TelBench: A Benchmark for Evaluating Telco-Specific Large Language Models

Sunwoo Lee, Dhammiko Arya, Seung-Mo Cho, Gyoung-eun Han,
Seokyoung Hong, Wonbeom Jang, Seojin Lee, Sohee Park,
Sereimony Sek, Injee Song, Sungbin Yoon, Eric Davis

SK Telecom, South Korea

{sunwoo.lois, dhammikoarya, seungmo, gyoungeun.han, seokyoungh, wonbeom.jang, skt.kaylee, sparkling, mony, injee.song, sungbinyoon, eric.davis}@sk.com

Abstract

The telecommunications industry, characterized by its vast customer base and complex service offerings, necessitates a high level of domain expertise and proficiency in customer service center operations. Consequently, there is a growing demand for Large Language Models (LLMs) to augment the capabilities of customer service representatives. This paper introduces a methodology for developing a specialized Telecommunications LLM (Telco LLM) designed to enhance the efficiency of customer service agents and promote consistency in service quality across representatives. We present the construction process of TelBench, a novel dataset created for performance evaluation of customer service expertise in the telecommunications domain. We also evaluate various LLMs and demonstrate the ability to benchmark both proprietary and open-source LLMs on predefined telecommunications-related tasks, thereby establishing metrics that define telcommunications performance.

1 Introduction

Recent advancements in Large Language Models (LLMs) have significantly enhanced natural language understanding and generation capabilities, leading to increased development of domain-specific LLMs across various sectors, including law, finance, and science. These specialized models aim to leverage LLMs' general linguistic proficiency while incorporating domain-specific knowledge (Colombo et al., 2024; Yang et al., 2023; Zhang et al., 2024).

The telecommunications (telco) industry is characterized by its large subscriber base, complex network infrastructure, diverse service offerings, and 24-hour global connectivity. This complexity results in a wide variety of customer inquiries, requiring extensive training for customer service representatives.

To enhance customer satisfaction in call center interactions, we leveraged LLMs to augment the expertise of customer service representatives and reduce response times. Our approach improved service efficiency by enabling LLMs to perform post-interaction tasks that previously required manual searching, reasoning, and documentation.

This paper's primary contributions are:

- **TelTask Dataset**: Evaluates telco service terminology and language proficiency for customer service applications. We detail the identification of key telco tasks and the methodology for dataset construction.
- **TelInstruct Dataset**: Assesses LLM agentic abilities in retrieving and utilizing database information, as well as deeper telecommunications knowledge. We propose essential skills for a Telco LLM Agent.
- **Telco LLM Evaluation**: We evaluate proprietary and open-source LLMs using our telco-specific benchmarks and existing general LLM capability tests, demonstrating the importance of domain-specific datasets.

This paper is structured as follows: Section 2 reviews related research, Section 3 details the dataset composition and development methodology, Section 4 presents LLM evaluation results on the dataset, and Section 5 summarizes findings and proposes future research directions.

2 Related work

Evaluating Large Language Models (LLMs) for domain-specific knowledge and task performance is crucial when considering their deployment as a service. To address this, domain-specific datasets have emerged across various fields, including medicine(Guha et al., 2023; Pal et al., 2022; Antaki et al., 2023), science(Zhang et al., 2024), law(Guha

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 609–626
November 12-16, 2024

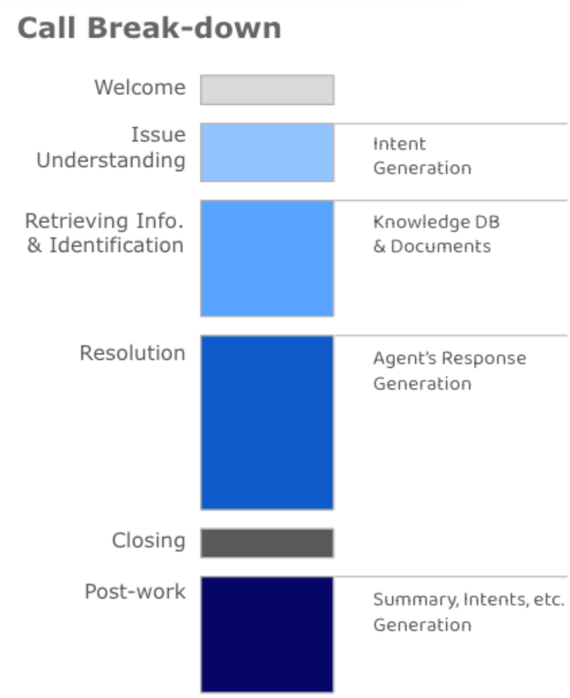

Figure 1: Breakdown of call center work

et al., 2023), finance(Son et al., 2023), education(Demszky and Hill, 2023), and coding(Chen et al., 2021; Liu et al., 2023a). Recently, this trend also has extended to telecommunications network infrastructure (Maatouk et al., 2023; Zou et al., 2024). All of these tailored datasets highlight the growing importance of adapting LLMs to specialized domains.

With the rapid advancement of LLMs' instruction-following capabilities, novel evaluation datasets have been developed to assess agentic behavior(Zhou et al., 2023) and responses to toxic language(Li et al., 2024). Evaluation approaches now integrate conventional automatic evaluation frameworks(Liang et al., 2023) with LLMs-based, reference-free approaches(Zheng et al., 2023; Liu et al., 2023b), enhancing the effectiveness and diversity of evaluating LLMs.

3 Dataset Construction and Development Methodology

Customer interactions with contact center agents involve issue comprehension, information retrieval, problem resolution, and post-call activities (Figure 1). These interactions often span multiple encounters with different agents, emphasizing the importance of the post-work phase. Large Language Models (LLMs) can enhance this phase by improving accuracy, reducing time spent, and standardizing practices, ultimately reducing human agent workload and improving service levels.

To support telco customer service with LLMs, we defined essential tasks and categorized them into two groups: TelTask and TelInstruct. TelTask is a comprehensive and involves contextual language understanding capability in conversations for post-interaction tasks, while TelInstruct is a benchmark set that assesses telco domain knowledge and instruction following capability. Sample data is available in Appendix A.

Two common pre-processing stages are applied to each dataset, followed by additional dataset-specific processing:

- **Heuristic Data Cleaning**: Rule-based methods and internal models eliminate excessively long or short dialogues and remove filler words. We sample and modify successful consultation logs, using stratified sampling to maintain data balance across consultation topics and types.
- **Anonymization**: To prevent the model from learning sensitive personal information, we replace Personally Identifiable Information (PII) with pseudonyms, maintaining data quality and coherence.

Detailed explanations of each task and associated development methodologies are provided in subsequent sections.

3.1 TelTask

We constructed the TelTask benchmark dataset using a balanced mixture of clean and slightly noisy data to reflect real-world usage in the telco industry. The dataset comprises between 100 and 963 instances per task category that are manually reviewed and validated by human annotators.

3.1.1 Sentiment

This task assesses customer sentiment in customer service dialogues. The goal is for an LLM to automatically classify sentiments into positive, negative, and neutral categories. The dataset facilitates precise prediction of customer emotions by capturing nuanced understanding of conventional expressions and context-specific phrases in telco interactions. For instance, the phrase "Thank you" appearing at the end of a conversation should be interpreted as a customary closing remark rather than an indication of satisfaction with a specific service.

Task	Volume	Reviewed[a]	Validated[b]
Sentiment	500	✓	✓
Entity	500	✓	
Intent	500	✓	
To-do	100	✓	✓
Topic	500	✓	✓
Summary	500	✓	✓
Safety	963	✓	

[a] Data reviewed and modified by human annotators to improve quality
[b] (Mainly) Lightly inspected automatically generated data

Table 1: TelTask Data Statistics

3.1.2 Entity

The entity benchmark evaluates recognition of telco-specific nomenclature, including product names, rate plans, and domain-specific proper nouns. To properly benchmark the complexities of entities in the telco domain, the test set incorporates various forms of each entity, including synonyms, and considers entity prevalence in utterances reflective of actual user speech patterns.

The evaluation process considers the prevalence of entities in utterances and incorporates cases reflective of actual user speech patterns. For instance, to evaluate rate plan name recognition (MOBILE_NAME entity), we included both current service plan names in their representative forms '5GX Regular' ("Is my current plan the 5GX Regular tariff?") and informal forms commonly used in actual customer utterances 'Regular' ("Is Regular any good?"). Our end goal was to evaluate entity recognition performance across diverse linguistic manifestations.

3.1.3 Intent

The intent dataset categorizes customer utterances into four primary types: Ask, Check, Cancellation, and Apply, mapped to specific service details. It includes both canonical examples and variations resembling real-world customer interactions to comprehensively evaluate the model's classification accuracy. For instance, the Check.RoamingPlan intent category includes well-formed, representative utterances such as, "I would like to subscribe to an international roaming plan," as well as more colloquial, abbreviated forms like "Sign up for roaming." This approach allows for a more comprehensive evaluation of the model's ability to accurately classify both canonical and real-world customer inputs.

3.1.4 Topic

The topic task extracts concise, noun-based key themes from customer service dialogues, specific to telco services. For instance, in a dialogue about roaming services for travel to Thailand, the ideal topics would include the specific tariff name, such as "Baro 3GB Plan," rather than generic terms like 'Thailand' or 'Travel'. The benchmark balances dialogues across various telco domains and was developed by using an LLM to generate representative topics and then having human annotators review and modify the outputs.

3.1.5 Summary

This task summarizes customer service dialogues, incorporating specific telco terminology. These dialogues are often length and contain domain-specific terminology and phrases, making it difficult for base LLMs to produce good summaries. The resulting summaries are also intended to be "action focused", so customer service agents can quickly glean key information about the call. As such, the benchmark set evaluates key metrics, including specificity, fluency, factuality, completeness, conciseness, and inclusion of key content.

3.1.6 To-do

This task generates follow-up actions for customer service representatives after conversations with customers. Common types of to-do items include sending multimedia messages (MMS) to convey additional information, making calls to obtain consent from account holders, conducting further research before responding, and transferring tasks to relevant departments. The benchmark includes consultations both requiring and not requiring follow-up actions, enabling accurate distinction between the two scenarios.

3.1.7 Safety

The safety data includes potentially unsafe situations in customer service interactions, addressing Korean language and cultural context-specific concerns. The benchmark comprises balanced sensitive expressions extracted from actual consultations, aiming to evaluate a model's ability to detect unsafe situations.

3.2 TelInstruct

The TelInstruct benchmark set comprises tasks containing 100 to 2,300 instances each. It evaluates a range of skills, from basic telco domain knowledge to complex scenarios requiring consideration of conversational context and relevant documents.

Task	Volume	Reviewed	Validated
Workflow	400[a]	✓	✓
TelcoQ&A	2,300[b]	✓	✓
MRC	120[c]	✓	

[a] 130 allocated for evaluation purposes
[b] 1,500 entries focused on customer service scenarios and 800 entries dedicated to infrastructure management
[c] equally distributed between simpleQA and Word-to-Text (60/60)

Table 2: TelInstruct Data Statistics

3.2.1 Workflow

The Workflow task evaluates an LLM's ability to respond appropriately to customer inquiries. It assesses the model's comprehension of telco consultation dialogue flows and its capacity to generate useful responses based on a knowledge database. The dataset comprises multi-turn dialogue data, telco knowledge documents, and generated responses.

Response quality is assessed on a 5-point Likert scale, considering relevance, specificity, factual accuracy, and fluency. The dataset closely resembles real-world telco consultation scenarios to ensure benchmark validity.

3.2.2 Telco Q&A

The Telco Q&A benchmark evaluates the LLM's understanding of customer service center operations and infrastructure management. It consists of open-ended questions simulating customer inquiries and infrastructure operator queries. Both sections contain concise, domain-specific questions and answers.

The evaluation process selected high-scoring question-answer pairs based on utility, factual accuracy, and user satisfaction. The Infrastructure Q&A set was developed using an LLM to generate questions and answers from infrastructure documentation, focusing on operation commands and troubleshooting procedures.

The customer service component comprises open-ended questions and answers that simulate typical customer inquiries. Concurrently, the infrastructure management section contains questions and answers that an infrastructure operator might encounter in their daily operations. Both topics contain concise, domain-specific questions and their corresponding answers.

3.2.3 MRC

The Machine Reading Comprehension (MRC) task is based on telco product and policy documents and instruction manuals. It includes two formats: SimpleQA and Word-to-Text. SimpleQA consists of concise questions answerable with a noun or noun phrase, designed to elicit answers from various text locations. The Word-to-Text task, inspired by (Cheng et al., 2024), involves generating sentences containing domain-specific terminology. Both formats follow a structure of reference document, question, and answer.

4 Evaluation of LLMs

This section presents the evaluation methodology and results for various Large Language Models (LLMs) to validate the utility of the TelBench benchmark set.

4.1 Evaluation Methodology

The evaluation of LLMs employs two primary methodologies: automatic evaluation and LLM-as-a-judge evaluation.

4.1.1 Automatic Evaluation

Automated assessment forms the initial phase of evaluation. While imperfect, this cost-effective method is crucial for facilitating a feedback loop of model tuning, evaluation, dataset improvement, and re-tuning. Task-specific metrics are selected based on the characteristics of each task.

For extensive response generation tasks like summarization, the ROUGE score is employed. Classification tasks utilize accuracy for balanced class frequencies and the F1 score for imbalanced cases. Topic-related tasks, which require detection of all positive instances, use hit rate (recall) as the primary metric.

4.1.2 LLM-as-a-Judge Evaluation

Domain-specific benchmarks like TelBench typically require costly human evaluation by domain experts. On the other hand, LLM-based evaluations,

Task	Spearman Correlation	Cohen's Kappa
Summary	0.72	0.35
Topic	0.84	0.31

Table 3: Spearman correlation coefficient and Cohen's Kappa coefficient between Human Evaluation and LLM-as-a-Judge methodologies

		Proprietary LMs			Open-Sourced LMs		
		GPT-4-Turbo	Claude 3.5 Sonnet	Claude 3 Haiku	Llama-3.1-405B-Instruct-FP8	Mistral-Large-Instruct	Mistral-Small-Instruct
Sentiment	F1-Score	0.744	0.860	0.772	0.870	0.886	0.714
Entity	F1(Micro)	0.303	0.368	0.451	0.258	0.258	0.181
Intent	Accuracy	0.632	0.663	0.606	0.570	0.659	0.234
Topic	Recall	0.228	0.252	0.254	0.278	0.261	0.131
Summary	ROUGE-L	0.441	0.443	0.424	0.437	0.410	0.369
To-Do	ROUGE-L	0.671	0.654	0.167	0.650	0.714	0.715
Safety (Harmless)	F1-Score	0.598	0.652	0.649	0.336	0.630	0.398
Safety (Privacy)	F1-Score	0.875	0.894	0.996	0.872	0.940	0.959
Telco Q&A (AICC)	ROUGE-L	0.353	0.330	0.345	0.422	0.293	0.314
Telco Q&A (Infra)	Accuracy	0.774	0.776	0.482	0.788	0.732	0.410
MRC (SQA)	ROUGE-L	0.618	0.662	0.353	0.574	0.691	0.588
MRC (WTT)	Accuracy	0.455	0.557	0.471	0.559	0.546	0.473

Table 4: Automatic evaluation results

if correlated with human assessments, enable more frequent performance assessments at reduced costs. This can then facilitate dataset refinement and more frequent model tuning.

TelTask Experiments were conducted to validate the LLM-as-a-Judge approach for two TelBench tasks: topic identification and summary generation. The experimental design included both human evaluation and LLM-as-a-Judge assessment for each task, with correlation between human ratings and LLM-as-a-Judge ratings serving as a validity measure.

Human evaluation involved assessing 100 sessions on a 5-point scale per task, using a two-way evaluation method to ensure inter-rater reliability. The LLM-as-a-Judge experiment utilized the GPT-4-turbo model to evaluate the same 100 sessions using the prompt framework outlined in Appendix Table 6.

The evaluation prompt framework, adapted from Liu et al., 2023b, comprised three components: task description, evaluation rubric, and evaluation steps. A chain-of-thought approach in the evaluation step, detailing key points and potential deductions emphasized in human evaluation, demonstrated modest improvement in assessment performance.

These results, shown in Table 3, indicate strong correlation and substantial agreement between human evaluations and LLM-as-a-Judge assessments for these tasks.

TelInstruct Given the complex, agent-like characteristics of TelInstruct, LLM-as-a-Judge evaluation is more appropriate than automatic evaluation methods. To address diversity and scalability challenges in agent benchmarking, a system based on PairEval(Park et al., 2024), a reference-free method, was designed. Evaluation prompts were developed drawing inspiration from Prometheus2(Kim et al., 2024) and G-Eval(Liu et al., 2023b) frameworks for assessing generated responses.

Recent studies(Wang et al., 2023; Zheng et al., 2023) have identified a position bias in LLMs when evaluating pairs of model responses. To mitigate this bias, we implemented a two-stage evaluation process, where the Eval LLM assesses Response A followed by Response B and then evaluates Response B followed by Response A. If evaluations across both orderings are consistent, we classify the case as a "WIN", and inconsistent cases are deemed comparable in response quality and classified as a "TIE".

4.2 Evaluation Results

This section presents the results of evaluating the telco-specific performance of various proprietary and open source LLMs using the TelBench framework.

Table 4 demonstrates that while Claude 3.5 Sonnet shows the best overall performance among proprietary LLMs, recently released open-sourced models, such as Llama-3.1-405B-Instruct-FP8 and Mistral-Large-Instruct, exhibit performance that is comparable to proprietary models. The distribution of results varies significantly between tasks. For sentiment classification and summary generation, models show similar performance with minimal variance. However, tasks requiring specialized telco knowledge, such as entity and intent recognition, still highlight limitations in open-source models, with scores trailing about 0.1 behind the top-

Language Model	Summary	Topic
GPT-4-Turbo	4.35	3.63
Claude 3.5 Sonnet	4.05	3.83
Claude 3 Haiku	3.89	3.15
Llama-3.1-405B-Instruct-FP8	4.09	3.14
Mistral-Large-Instruct	2.77	3.24
Mistral-Small-Instruct	2.47	2.01

Table 5: LLM-as-a-Judge results for Summary and Topic

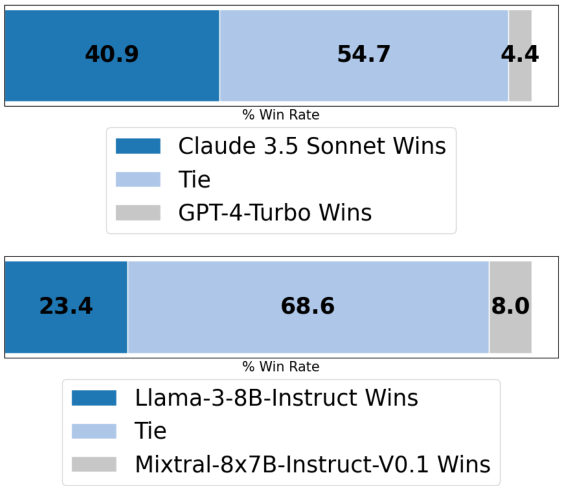

Figure 2: LLM-as-a-Judge results for Workflow

performing proprietary models.

The telco-related tasks demand a nuanced understanding and appropriate application of domain-specific knowledge and terminology. While many open-source models continue to demonstrate notable limitations in telco-specific tasks, recent evaluations of Llama-3.1-405B-Instruct-FP8 and Mistral-Large-Instruct show encouraging improvements, particularly in Q&A tasks. These models exhibit strong comprehension abilities, allowing them to generate more contextually appropriate responses, especially in customer service scenarios. This performance narrows the gap between proprietary and open-source language models. However, despite these gains, smaller open-source models still struggle with comprehending and responding to domain-specific queries effectively.

Table 5 illustrates that GPT-4-Turbo maintains superior performance in the LLM-as-a-Judge summary evaluation, with Claude 3.5 Sonnet showing the best results in topic-related tasks. Interestingly, the Llama-3.1-405B-Instruct-FP8 model also performs competitively in summary evaluation, outperforming several other open-source models, though it still falls behind in topic-based tasks. These findings do not fully align with results from automated evaluation methods. Nonetheless, the strong correlation between the proposed LLM-as-a-Judge methodology and human evaluation highlights the importance of combining both quantitative metrics and qualitative insights for a comprehensive understanding of an LLM's capabilities. This suggests that evaluations should integrate both objective measurements and subjective judgments to capture a more nuanced picture of model performance.

Figure 2 reveals that while the quality of workflow responses is generally comparable across models, notable differences exist. Claude 3.5 Sonnet demonstrates superior performance compared to GPT-4-Turbo, and Llama-3-8B-Instruct out performs Mixtral-8x7B-Instruct-v0.1. These nuanced differences in model performance provide valuable insights into the relative strengths (and weaknesses) of various LLMs in the context of telco-specific tasks.

5 Conclusion

This paper introduces TelBench, the first (to our knowledge) benchmark dataset focused on telco customer service centers, and evaluates the performance of Large Language Models (LLMs) using the dataset we designed and built. Leveraging proprietary assets and domain expertise, we have created a benchmark dataset to measure the telco-specific performance of various LLMs, both proprietary and open source.

The methodology employed in developing TelBench can be extended to create specialized training datasets for LLMs in the telco sector, and such datasets can help facilitate the development of LLMs optimized for telco-specific tasks. Future research will focus on the development and performance evaluation of these specialized telco LLMs. Additionally, we plan to expand the scope of TelBench to include other areas of the telco industry, such as infrastructure operations, task planning, and contract reviews. Furthermore, we are preparing further LLM-based evaluation methods to reduce the burden of human assessment and also developing an Evaluation-as-a-Service platform.

References

Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, and Renaud Duval. 2023. Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmology Science*, 3(4):100324.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. Adapting large language models via reading comprehension. *Preprint*, arXiv:2309.09530.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. Saullm-7b: A pioneering large language model for law. *Preprint*, arXiv:2403.03883.

Dorottya Demszky and Heather Hill. 2023. The ncte transcripts: A dataset of elementary math classroom transcripts. *Preprint*, arXiv:2211.11772.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023a. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Preprint*, arXiv:2305.01210.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Ali Maatouk, Fadhel Ayed, Nicola Piovesan, Antonio De Domenico, Merouane Debbah, and Zhi-Quan Luo. 2023. Teleqna: A benchmark dataset to assess large language models telecommunications knowledge. *Preprint*, arXiv:2310.15051.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. 2024. Paireval: Open-domain dialogue evaluation with pairwise comparison. *arXiv preprint arXiv:2404.01015*.

Guijin Son, Hanearl Jung, Moonjeong Hahm, Keonju Na, and Sol Jin. 2023. Beyond classification: Financial reasoning in state-of-the-art language models. *Preprint*, arXiv:2305.01505.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *Preprint*, arXiv:2306.06031.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, Dongzhan Zhou, Shufei Zhang, Mao Su, Han-Sen Zhong, and Yuqiang Li. 2024. Chem-llm: A chemical large language model. *Preprint*, arXiv:2402.06852.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Hang Zou, Qiyang Zhao, Yu Tian, Lina Bariah, Faouzi Bader, Thierry Lestable, and Merouane Debbah. 2024. Telecomgpt: A framework to build telecom-specfic large language models. *Preprint*, arXiv:2407.09424.

A Benchmark Samples

- TelTask
  - **Sentiment**: Classification of customer sentiment (Positive/Negative/Neutral) in consultation dialogues
  - **Entity**: Extraction of essential entities and categories in customer utterances
  - **Intent**: Classification of customer's intent(s) into broad categories and specific services from customer inquiries
  - **To-do**: Generation of task-oriented to-do lists derived from customer service dialogues
  - **Topic**: Generation of topics from customer service dialogues
  - **Summary**: Generation of summaries from customer service dialogues
  - **Safety**: Determine unsafe utterances that may occur during consultation
- TelInstruct
  - **Workflow**: Extraction of essential function calls, such as database searches from customer service dialogues to facilitate the development of LLM-based customer service agents
  - **Telco Q&A**: A task involving the generation of responses to potential telco-related customer inquiries
  - **MRC**: A task designed to evaluate the LLM's ability to provide accurate responses to queries based on telco product guides and documentation

A.1 Sentiment

A.1.1 Korean (original)


```
{
"dialog": [
  {
  "channel": "상담사",
  "text": "반갑습니다 김지원입니다."
  }, {
  "channel": "고객",
  "text": "네, 제가 미납 요금을 가상계좌로 입금했는데 통장에서도 빠져나갔습니다."
  }, {
  "channel": "상담사",
  "text": "아 그렇군요, 고객님 번호가 010-1234-5678인 고객님 본인 맞으세요?"
  }, {
  "channel": "고객",
  "text": "네, 맞습니다."
  }, {
  "channel": "상담사",
  "text": "네, 감사합니다. 얼른 확인해보겠습니다. 잠시만 기다려 주세요."
  }, {
  "channel": "상담사",
  "text": "기다려 주셔서 감사합니다. 고객님 말씀처럼 어제 날짜에 납부하셨던 건 정확하게 확인되는데요. 월정액 통장에서 저희가 32400원을 다시 인출 시도 들어갔던 날짜도 있습니다. 아직 은행쪽에서 결과가 확인되지 않아서 이 부분은 내일 저희 전산에 반영이 될 예정입니다. 만약에 전산이 반영이 되서 두 번 납부된 금액이 맞다면 고객님께 그 계좌로 다시 자동송금이나 연락 주시면 저희가 다시 접수해서 개별송금으로 진행을 해 드리겠습니다. 그래서요, 두 번 받았던 금액을 돌려드리니까요 염려 안 하셔도 되지만 일단 오늘 확인은 안 됩니다."
  }, {
  "channel": "고객",
```

"text": "내일 확인되면 그냥 자동이체 계좌로 입금된다는 거죠?"
},{
"channel": "상담사",
"text": "네, 그렇습니다. 내일 다시 한 번 확인 부탁 드리겠습니다. 김지원이었습니다. 바로 도움드리지 못해 죄송합니다."
},{
"channel": "고객",
"text": "네, 알겠습니다."
}
],
"sentiment": "neutral"
}

A.1.2 English (translated)

{
"dialog": [
{
"channel": "Agent",
"text": "Hi, I'm Jiwon Kim."
},{
"channel": "Customer",
"text": "Yes, I deposited the unpaid amount into the virtual account and it has also been deducted from my bank account."
},{
"channel": "Agent",
"text": "Oh, right, are you the customer whose number is 010-1234-5678?"
},{
"channel": "Customer",
"text": "Yes, that's right."
},{
"channel": "Agent",
"text": "Okay, thank you. I'll check it out, just give me a second."
},{
"channel": "Agent",
"text": "Thank you for your patience. As you said, the payment you made on yesterday's date is correct, and there is also a date when we tried to withdraw 32400 won from your monthly account again. The bank hasn't confirmed the result yet, so this will be reflected in our system tomorrow. If it does, and it's the correct amount that was paid twice, we'll send you a direct deposit back to that account or you can contact us and we'll take it back and process it as a separate payment. So, yes, we'll refund the amount that you received twice, so don't worry, but we won't be able to confirm it today."
},{
"channel": "Customer",
"text": "If it's confirmed tomorrow, it'll just go into my direct deposit account, right?"
},{
"channel": "Agent",
"text": "Yes, that's right, we'll check back with you tomorrow. This was Jiwon Kim. I apologize for not being able to help you right away."
},{

```
  "channel": "Customer",
  "text": "OK. Thank you."
  }
  ],
"sentiment": "neutral"
}
```

A.2 Entity

A.2.1 Korean (original)

```
{
"input": "T다이렉트샵에서 갤럭시 Z 폴드5 사기",
"output": [
  {
  "name": "T다이렉트샵",
  "entity_type": "SERVICE_NAME"
  }, {
  "name": "갤럭시 Z 폴드5",
  "entity_type": "DEVICE_NAME"
  } ]
}
```

A.2.2 English (translated)

```
{
"input": "Buying Galaxy Z Fold 5 at T-Direct Shop",
"output": [
  {
  "name": "T-Direct Shop",
  "entity_type": "SERVICE_NAME"
  }, {
  "name": " Galaxy Z Fold 5",
  "entity_type": "DEVICE_NAME"
  } ]
}
```

A.3 Intent

A.3.1 Korean (original)

```
{
"text": "현재 남은 음성 통화 잔여량 확인",
"intent": "Check.VoiceRemaining"
},
```

A.3.2 English (translated)

```
{
"text": "See how much voice call time you have left",
"intent": "Check.VoiceRemaining"
},
```

A.4 To-do

A.4.1 Korean (original)

```
{
"dialog": [
  {
```

"channel": "상담사",
"text": "반갑습니다 김지원입니다"
}, {
"channel": "고객",
"text": "제가 로밍을 했어요 시월초에 티전화 설치를 취소를 할려면 어떻게 해야 하나요"
}, {
"channel": "상담사",
"text": "아 그럼 티 전화 고객님 가입하신 거 취소하신다는 말씀이세요"
}, {
"channel": "고객",
"text": "네"
}, {
"channel": "상담사",
"text": "아 그럼 고객님 이게 통화 중에는 설정이 안 되다 보니까 제가 경로 문자 넣어드리면 그대로 고객님 한 번 처리해 주시겠어요?"
}, {
"channel": "고객",
"text": "알겠습니다"
}, {
"channel": "상담사",
"text": "네 그럼 제가 고객님 경로 지금 메모해서 고객님께 문자 남겨 놓도록 하겠습니다"
}, {
"channel": "고객",
"text": "알겠습니다 감사합니다"
}, {
"channel": "상담사",
"text": "네 감사합니다 김지원이었습니다"
}],
"todo": [
"- 문자 발송: 티전화 설치 취소 경로"]
},

A.4.2 English (translated)

{
"dialog": [
{
"channel": "Agent",
"text": "Hi, I'm Jiwon Kim."
}, {
"channel": "Customer",
"text": "I was roaming in early October. How can I cancel my T-Phone installation?"
}, {
"channel": "Agent",
"text": "Oh, so you're canceling the T-phone subscription?"
}, {
"channel": "Customer",
"text": "Yes"
}, {
"channel": "Agent",
"text": "Oh, as you can't set this up while you're on a call, I'll just put in the route text and you can just take care of it?"
}, {

"channel": "Customer",
"text": "Ok"
}, {
"channel": "Agent",
"text": "Okay. I'll take note of your route now and text you a message."
}, {
"channel": "Customer",
"text": "Fine. Thank you."
}, {
"channel": "Agent",
"text": "Yes, thank you. It was Jiwon Kim."
}],
"todo": [
"- Send a text: Route to cancellation of T-phone installation"]
},

A.5 Topic

A.5.1 Korean (original)

{
"dialog": [
{
"channel": "상담사",
"text": "반갑습니다 김지원입니다."
}, {
"channel": "고객",
"text": "여보세요."
}, {
"channel": "상담사",
"text": "네 안녕하세요."
}, {
"channel": "고객",
"text": "아 네 그 저 요금 내역 확인해서 연말정산으로 보낼려고 하는데 작년 꺼 좀 팩스로 받을려고요."
}, {
"channel": "상담사",
"text": "아 네 확인해 도움드리겠습니다. 문의하시는 번호가 010-1234-5678 번 송정하 고객님 본인 되십니까? 네 그러시면 받아보실 팩스 번호 한 번 천천히 불러주시겠어요?"
}, {
"channel": "고객",
"text": "네 서울이고, 02-123-4567이요"
}, {
"channel": "상담사",
"text": "예 말씀해주셨던대로 작년 수납내역서 팩스로 발송처리 해드리도록 하겠습니다."
}, {
"channel": "고객",
"text": "네 감사합니다."
}, {
"channel": "상담사",
"text": "네 감사합니다. 좋은 하루 되세요." }
],
"topics": ["요금 내역", "팩스 발송", “수납내역서"],

```
},
```


A.5.2 English (translated)


```
{
"dialog": [
{
  "channel": "Agent",
  "text": "Hi, I'm Jiwon Kim."
  },   {
  "channel": "Customer",
  "text": "Hello."
  },   {
  "channel": "Agent",
  "text": "Yes, how can I help you?"
  },   {
  "channel": "Customer",
  "text": "Oh yeah, I'm going to check last year's bills and send them to the year-end reconciliation. Can I get them via fax?"
  },   {
  "channel": "Agent",
  "text": "OK, let me check. The number you're calling is 010-1234-5678. Is that you? And, if it is, can you please say the fax number slowly?"
  },   {
  "channel": "Customer",
  "text": "Yes, it's Seoul, and 02-123-4567."
  },   {
  "channel": "Agent",
  "text": "Yes, we will fax you last year's statement as you mentioned."
  },   {
  "channel": "Customer",
  "text": "Ok, thank you."
  },   {
  "channel": "Agent",
  "text": "Thanks. Have a nice day."   }
  ],
"topics": ["Bills", "Fax request", "Statement" ],
}
```


A.6 Summary

A.6.1 Korean (original)


```
{
"dialog": [
  {
  "channel": "상담사",
  "text": "반갑습니다, 이지훈입니다. 무엇을 도와드릴까요?"
  }, {
  "channel": "고객",
  "text": "네, 수고하십니다. 인터넷 연결을 하고 싶은데요."
  }, {
  "channel": "상담사",
  "text": "인터넷 연결이라면 혹시 지금 인터넷이 잠깐 끊겨져 있는 건가요?"
```

}, {
"channel": "고객",
"text": "아니요, 신규로 가입하고 싶습니다."
}, {
"channel": "상담사",
"text": "아, 김유연 고객님 본인이신가요? 문의 주신 번호가 010-1234-5678인데, 자택에 지금 고객님이 결합해서 인터넷 개통이 되어 있다라고 나오는데 혹시 다른 가족분 자택에 인터넷이 설치돼 있는 건가요?"
}, {
"channel": "고객",
"text": "네, 그렇습니다. 지금 제가 따로 살고 있어서요"
}, {
"channel": "상담사",
"text": "그러면 새로운 주소에 인터넷을 새로 설치해야 하는 거군요."
}, {
"channel": "고객",
"text": "네, 그렇습니다."
}, {
"channel": "상담사",
"text": "알겠습니다. 바로 연결하겠습니다. 잠시만 기다려 주세요. 김유연 고객님, 좋은 상담 이어가시길 바랍니다."
}],
"summary": "고객이 자택이 아닌 새로운 주소에 인터넷 신규 설치를 원하였고, 상담사는 이를 위해 상담을 연결해주겠다고 말하였다."
}

A.6.2 English (translated)

{
"dialog": [
{
"channel": "Agent",
"text": "Hi, it's Ji-Hoom Lee. How can I help you?"
}, {
"channel": "Customer",
"text": "Yes, thank you. I'd like to connect to the internet."
}, {
"channel": "Agent",
"text": "You mean you are experiencing temporary trouble in connecting to the internet?"
}, {
"channel": "Customer",
"text": "No, I want to sign up for a new subscription."
}, {
"channel": "Agent",
"text": "Oh, is this you, Yoo Yeon Kim, and the number you gave us is 010-1234-5678? It says that you have a combined internet connection at your home. Do you have the internet at another family member's home?"
}, {
"channel": "Customer",
"text": "Yes, I do, because I'm living on my own now."
}, {
"channel": "Agent",

```
  "text": "Then you'll need to reinstall the internet at the new address."
  },  {
  "channel": "Customer",
  "text": "Yes, that’s right."
  },  {
  "channel": "Agent",
  "text": "Okay, we'll connect you right away. Please hold on a moment. Thank you, Yoo Yeon Kim, and good luck with your call."
  }    ],
"summary": "The customer wanted a new internet installation at a new address that was not their home, and the agent said he would set them up with a consultation."
}
```

A.7 Workflow

A.7.1 Korean (original)

```
<|im_end|>
<|im_start|>assistant
반갑습니다. 강하루입니다.<|im_end|>
<|im_start|>user
카드를 갱신했더니 자동이체 설정을 다시 하라고 하셔서요. 어떻게 변경하는 거예요?<|im_end|>
<|im_start|>assistant
<|action_start|><|plugin|>
{"name": "db_search", "parameters":
{"query": "카드 갱신 후 자동이체 설정 방법"}}
<|action_end|><|im_end|>
<|im_start|>environment name=<|plugin|>
{"output": [{"title": "자동이체(납부)구비서류 > 자동이체 신청/변경_개인",
"text": "참고사항 해지 고객 자동납부 신청/변경/취소는 고객센터만 가능..."}]}
<|im_end|>
<|im_start|>assistant
자동이체를 변경하실 번호가 010-1234-5678번이 맞습니까?<|im_end|>
```

A.7.2 English (translated)

```
<|im_end|>
<|im_start|>assistant
Hi, it’s Haru Kang.<|im_end|>
<|im_start|>user
I renewed my card and it says I need to set up direct debit again.
How do I change it?<|im_end|>
<|im_start|>assistant
<|action_start|><|plugin|>
{"name": "db_search", "parameters":
{"query": "Setting up direct deposit after card renewal"}}
<|action_end|><|im_end|>
<|im_start|>environment name=<|plugin|>
{"output": [{"title": "Direct Debit (Payment) Documents > Direct Debit Request/Change_Individual",
"text": "Note: Only customer center can apply for/change/cancel automatic payment for terminated customers..."}]}
<|im_end|>
<|im_start|>assistant
Is 010-1234-5678 the correct number to change your direct deposit?<|im_end|>
```

A.8 Telco Q&A

A.8.1 Korean (original)

{
"question":"FLO 고객센터 연락처 알려주세요",
"answer":"FLO 고객센터 연락처는 1234-5678입니다. 운영시간은 평일 09:00~18:00입니다. [1]"
}

A.8.2 English (translated)

{
"question":"What is the FLO customer support number?",
"answer":"The FLO customer support number is 1234-5678. The hours of operation are 09:00-18:00 on weekdays. [1]"
}

A.9 MRC

A.9.1 Korean (original)

{
"document": "22. 7. 1(금) 신규가입 중단 예정\n☞ [가입중단] 미주/오세아니아 주요국 내 제휴 사업자망을 통해 고객이 설정한 개시일시(한국시간 기준)부터 30일 동안 3G/LTE 데이터를 저렴하게 이용할 수 있는 요금제\n\n 홈 > 상품매뉴얼 > 이동전화 > 요금제 > Outbound 로밍\n2022-05-31T10:15:20.000Z\n\n\n 1) Swing [요금제/부가상품변경] 화면 > '부가상품' 탭> ‘T로밍 미주패스' 입력 > 추가 버튼 > ‘개시일시’ 설정 후 가입\n\n image.jpg \n\n 2) Swing [OB로밍서비스 관리] 화면 > 부가정보 > 부가상품 > ‘T로밍 미주패스' 입력 > 개시일시 설정 후 가입 가능\n\n image.jpg \n\n \n\n**2. 참고사항** \n 1) 데이터로밍무조건차단 서비스 가입 상태에서 T로밍 미주패스 가입 시 데이터로밍무조건차단 서비스 자동 해지 \n 2) T로밍 미주패스 해지 시 데이터로밍무조건차단 서비스 자동 가입 \n\n 3) T로밍 미주패스 해지 당일 재가입 가능 \n\n \n\n \n\n\n",
"question": "T로밍 미주패스 기간 얼마나 돼요?",
"answer": "30일"
}

A.9.2 English (translated)

{
"document": "22. 7. 1(Fri) New subscriptions will be discontinued\n☞ [Discontinued] A plan that allows you to use 3G/LTE data at a low price for 30 days from the start date (Korea time) set by the customer through a network of partner operators in major countries in the Americas/Oceania\n\n Home > Product Manual > Mobile Phone > Plan > Outbound Roaming\n2022-05-31T10:15:20. 000Z\n\n\n 1) Swing [Change plan/add-on] screen > 'Add-on' tab > Enter 'T-Roaming Americas Pass' > Add button > Set 'Start date' and sign up\n\n image. jpg \n\n 2) Swing [OB Roaming Service Management] screen > Additional Information > Additional Products > Enter 'T Roaming Americas Pass' > Set the start date and sign up\n\n image.jpg \n\n \n\n**2. Notes** \n 1) Data roaming unconditional blocking service is automatically canceled when subscribing to T-Roaming Americas Pass while subscribed to data roaming unconditional blocking service \n 2) Data roaming unconditional blocking service is automatically subscribed when canceling T-Roaming Americas Pass \n\n 3) Re-subscription is possible on the day of T-Roaming Americas Pass cancellation \n\n \n\n \n\n\n",
"question": "How long is the T-Roaming Americas Pass valid for?",
"answer": "30 Days"
}

B Evaluation Prompt Framework sample

Task description
The evaluation process comprises the following components: instructions, prompts, responses to be assessed, a scoring rubric delineating assessment criteria, and evaluation steps. 1. Construct detailed feedback evaluating the quality of the response, adhering strictly to the provided scoring rubric. 2. Based on the feedback, assign an integer score between 1 and 5, referencing the scoring rubric for guidance. ...
Evaluation rubric
5 (Excellent Topic Quality): The topics effectively encapsulate the essential information representative of the consultation dialogue. They are concisely generated in consistent (compound) nouns, accurately and specifically reflecting the content of the consultation. 4 (Good Topic Quality): The majority of topics effectively capture the key information representative of the consultation dialogue. They accurately and specifically reflect the consultation content and consiste of (compound) nouns. ...
Evaluation Steps
Carefully analyze the consultation dialogue, understand the content, and subsequently identify key topics that encapsulate the essential elements in the dialogue. ...

Table 6: Evaluation Prompt Framework