KoChatBench: A Korean Multi-turn Chatbot Benchmark for Conversational Capability Evaluation

Anonymous ACL submission

Abstract

We introduce KoChatBench, a capability-based benchmark for evaluating Korean generative multi-turn dialogue. Existing evaluations often rely on single-turn or domain-specific tasks, limiting their ability to diagnose interaction-level failures. KoChatBench defines four core capabilities and constructs 600 sessions spanning 3–6 turns. We evaluate six commercial LLMs using an LLM-as-a-judge framework with GPT-5-mini, adopting session-level minimum aggregation to capture critical failures. Results show that Gemma-4-31B-IT achieves the strongest overall performance, while Nemotron-3-Super-120B-A12B exhibits weaknesses in conversational robustness. These findings highlight the importance of capability-level analysis and provide a structured framework for assessing stability in multi-turn interactions.

1 Introduction

Large Language Models (LLMs) are increasingly expected to support real user conversations that extend beyond single-turn question answering. In practice, interactions often span three or more turns, requiring models to retain prior information, resolve anaphora and ellipsis, incorporate cumulative constraints, revise responses, and handle topic shifts. However, existing evaluations of Korean LLMs are still largely based on single-turn settings. Even when multi-turn evaluation is considered, benchmarks are often limited to short interactions or domain-oriented tasks, making it difficult to distinguish whether failures stem from insufficient domain knowledge or limitations in multi-turn interaction capabilities.

This issue is particularly important for Korean dialogue, where subject omission, demonstratives, case particles, honorifics, and style shifts are tightly coupled with conversational context. As a result, simply translating English benchmarks is insufficient to capture realistic failure patterns in Korean multi-turn interactions.

To address this gap, we propose KoChatBench, a capability-based benchmark for evaluating Korean generative multi-turn dialogue. The benchmark is built on four core capabilities: (1) context and reference management, (2) instruction constraint maintenance and response execution, (3) revision and recovery, and (4) dialogue recovery and robustness. Each capability is further decomposed into mutually exclusive sub-tasks, enabling fine-grained diagnosis of interaction-level failures.

Figure 1 illustrates the benchmark construction pipeline. We define a turn as a query–response pair and a session as a sequence of turns, and construct 600 Korean sessions spanning 3–6 turns with an evaluation rubric. We further evaluate six commercial LLMs using an LLM-as-a-judge framework with `gpt-5-mini`, adopting session-level minimum aggregation to capture critical failures. Results show that Gemma-4-31B-IT achieves the strongest overall performance, while Nemotron-3-Super-120B-A12B exhibits weaknesses in conversational robustness. These findings demonstrate that KoChatBench provides a structured framework for diagnosing stability and capability-level limitations in Korean multi-turn LLMs.

2 Related Works

Dialogue evaluation is challenging because conversational quality depends on system goals, context, and human preference, while human evaluation is costly to scale (Deriu et al., 2021). Prior work shows that good conversations require more than locally plausible responses: they must be specific, related to prior turns, and coherent over interaction (See et al., 2019; Adiwardana et al., 2020). Open-domain chatbots further require persistent control of knowledge, persona, empathy, and long-term memory across turns (Roller et al., 2021; Xu

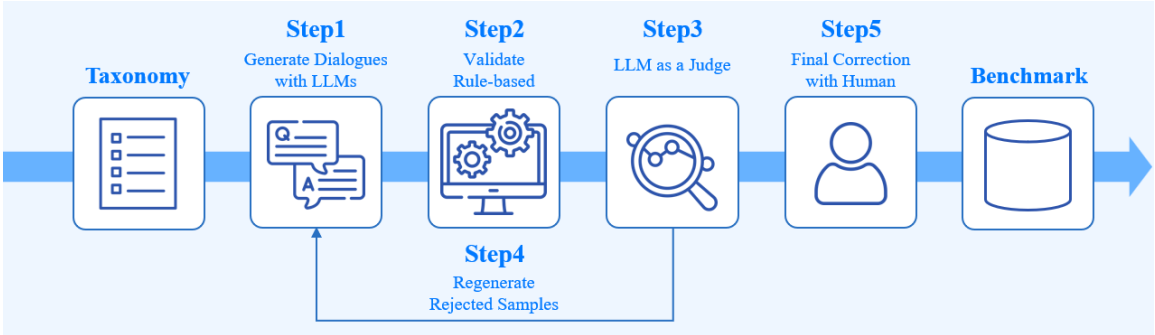

Figure 1: A Pipeline for Benchmark Dataset Generation

Benchmark	#Dialogues	#Turns	#Tasks
MT-Bench-101	1388	4208	13
MT-Bench++	80	640	8
MT-Eval	168	1170	4
MultiChallenge	273	1108	4
StructFlowBench	155	643	6
MMMT-IF	71	990	6
TOD-ProcBench	1004	6953	3
KoChatBench (Ours)	**600**	**2518**	**12**

Table 1: Comparison of Multi-turn Dialogue Benchmarks.

et al., 2022). These findings motivate benchmarks that evaluate multi-turn interaction itself, rather than isolated response quality.

Recent benchmarks operationalize this direction from different perspectives. MT-Bench-101, MT-Bench++, and MT-Eval evaluate context memory, anaphora, ellipsis, topic transition, recollection, and refinement (Bai et al., 2024; Sun et al., 2024; Kwan et al., 2024). MultiChallenge and StructFlowBench focus on long-term instruction retention, self-coherence, dialogue flow, and state transitions (Sirdeshmukh et al., 2025; Li et al., 2025), while MMMT-IF and TOD-ProcBench study instruction retrieval and constraint satisfaction in multimodal and task-oriented settings (Epstein et al., 2024; Ghazarian et al., 2025). Building on these studies, KoChatBench reorganizes prior tasks into four Korean-specific capabilities: context and reference management, instruction constraint maintenance and response execution, revision and recovery, and dialogue recovery and robustness. This design enables fine-grained diagnosis of Korean multi-turn failures involving ellipsis, demonstratives, particles, honorifics, and style shifts. Table 1 compares KoChatBench with prior benchmarks.

3 Benchmark Setup

We construct KoChatBench through a five-stage pipeline for dataset generation and quality control. As shown in Figure 1, the pipeline consists of: (1) LLM-based dataset generation, (2) rule-based validation, (3) LLM-as-a-judge filtering, (4) regeneration of rejected samples, and (5) human verification.

Taxonomy-guided generation. KoChatBench is based on the capability taxonomy in Table 2, which includes four high-level capabilities: context and reference management, instruction constraint maintenance and response execution, revision and recovery, and dialogue recovery and robustness. Each capability is divided into mutually exclusive subtasks and used as a generation target. For each sample, the prompt specifies the target capability, sub-task, number of turns, scenario, and expected interaction pattern. We define a turn as a user query–model response pair and a session as a sequence of turns.

Generation and validation. We use GPT-5.5 and Claude Sonnet 4.6 to generate candidate Korean multi-turn sessions. Rule-based validation filters structurally invalid samples, such as incorrect turn counts, missing roles, malformed outputs, or task-structure mismatches. The same models are then used as LLM judges to assess consistency, naturalness, task alignment, and evaluability. Rejected samples are regenerated, and accepted samples are finally verified by human annotators.

Dataset composition and evaluation. KoChatBench consists of 600 Korean multi-turn dialogue sessions spanning three to six turns. Table 3 shows the turn-count distribution. We evaluate commercial LLMs through the OpenRouter API, enabling comparison across model providers under the same benchmark instances and evaluation criteria. The

Capability ID	Capability Name	Task ID	Task Name	Evaluation Focus
C1	Context & Reference Management	C1-1	Explicit Information Recall	Recall and reuse explicitly stated information from previous turns
		C1-2	Distributed Information Integration	Integrate scattered information across turns via multi-hop reasoning
		C1-3	Referential & Ellipsis Resolution	Resolve pronouns, references, and omitted entities from context
		C1-4	State Change Tracking	Track structural and state changes throughout the dialogue
C2	Instruction & Constraint Maintenance	C2-1	Single Instruction Maintenance	Maintain predefined response rules and output formats across turns
		C2-2	Accumulated Constraints	Satisfy multiple constraints incrementally added over turns
		C2-3	Conflict Resolution	Resolve conflicting constraints using rule priority or recency
C3	Refinement & Recovery	C3-1	Iterative Revision and Preservation	Refine responses iteratively while preserving unaffected content and prior versions
		C3-2	Self-Correction	Detect and correct the assistant's own previous errors
C4	Dialogue Recovery & Robustness	C4-1	Error Propagation Blocking	Prevent propagation of incorrect premises and detect inconsistent restatements
		C4-2	Ambiguity Resolution	Ask clarification questions when instructions are ambiguous
		C4-3	Topic Interruption Recovery	Resume the original task after topic interruptions

Table 2: Mapping Between Capabilities and Tasks

Number of Turns	Number of Sessions
3 turns	10
4 turns	502
5 turns	48
6 turns	40
Total	600

Table 3: Distribution of KoChatBench sessions by number of turns.

evaluation supports both overall performance comparison and capability-level analysis of Korean multi-turn conversational behavior.

4 Experiments

In this section, we evaluate the Korean multi-turn conversational capabilities of several commercial LLMs using KoChatBench. We consider six models: Gemma-4-31B-IT, Mistral Medium 3.5, GPT-OSS-120B, GPT-OSS-20B, Nemotron-3-Nano-30B-A3B, and Nemotron-3-Super-120B-A12B. Each model generates responses for the same KoChatBench sessions, and the generated responses are evaluated using an LLM-as-a-judge approach. We use GPT-5-mini as the judge model and assign a score from 0 to 100 to each turn-level response.

4.1 Evaluation Protocol

Since KoChatBench consists of multi-turn dialogue sessions, evaluating only the average performance across turns may not sufficiently capture the stability of a model in realistic conversations. For example, even if a model performs well in most turns, it may still fail critically in a specific turn by forgetting previous context, violating accumulated constraints, or incorrectly resolving references. In such cases, a simple average score can obscure severe turn-level failures and overestimate the model's practical multi-turn reliability.

To address this issue, we use the lowest turn-level score within each session as the final score of that session. Formally, given a session i consisting of T_i turns, where $s_{i,t}$ denotes the evaluation score of turn t, the session score S_i is defined as follows:

$$S_i = \min_{t \in \{1,\ldots,T_i\}} s_{i,t} \quad (1)$$

This session-level minimum aggregation penalizes a model when it exhibits a severe failure at any point in a multi-turn interaction. This evaluation protocol follows a similar motivation to MT-Bench-101 (Bai et al., 2024), where the lowest turn score is used to reflect the weakest point of a dialogue session.

4.2 Overall Task-Level Performance

We first compute the session-level minimum score for each task and then average these scores for each model to measure overall performance on KoChatBench. Table 4 reports the task-level performance and overall average score of each model.

In terms of the overall average score, Gemma-4-31B-IT achieves the best performance, indicating that it handles multi-turn interactions relatively

Table 4: Model Performance by Task using Session-level Minimum Scores. (0–100 scale)

Task ID	Task Name	Gemma-4 31B it	Mistral Medium 3.5	GPT-OSS 120B	GPT-OSS 20B	Nemotron Nano	Nemotron Super
C1-1	Explicit Information Recall	92.6	97.6	96.4	93.2	94.4	95.2
C1-2	Distributed Information Integration	100.0	100.0	92.0	91.6	96.8	97.6
C1-3	Referential & Ellipsis Resolution	95.6	95.4	93.8	67.0	83.0	76.4
C1-4	State Change Tracking	92.4	89.2	77.4	76.6	65.6	40.6
C2-1	Single Instruction Maintenance	94.2	91.6	86.0	85.2	90.4	73.0
C2-2	Accumulated Constraints	70.0	71.0	59.8	55.4	66.2	55.0
C2-3	Conflict Resolution	95.0	95.6	96.2	96.0	99.8	99.0
C3-1	Iterative Revision and Preservation	100.0	98.8	98.8	93.8	98.6	77.4
C3-2	Self-Correction	87.2	81.6	75.6	75.6	83.4	87.6
C4-1	Error Propagation Blocking	72.8	60.8	93.6	87.8	52.6	71.6
C4-2	Ambiguity Resolution	94.6	94.4	95.4	85.2	81.6	35.8
C4-3	Topic Interruption Recovery	98.0	93.2	92.8	92.8	89.4	49.8
Overall Session-Minimum Average		91.0	89.1	88.2	83.3	83.5	71.6

robustly across most capabilities and tasks. In particular, Gemma-4-31B-IT maintains high scores across multiple tasks, suggesting that it is less prone to abrupt failures in individual turns. In contrast, Nemotron-3-Super-120B-A12B shows the lowest overall performance. Its performance is particularly weak in C4, Conversational Recovery & Robustness, suggesting limitations in preventing error propagation, resolving ambiguity, and resuming the original task after topic interruptions.

Interestingly, the strongest model in terms of overall performance does not outperform all other models on every task. In C1-1, Explicit Information Recall, Gemma-4-31B-IT records the lowest score among the evaluated models. This result shows that even a model with strong aggregate performance can exhibit task-specific weaknesses in particular multi-turn capabilities. Therefore, KoChatBench enables a more fine-grained analysis of model failures beyond a single overall score.

Additional analyses of the gap between mean and minimum session-level scores and the performance degradation from the first to the last turn are provided in Appendix A.

5 Discussion

KoChatBench provides a capability-level analysis of Korean multi-turn dialogue beyond aggregate scores. Results show that strong overall performance does not ensure consistent task-level behavior: Gemma-4-31B-IT performs best overall but is weaker on C1-1, whereas Nemotron-3-Super-120B-A12B shows notable weakness in C4. The mean–minimum gap further reveals turn-level instability, since a single failure can degrade the whole session. We also observe lower scores in later turns, indicating difficulty in maintaining context, constraints, and revisions. Future work will compare LLM-as-a-judge results with human evaluation and expand the benchmark to more diverse scenarios.

6 Conclusion

This work presents KoChatBench, a capability-based benchmark for evaluating Korean generative multi-turn dialogue. We define four core capabilities: context and reference management, instruction constraint maintenance and response execution, revision and recovery, and dialogue recovery and robustness, each divided into mutually exclusive sub-tasks. Based on this taxonomy, we construct 600 Korean sessions spanning 3–6 turns and evaluate multiple commercial LLMs using an LLM-as-a-judge framework. Experimental results show that aggregate scores can hide task-specific weaknesses and turn-level failures. By analyzing minimum session scores and first-to-last turn degradation, KoChatBench provides a fine-grained framework for diagnosing the stability and robustness of Korean multi-turn LLMs.

7 Limitations

KoChatBench has several limitations. First, although LLM-as-a-judge enables scalable evaluation, it may inherit biases from the judge model. Second, the benchmark contains 600 sessions spanning 3–6 turns, which may not fully cover longer or highly domain-specific conversations. Finally, although samples are human-verified, generated dialogues may still differ from real user interactions in diversity and naturalness.

References

Daniel Adiwardana and 1 others. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. 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)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Elliot L. Epstein, Kaisheng Yao, Jing Li, Xinyi Bai, and Hamid Palangi. 2024. MMMT-IF: A challenging multimodal multi-turn instruction following benchmark. *arXiv preprint arXiv:2409.18216*.

Sarik Ghazarian, Abhinav Gullapalli, Swair Shah, Anurag Beniwal, Nanyun Peng, Narayanan Sadagopan, and Zhou Yu. 2025. TOD-ProcBench: Benchmarking complex instruction-following in task-oriented dialogues. *arXiv preprint arXiv:2511.15976*.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. 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*, pages 20153–20177, Miami, Florida, USA. Association for Computational Linguistics.

Jinnan Li, Jinzhe Li, Yue Wang, Yi Chang, and Yuan Wu. 2025. StructFlowBench: A structured flow benchmark for multi-turn instruction following. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9322–9341, Vienna, Austria. Association for Computational Linguistics.

Stephen Roller and 1 others. 2021. Recipes for building an open-domain chatbot. In *EACL*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *NAACL*.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. 2025. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. *arXiv preprint arXiv:2501.17399*.

Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Parrot: Enhancing multi-turn instruction following for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9729–9750, Bangkok, Thailand. Association for Computational Linguistics.

Jing Xu and 1 others. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *ACL*.

A Appendix

This section provides additional analyses that complement the main experimental results. We examine two aspects of multi-turn stability: the gap between mean and minimum session-level aggregation, and performance degradation from the first to the last turn within each session.

A.1 Mean vs. Minimum Session-Level Aggregation

We compare the turn-level mean score and the minimum score within each session. The mean score reflects the model's overall response quality across turns, whereas the minimum score captures the weakest turn in the session. Therefore, a larger gap between the two scores indicates that the model may perform well in some turns but fail substantially in at least one turn. This gap reflects instability in multi-turn interaction.

Table 5 presents the difference between session-level mean and session-level minimum scores for each model. Gemma-4-31B-IT shows the smallest gap between the mean and minimum scores, while also achieving the highest final performance. This suggests that the model not only produces high-quality responses on average but also maintains relatively consistent performance throughout a session. In contrast, Nemotron-3-Super-120B-A12B shows the lowest overall performance and a substantial drop under minimum-score aggregation. This indicates that the model is more vulnerable to turn-level failures and exhibits lower stability across multi-turn sessions.

A.2 Performance Degradation across Turns

We further analyze how model performance changes as the number of turns increases. In multi-turn dialogue, later turns require the model to jointly process more previous context, accumulated constraints, revision history, and topic transitions. Therefore, performance degradation is more likely

Table 5: Within-Session Score Gap between Mean and Minimum Turns

Model	Mean	Min	Gap
Gemma-4 31B it	96.1	91.0	5.1
Mistral Medium 3.5	95.3	89.1	6.2
GPT-OSS 120B	93.9	88.2	5.7
GPT-OSS 20B	92.3	83.3	9.0
Nemotron Nano 30B A3B	94.0	83.5	10.5
Nemotron Super 120B A12B	89.0	71.6	17.4

Table 6: Within-Session Score Gap between First and Last Turns

Model	First Turn	Last Turn	Gap
Gemma-4 31B it	97.4	94.4	3.0
Mistral Medium 3.5	97.3	92.5	4.8
GPT-OSS 120B	95.2	93.1	2.1
GPT-OSS 20B	92.9	90.8	2.1
Nemotron Nano 30B A3B	95.5	91.2	4.3
Nemotron Super 120B A12B	91.8	86.7	5.2

to occur in later stages of a session. The score difference between the first and last turns serves as an important indicator of how reliably a model maintains multi-turn context as the dialogue progresses.

Table 6 shows the change in performance from the first turn to the last turn within each session. Overall, most models tend to receive lower scores in later turns. This suggests that models face increasing difficulty in maintaining previous information and accumulated constraints as the conversation becomes longer.

Among the evaluated models, Gemma-4-31B-IT shows the smallest performance gap between the first and last turns, indicating that it maintains relatively stable response quality even in longer conversations. Other models exhibit larger drops in later turns, suggesting that failures in specific capabilities may accumulate as multi-turn interactions progress. These results demonstrate that KoChat-Bench is useful not only for measuring overall task performance but also for analyzing performance stability and robustness with respect to dialogue length.

A.3 Sample Case.

As shown in Figures 2 and 3, the successful case demonstrates that the model correctly integrates information distributed across previous dialogue turns to generate an appropriate response, whereas the failure case shows that the model produces an incorrect response by ignoring relevant context from earlier turns.

Figure 2: Successful case for the C1-2 Distributed Information Integration task.

Figure 3: Failure case for the C1-2 Distributed Information Integration task.