	Holm p
`pass^3`	0.844 [0.789, 0.904]	2.23×10^{-6}	0.932 [0.858, 1.012]	0.0935
Individual run	0.853 [0.800, 0.909]	2.18×10^{-6}	0.931 [0.860, 1.007]	0.0738

Configuration-specific associations. Separate models estimate the same two coefficients for each configuration, retaining run intercepts for individual-run outcomes. Each outcome has a separate Holm family of twelve tests across six configurations and two coefficients (Table 6). Four configurations retain negative U associations after correction under both outcomes: Qwen, DeepSeek, Kimi, and GLM. GPT-5.5 has a pointwise interval below one but does not pass Holm correction; the same holds for GPT-5.6 under the three-run outcome. For individual-run outcomes, GPT-5.6 and DeepSeek retain negative message-count associations after correction, whereas none of the six three-run message-count tests passes its Holm correction. A pooled interval containing one therefore does not establish that message count is irrelevant.

In the individual-run model, configuration-by-U and configuration-by-$\log_2 M$ interactions yield cluster-robust Wald statistics of $\chi^2_5 = 24.878$ and 24.458, respectively; both have Holm-adjusted $p = 0.000294$ across the two tests. These are omnibus heterogeneity tests, not pairwise comparisons between configurations.

Functional forms and observation restrictions. Table 7 checks the individual-run analysis with raw $M/10$, a cubic spline of $\log_2 M$ with four degrees of freedom, categorical U, controls for Golden action, and three restrictions. The restrictions retain only $U \geq 2$ tasks, retain $M \leq 44$ (the cohort's 95th percentile), or exclude the 482 anomaly-flagged records. The first two retain 1,330 and 1,882 tasks; excluding anomalies retains 35,032 observations from all 1,973 tasks. Exclusion is a selection-sensitivity check, not a change to the main zero-score policy. The negative U estimates persist in these specifications, but the message-count estimate becomes more negative when controlling for Golden action. Action adjustment and action-balanced weighting are distinct operations; neither establishes causal identification.

Exact message-count strata and conditional shape. A supplementary linear model of each task's mean over eighteen runs includes an intercept for every observed value of M. Within this specification, the common slope per additional intent is -3.91 percentage points (95% CI $[-5.48, -2.33]$;

Table 6: Configuration-specific models including U and $\log_2 M$, with run intercepts for individual-run outcomes. All use the full 1,973-task cohort and source-ticket cluster-robust covariance. Each outcome has a separate Holm family of twelve tests (six configurations, two coefficients). Pointwise intervals can exclude one even when the corrected test does not reject.

Configuration	U OR [95% CI]	Holm p	M OR [95% CI]	Holm p
Three-run pass indicator				
GPT-5.5 medium	0.893 [0.822, 0.971]	0.0563	0.964 [0.868, 1.071]	1.0000
GPT-5.6 sol medium	0.895 [0.821, 0.975]	0.0629	0.862 [0.775, 0.960]	0.0527
Qwen3.7-Max	0.816 [0.749, 0.889]	3.33×10^{-5}	0.946 [0.850, 1.053]	1.0000
DeepSeek-V4-Pro	0.833 [0.763, 0.909]	0.0004	0.867 [0.778, 0.967]	0.0629
Kimi K3	0.860 [0.790, 0.937]	0.0049	0.980 [0.882, 1.090]	1.0000
GLM 5.1	0.769 [0.703, 0.841]	1.19×10^{-7}	0.978 [0.879, 1.089]	1.0000
Individual-run outcome				
GPT-5.5 medium	0.901 [0.833, 0.974]	0.0505	0.956 [0.868, 1.053]	1.0000
GPT-5.6 sol medium	0.931 [0.862, 1.007]	0.3621	0.839 [0.761, 0.926]	0.0036
Qwen3.7-Max	0.820 [0.762, 0.883]	1.55×10^{-6}	0.953 [0.867, 1.047]	1.0000
DeepSeek-V4-Pro	0.838 [0.777, 0.904]	4.52×10^{-5}	0.853 [0.774, 0.940]	0.0093
Kimi K3	0.863 [0.800, 0.932]	0.0014	0.993 [0.900, 1.096]	1.0000
GLM 5.1	0.772 [0.713, 0.835]	1.92×10^{-9}	1.005 [0.910, 1.109]	1.0000

Table 7: Individual-run specification checks, with configuration–run intercepts and source-ticket clustered uncertainty. The base controls are U and $\log_2 M$; each row replaces a functional form, adds Golden action, or restricts observations as specified in the text. The spline row tests U while controlling a cubic spline of $\log_2 M$. Holm correction covers the twelve coefficients in this table.

Specification	Change	OR [95% CI]	Holm p
Raw message count	$U + 1$	0.844 [0.793, 0.898]	8.71×10^{-7}
Raw message count	$M + 10$	0.963 [0.908, 1.022]	0.3487
Adjustment for Golden action	$U + 1$	0.847 [0.794, 0.904]	5.13×10^{-6}
Adjustment for Golden action	$M \times 2$	0.854 [0.786, 0.927]	0.0011
Spline message count	$U + 1$	0.856 [0.802, 0.913]	1.89×10^{-5}
Categorical intent count	$M \times 2$	0.929 [0.858, 1.006]	0.3487
$U \geq 2$ only	$U + 1$	0.810 [0.743, 0.884]	1.89×10^{-5}
$U \geq 2$ only	$M \times 2$	0.900 [0.803, 1.009]	0.3487
$M \leq 44$ only	$U + 1$	0.848 [0.794, 0.906]	9.65×10^{-6}
$M \leq 44$ only	$M \times 2$	0.929 [0.854, 1.010]	0.3487
Exclude anomaly records	$U + 1$	0.868 [0.814, 0.926]	0.0001
Exclude anomaly records	$M \times 2$	0.936 [0.864, 1.013]	0.3487

Holm $p = 2.41 \times 10^{-6}$). The reciprocal model uses an intercept for each observed U and estimates -1.75 points per doubling of M ($[-3.64, 0.14]$; $p = 0.0699$). Both use source-ticket cluster-robust t inference with 1,765 degrees of freedom, and Holm correction covers the two tests. These are descriptive mean slopes, not bounded single-task probability predictions.

To examine shape without imposing a linear U slope, we fit the individual-run logit model with configuration–run intercepts, categorical U, and $M/10$. Figure 7 fixes $M = 20$ and averages the eighteen run-specific probabilities with equal weights. All $U = 1$–9 tasks enter the fit; the plot shows exact $U = 1$–5. Intervals for conditional means use the delta method with the clustered covariance. At $M = 20$, the changes for $U = 2 \to 3$ and $3 \to 4$ are -7.15 points ($[-11.95, -2.36]$; Holm $p = 0.0138$) and -8.26 points ($[-14.35, -2.16]$; $p = 0.0238$), respectively. The $1 \to 2$ and $4 \to 5$ comparisons are inconclusive (both adjusted $p = 1$). This four-test family concerns individual-run conditional predictions; it is distinct from the 24 action-balanced three-run contrasts in Appendix C. Neither analysis supports a strictly monotone relationship across the full U range.

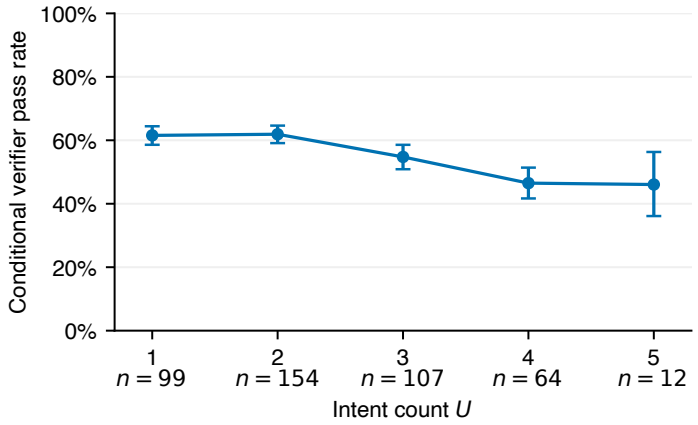


Figure 7: Conditional individual-run verifier pass rate at $M = 20$, averaging predictions over eighteen registered runs on 1,973 tasks. Bars are pointwise 95% source-ticket cluster-robust delta-method intervals for the expected rate, not single-task prediction intervals. Labels give task counts within $M = 20 \pm 3$. The final point is exact $U = 5$ (55 tasks in the full cohort), not the pooled $U \geq 5$ stratum. These are conditional predictions rather than action-balanced `pass^3` rates.

C Complete Intent-Count Results

Tail support and exact counts. The main figure retains every task and pools $U = 5$–9 into one 88-task group. Table 8 shows its sparse action cells. Table 9 reports every exact count using task weighting, since some high-U strata lack one or more action categories and do not support a four-action average. These rates use a different weighting from the main figure.

Table 8: Task counts in each cell defined by predicted U and primary Golden action.

U	Direct reply	Clarification	Knowledge query	Tool call
1	186	157	141	159
2	287	47	182	114
3	183	33	105	67
4	102	14	76	32
≥ 5	47	6	29	6

Adjacent action-balanced comparisons. The adjacent contrasts use action-balanced rates and therefore cannot be obtained by subtracting the task-weighted rates in Table 9. The exact-U analysis uses the 1,940 tasks at $U = 1, \ldots, 5$, from 1,752 source tickets. It fits saturated U-group-by-action cell means and contrasts the four action means with equal weights. CR1 source-ticket covariance and a two-sided t reference with 1,751 degrees of freedom give the pointwise intervals in Table 10.

Table 9: Task-weighted pass^3 (%) at every exact predicted intent count. GPT-5.5 and GPT-5.6 denote the medium configurations in Figure 3. Single-task tail cells are descriptive observations.

U	Tasks	GPT-5.5	GPT-5.6	Qwen	DeepSeek	Kimi	GLM
1	643	51.79	54.28	49.46	50.54	53.50	52.41
2	630	51.11	49.05	50.16	50.32	55.87	51.75
3	388	46.91	43.81	40.72	41.24	45.10	38.66
4	224	42.41	38.39	31.25	30.80	40.18	29.91
5	55	38.18	41.82	29.09	27.27	36.36	29.09
6	24	41.67	50.00	29.17	29.17	54.17	37.50
7	7	28.57	14.29	28.57	28.57	42.86	28.57
8	1	0.00	0.00	0.00	0.00	0.00	0.00
9	1	0.00	0.00	0.00	0.00	0.00	0.00

Holm correction covers all 24 adjacent tests. These are exploratory comparisons performed after observing the curves. The 33 tasks with $U \geq 6$ remain in the main figure and full cohort. Exact $U = 5$ has 3 clarification, 27 direct-reply, 20 knowledge-query, and 5 tool-call tasks, limiting tail inference. Using the pooled $U \geq 5$ group on all 1,973 tasks yields the same rejection pattern after Holm correction within its separate 24-test family: four declines for $2 \to 3$, two for $3 \to 4$, and none for the remaining contrasts, although tail effect estimates can differ in magnitude and direction. Neither analysis controls message count.

Table 10: Adjacent differences in action-balanced pass^3 for exact $U = 1, \ldots, 5$. Differences are higher minus lower U, in percentage points; intervals are pointwise 95% CR1 intervals and are not adjusted for multiple comparisons. Significance is assessed using p-values Holm-adjusted over all 24 rows.

Configuration	Contrast	Difference	95% CI	Holm p
GPT-5.5 medium	2-1	-1.59	[-7.47, +4.30]	1.0000
GPT-5.5 medium	3-2	-9.26	[-16.38, -2.15]	0.1724
GPT-5.5 medium	4-3	-7.32	[-16.21, +1.58]	1.0000
GPT-5.5 medium	5-4	-4.09	[-18.53, +10.36]	1.0000
GPT-5.6 sol medium	2-1	-2.15	[-8.23, +3.93]	1.0000
GPT-5.6 sol medium	3-2	-10.27	[-17.78, -2.76]	0.1254
GPT-5.6 sol medium	4-3	-10.50	[-19.66, -1.34]	0.3464
GPT-5.6 sol medium	5-4	+8.26	[-11.68, +28.19]	1.0000
Qwen3.7-Max	2-1	-2.42	[-8.28, +3.44]	1.0000
Qwen3.7-Max	3-2	-11.29	[-18.28, -4.30]	0.0298
Qwen3.7-Max	4-3	-11.58	[-19.55, -3.61]	0.0799
Qwen3.7-Max	5-4	-3.53	[-15.94, +8.87]	1.0000
DeepSeek-V4-Pro	2-1	-2.06	[-8.20, +4.07]	1.0000
DeepSeek-V4-Pro	3-2	-13.92	[-21.15, -6.70]	0.0032
DeepSeek-V4-Pro	4-3	-14.06	[-20.81, -7.31]	0.0010
DeepSeek-V4-Pro	5-4	-4.35	[-12.11, +3.41]	1.0000
Kimi K3	2-1	-1.14	[-7.05, +4.78]	1.0000
Kimi K3	3-2	-14.23	[-21.26, -7.20]	0.0016
Kimi K3	4-3	-9.19	[-16.66, -1.72]	0.2384
Kimi K3	5-4	-3.96	[-16.32, +8.39]	1.0000
GLM 5.1	2-1	-4.88	[-10.65, +0.89]	1.0000
GLM 5.1	3-2	-14.00	[-20.58, -7.43]	0.0007
GLM 5.1	4-3	-13.74	[-20.20, -7.28]	0.0007
GLM 5.1	5-4	+0.95	[-11.07, +12.96]	1.0000

D External Sources, Models, and Linkage

Samples and outcomes. The archived τ collection has Airline (50), Retail (114), Telecom (114), and Banking Knowledge (97) trajectories, each from one local DeepSeek-V4-Pro run. Success means source reward equal to one; 270 of 375 trajectories succeed. FDS contains 256 Easy and 256 Hard generated test conversations with intent predictions, corresponding to 235 and 234 source

conversations, respectively. The settings share 106 source conversations, giving 363 unique sources overall. Satisfaction is generated by the simulator on a one-to-five scale, not by a human evaluator.

External intent annotation. A fine-tuned 32-billion-parameter model annotates the intent structure of the completed dialogue trajectories, predicting intent-chunk boundaries and semantic-group identities. We compute U by counting distinct semantic groups, with recurring groups counted once, following Section 3.2. M counts messages in the same completed trajectory. These measurements describe the observed interaction as a whole, whereas CloudIntent-Bench measures U within the visible dialogue history of a fixed decision task.

Intent-count distributions. Table 11 reports the complete distributions of predicted U in the archived trajectories, including the domain and setting breakdowns. The distributions vary across domains and sources: single-intent conversations account for 56.5% of τ and 76.8% of FDS. The former contains 63 trajectories with $U \geq 3$, while the latter contains only four and none with $U \geq 4$. These differences motivate the single-/multi-intent comparison in Section 5.4; the external samples do not provide comparable support for testing the $U = 2$–4 transition observed in CloudIntent-Bench. Domain, conversation-generation setting, and observation window differ across these collections, so the distributional differences cannot be attributed to domain alone.

Table 11: Predicted intent-count distributions in completed external conversations. Entries are conversation counts; overall rows sum the corresponding domain or setting rows. No observed count exceeds $U = 5$ in τ or $U = 3$ in FDS.

Source / domain	$U = 1$	$U = 2$	$U = 3$	$U = 4$	$U = 5$	Total
τ: Overall	**212**	**100**	**37**	**24**	**2**	**375**
Airline	28	14	8	0	0	50
Retail	50	43	15	6	0	114
Telecom	91	14	2	6	1	114
Banking Knowledge	43	29	12	12	1	97
FDS: Overall	**393**	**115**	**4**	**0**	**0**	**512**
Easy	200	54	2	0	0	256
Hard	193	61	2	0	0	256

Single- and multi-intent outcomes. Multi-intent conversations ($U > 1$) have poorer aggregate outcomes than single-intent conversations in both external collections (Table 1). In τ, success rates are 61.96% versus 79.72%; in FDS, mean simulated satisfaction is 2.168 versus 3.160. The same direction holds in both FDS settings and in three of the four τ domains, with Telecom as the exception. Thus, despite differences in intent-count distributions, both collections show a descriptive association between multiple intents and poorer average outcomes. These unadjusted comparisons do not establish a common effect size, a causal effect, or a uniform relationship within every domain.

Adjusted models and sensitivity. The τ binary logit includes U, $\ln M$, and domain fixed effects, with HC1 robust standard errors. The FDS ordered logit includes U, $\ln M$, setting, core need, user profile, and FAQ configuration; its corrected sandwich covariance clusters by source conversation across both settings. Both length terms use the natural logarithm. The two primary U tests use Holm correction. Their pointwise intervals do not include uncertainty from new model runs or intent reannotation. The adjusted U odds ratios are 0.696 [0.511, 0.949] for τ (Holm-adjusted $p = 0.044$) and 0.621 [0.365, 1.058] for FDS (Holm-adjusted $p = 0.080$).

In τ, replacing continuous U with a single-/multi-intent indicator gives OR=0.663 [0.341, 1.288]. Removing Banking Knowledge gives OR=0.747 [0.484, 1.154], and individual-domain single-/multi-intent contrasts do not survive Holm correction. Telecom has only one failed trajectory. These checks limit interpreting the pooled coefficient as a uniform cross-domain pattern. For FDS, the model without message count but with design controls gives OR=0.505 [0.304, 0.838]; adding $\ln M$ gives the wider interval reported above. Linear-message and spline sensitivity intervals also include one.

Unadjusted FDS associations. Spearman correlation between U and simulated satisfaction is -0.296 $[-0.366, -0.225]$ overall, -0.315 in Easy, and -0.278 in Hard. Split estimates are descriptive. Message count has a stronger pooled correlation of -0.626 $[-0.674, -0.567]$. The pooled 95% percentile intervals use 5,000 source-conversation bootstrap resamples (seed 2026091601), retaining all generated conversations from each sampled source across both settings. These associations and the adjusted odds ratios measure different quantities.

Analysis version and archive linkage. All external results reported here use one archived reanalysis snapshot, with input-file hashes and dialogue-level analysis records retained for verification. The τ predictions and outcomes align by domain and dialogue identity. FDS prediction-file hashes match the archived analysis, and source IDs, satisfaction, message counts, design covariates, and recomputed U agree. The FDS dialogue files used in the reanalysis differ in hash from those recorded in the original annotation analysis, however, and the prediction archive lacks the complete input-text and prompt/checkpoint hashes needed to establish text identity. The FDS results therefore describe the archived annotations and outcomes; they do not constitute a new end-to-end annotation run on the dialogue files used in the reanalysis.