

Jago Grahak Jago: Consumer Grievance Redressal through Large Language Models

Shrey Ganatra¹, Swapnil Bhattacharyya¹, Harshvivek Kashid¹, Spandan Anaokar¹,
Shruti Nair², Reshma Sekhar², Siddharth Manohar², Rahul Hemrajani²,
Pushpak Bhattacharyya¹

¹Indian Institute of Technology Bombay

²National Law School of India University, Bangalore

{ganatrashrey2002, harshvivek14, spandananao, pushpakbh}@gmail.com
swapnilbhyya@cse.iitb.ac.in

Abstract

Jago Grahak Jago is a consumer awareness initiative that translates to *Wake Up, Consumers, Wake Up*. Consumer grievance redressal remains a complex process due to procedural complexity, legal jargon, and barriers like jurisdiction and limitation periods, despite various initiatives aimed at simplifying it. To address this, we present **Grahak-Nyay**¹ (Justice-to-Consumers), a chatbot designed to simplify the consumer grievance redressal process for Indian consumers. Powered by open-source Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG), *Grahak-Nyay* is supported by a concise and up-to-date Knowledge Base. In addition, we release three key datasets: *GeneralQA*, based on general consumer laws; *SectoralQA*, focusing on sector-specific knowledge; and *SyntheticQA*, for evaluating RAG performance. We also introduce *NyayChat*, a dataset of 300 annotated conversations between users and the chatbot. We propose the HAB metrics to measure the chatbot’s performance, which evaluates conversations based on Helpfulness, Accuracy, and Brevity. Legal domain experts have tested the chatbot; reference-based and reference-free evaluations confirm its effectiveness. Demo video is available on github¹.

1 Introduction

Chatbots using Large Language Models (LLMs) such as GPT-4 (Achiam et al., 2023) and Llama-3 (Dubey et al., 2024) are gaining immense popularity due to their wide applications in domains such as finance (Zhao et al., 2024), tourism (Meyer et al., 2024), medical (Mishra et al., 2023), education (Lee et al., 2023), and customer support (Obadinma et al., 2022).

The use of LLMs in the legal domain has mainly been focused on Legal-Judgement Prediction (Shui et al., 2023), Summarization, Prior Case Retrieval, Legal Statute Identification, (Joshi et al., 2024; Feng et al., 2024) etc. There has been some development of legal LLMs (Zhou et al., 2024), but none focused on consumer law.

Consumer grievance redressal in India continues to be a challenge. Although the Department of Consumer Affairs² has implemented various initiatives, such as the National Consumer Helpline³, a comprehensive FAQ section, and the e-Daakhil⁴ portal to facilitate the process of filing consumer complaints, these measures do not assist people without legal training. Although filing a consumer complaint does not necessarily require legal representation, the complexity of the procedure, nuanced language of the law, and intricacies like jurisdiction and limitation periods act as barriers.

The e-Daakhil portal, while helpful, mainly provides generic guidelines that do not tailor advice to the unique facts of an individual’s case. The portal does not offer actionable guidance on crafting a specific notice or complaint, nor does it address essential variables such as time limitations for filing or the appropriate jurisdiction for particular grievances. Templates for complaints and notices are available on various platforms but often suffer from being either too generic to be useful or poorly drafted to be reliable.

1,02,976 cases have been registered every month in the financial year 2023-2024 on the National Consumer Helpline (Ministry of Consumer Affairs). Currently, these complaints are noted by trained staff. This indicates the need for our chatbot to accelerate the resolution of consumer grievances.

Publicly available chatbots like ChatGPT⁵ or

¹The full implementation code, data, and demo video are available at: <https://github.com/ShreyGanatra/GrahakNyay>

²<https://consumeraffairs.nic.in/>

³<https://consumerhelpline.gov.in/>

⁴<https://edaakhil.nic.in/>

⁵<https://chatgpt.com/>

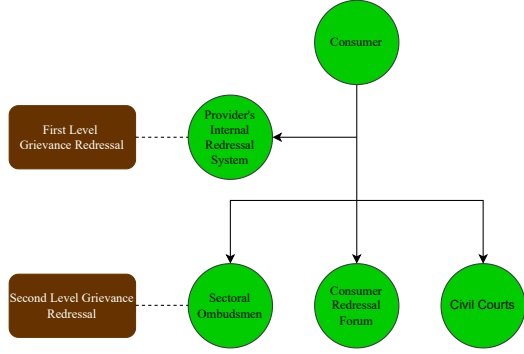


Figure 1: Consumer grievance redress in the financial sector

Claude⁶ also mainly provide generic guidelines and do not ask for more details about the problem, which can be seen in Fig. 10.

In India, a two-level consumer grievance redressal system is followed Fig.1 (Chivukula, 2021), where the consumer is encouraged at the first level to contact the provider for redress. If the consumer is not satisfied, they may take action in government bodies. Fig.2 (Reserve Bank of India) shows the reasons for the rejecting the complaint. These include wrong representation, jurisdiction or limitations.

Our chatbot is designed to offer helpful responses to standard consumer inquiries, making legal assistance available to everyone, regardless of location or financial status. In addition, the chatbot will help craft legal documents specific to individual scenarios, including forms and letters that can be sent directly to appropriate legal authorities. The chatbot will also provide detailed step-by-step guidance on the legal procedures that must be followed for legal complaints.

Our contributions are:

1. A consumer grievance redressal chatbot, **Grahak-Nyay**, tailored for Indian consumers, utilizing an open-source Large Language Model powered by Retrieval-Augmented Generation, aided by a concise Knowledge Base with the latest information (Section 3).
2. We release **GeneralQA**, a question-answer dataset based on general consumer laws; **SectoralQA**, a question-answer dataset based on sector-wise knowledge of consumer laws; and **SyntheticQA**, a question-answer dataset to

⁶<https://claude.ai/>

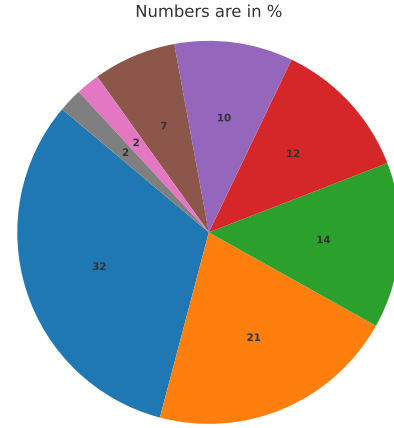


Figure 2: Reasons for Banking Ombudsman rejecting complaints in FY 2018-19

evaluate the performance of RAG (Section 2.1).

3. We release **NyayChat**, a dataset containing 300 annotated conversations between users and the chatbot, based on various issues and complaints (Section 2.2).
4. We propose the **HAB** metrics to assess the quality of chatbot conversations based on **Helpfulness**, **Accuracy**, and **Brevity**. We conduct experiments on reference-free automatic evaluation of conversations using various *LLM-based* evaluators based on the HAB metrics (Section 4).

2 Dataset

2.1 Knowledge Base

2.1.1 GeneralQA

This dataset provides an overview of the Consumer Protection Act 2019 (Ministry of Law and Justice). It presents the Act’s provisions in a structured format of 53 questions and answers, offering a comprehensive understanding of its applicability. Additionally, it defines key concepts such as ‘consumer’ and ‘goods,’ clarifying the scope within which the chatbot is designed to operate.

2.1.2 SectoralQA

This dataset encompasses detailed information from 23 sectors, including real estate, telecommu-

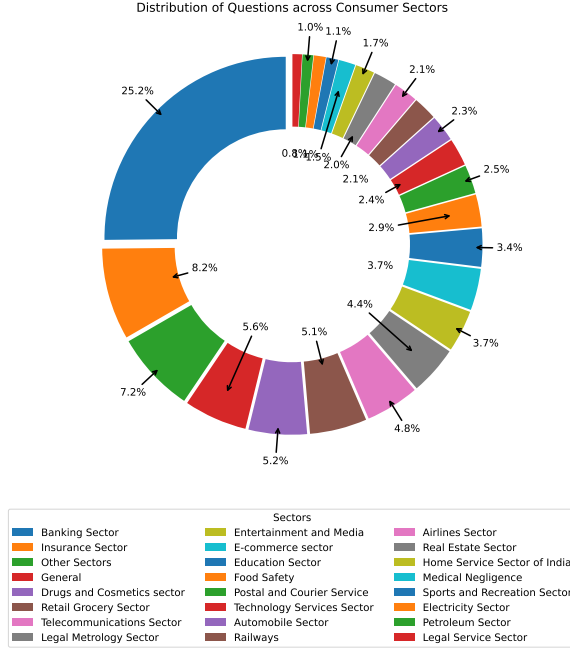


Figure 3: Distribution of corpus questions across different consumer sectors

nications, sports and recreation. The corpus comprises of 889 questions and answers, each tailored to address sector-specific concerns. A breakdown of the sector-wise distribution, along with the general corpus, is provided in Fig. 3.

2.1.3 SyntheticQA

To assess our RAG system, we generated five user questions that yield the same answer across both the general corpus and the sector-specific corpus, using the Llama-3.1-70B-Instruct⁷ model. This amounts to generating 4734 (in some cases, 6 instead of 5 questions) questions. The answers were generated using the Llama-3.1-8B-Instruct⁸ model, keeping corresponding answer in the context. We call this dataset as **SyntheticQA**. The prompt used for generation can be found in Fig.9.

2.2 NyayChat

This dataset includes 303 simulated conversations, meticulously crafted by a team of legal experts specializing in various sectors such as e-commerce, medical negligence, railways, airlines, and more. Each conversation mirrors a real-world interaction between a user and the chatbot, addressing specific grievances that fall under the purview of consumer

⁷<https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct>

⁸<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

law.

3 Methodology

3.1 Retrieval-Augmented Generation

Retrieval Augmented Generation (RAG) is a prominent approach used in real-world applications for grounding large language model (LLM) generations in up-to-date and domain-specific knowledge. It has been observed (Lazaridou et al., 2022; Shuster et al., 2021; Ren et al., 2023) that RAG reduces hallucinations and improves answer quality, without the need for highly expensive and sometimes fragile domain-specific fine-tuning.

A typical RAG framework involves a retrieval system that fetches documents that are relevant to the query. These documents are then used as context, prompting the LLM to generate the required response. For our chatbot, we also include the chat history as part of the prompt. The RAG framework is divided into the following parts in our architecture.

3.1.1 Chunking

We observed that fixed-length chunking proved inadequate for our purposes. Long answers were often split across multiple chunks, resulting in the loss of crucial information during retrieval. Conversely, grouping multiple short answers in a single chunk introduced unnecessary noise. As a result, we adopted a chunking strategy where each chunk contains only one Question-Answer pair, ensuring clarity and precision in information retrieval.

3.1.2 Query Rewriting

To implement Retrieval-Augmented Generation (RAG), it is essential for each query to be properly contextualized. To accomplish this, we utilized Llama-3.1-8B-Instruct along with the instruction provided in Fig.6. Initially, we observed that the Large Language Model (LLM) answered the query directly rather than rewriting it as intended. To improve performance, we utilized one-shot prompting by incorporating a single example in the prompt, which led to significantly better results.

3.1.3 Retriever

The chunks are embedded using mixedbread-ai/mxbai-embed-large-v1⁹ and stored in a vector-store. The retriever is responsible for extracting relevant chunks using the query. We rely on cosine

⁹<https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1>

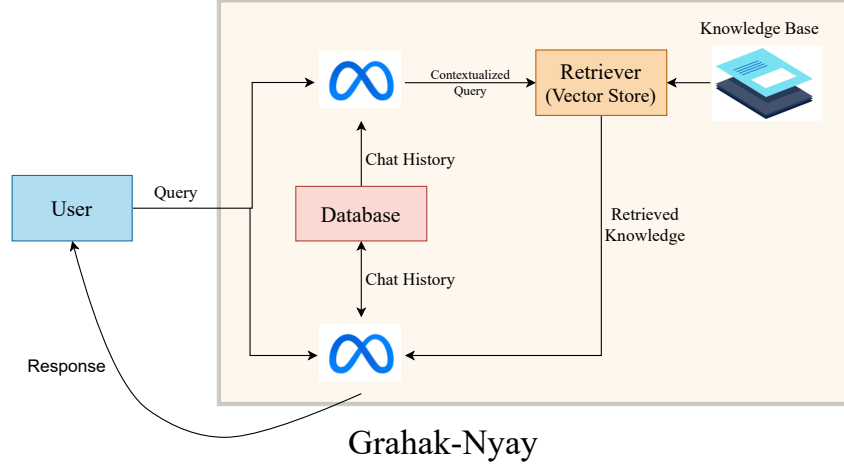


Figure 4: Architecture of *Grahak-Nyay* chatbot. The user query is first contextualised using an LLM to form an independent query to perform RAG. The retrieved knowledge is then given to the LLM along with the chat history to generate appropriate response to the user query.

Reference-based evaluation						Reference-free evaluation		
ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	METEOR	BLEU	Helpfulness	Accuracy	Brevity
66.9	41.1	33.2	90.9	41.9	37.4	4.51	3.89	3.22

Table 1: Performance of *Grahak-Nyay chatbot* on Reference-based and Reference-free evaluation. We evaluated Grahak-Nyay chatbot on 65 conversations for which reference was available. We performed LLM-based automatic evaluation on HAB metrics on the 5-point Likert scale using Llama-3.1-70B model.

similarity and extract the best four chunks and use it for response generation.

3.1.4 Generation

The chatbot is meant to be interactive and conversational. Hence, for each user input, while keeping the retrieved documents as a reference, the chatbot also takes into account the chat history. This ensures that all responses are relevant and grounded in the contextual history of the issue. We use Llama-3.1-8B-Instruct model to generate our responses.

3.2 Hallucination

A significant barrier to the wide use of LLMs in multiple domains is their tendency to hallucinate. It has been observed that in spite of clear instructions, LLM generates text which might be false or irrelevant. RAG-based approaches help to tremendously reduce this phenomenon, but even then, hallucination remains a major concern.

A domain-specific chatbot like our Grahak-Nyay chatbot implies that RAG is responsible for providing much of the context and domain knowledge that will be utilized for the chat. In such cases, the LLM must generate content based on the RAG Corpus only and never contradict it. For this pur-

pose, we modified the prompt where we specified the LLM to answer any out-of-corpus question by stating that it does not know the answer.

3.3 Chatbot

We utilize Langchain¹⁰ to build our chatbot. The link to the chatbot is: https://www.cfilt.iitb.ac.in/consumer_chatbot

3.4 Other methods

Along with the architecture used in the chatbot, techniques like long-context and fine-tuning were also explored. In long-context, we gave the entire general and sectoral corpus as the context to the LLM instead of the retrieved documents, thereby avoiding RAG. However we found out that this method led to huge latency and a lot of hallucination. We also attempted finetuning using *NyayChat* dataset but the generation was restricted. We will continue to explore this as our future work.

4 Evaluation

We strongly believe that any user-facing chatbot should help the user address the query, should be

¹⁰<https://langchain.com/>

Models	Helpfulness		Accuracy		Brevity	
	r_{pb}	ρ	r_{pb}	ρ	r_{pb}	ρ
Gemma-2-9B	0.256	0.242	0.113	0.102	0.183	0.182
Llama-3.1-8B	0.386	0.246	0.225	0.213	0.154	0.153
Llama-3.2-11B-text-preview	0.455	0.343	<u>0.342</u>	<u>0.232</u>	<u>0.239</u>	<u>0.204</u>
Mixtral-8x7B	<u>0.557</u>	<u>0.490</u>	0.205	0.207	0.159	0.141
Llama-3.1-70B	0.689	0.627	0.461	0.430	0.430	0.418

Table 2: Performance metrics for various models based on Helpfulness, Accuracy, and Brevity metrics. Each metric includes point biserial correlation (r_{pb}) and Spearman’s rank correlation coefficient (ρ) scores for each model. The best scores are bolded and the second-best scores are underlined.

accurate while doing so, and should keep the user engaged. We assess the quality of chatbot conversations using **HAB metrics: Helpfulness, Accuracy, and Brevity**. HAB metrics allow us to assess not only how effectively the chatbot addresses user issues and provides accurate information but also how concisely it communicates these responses. The HAB metrics are defined as follows:

- **Helpfulness:** This metric assesses how helpful the chatbot was in resolving the user’s issue or query. It evaluates chatbot’s ability to understand the user’s problem and provide actionable, relevant, and clear resolution.
- **Accuracy:** This metric evaluates the correctness of the information provided by the chatbot in response to user queries, ensuring that the responses are factually accurate and reliable.
- **Brevity:** This metric measures the conciseness of the chatbot’s responses, ensuring efficient communication without unnecessary elaboration. It ensures efficient communication by focusing on delivering essential information straight to the point, while avoiding excessive questioning or verbosity.

Through this comprehensive evaluation framework, we aim to enhance the effectiveness of chatbots in addressing consumer grievances and improving overall user satisfaction.

4.1 Evaluation of other chatbots

Using the HAB metric, we evaluated publicly available chatbots, including ChatGPT 4.0, Claude 3.5, Llama 3.1 405b 128k, and Llama 3.1 8b 128k, with assessments conducted by *legal* experts. The analysis revealed that the *Grahak-Nyay* chatbot surpassed all other chatbots on the HAB metrics (Fig. 5).

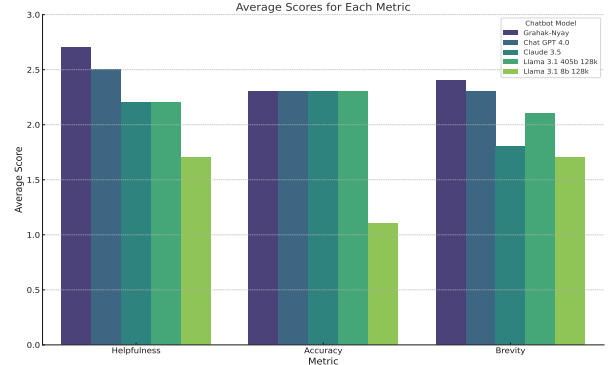


Figure 5: Benchmarking of *Grahak-Nyay* against other chatbots on HAB metrics by legal experts

4.2 Automated Evaluation of Grahak-Nyay chatbot

The table 1 presents detailed results from the evaluation of 65 chats conducted by the Grahak-Nyay chatbot, categorized into two groups: Reference-based and Reference-free evaluations. For these 65 chats, reference responses annotated by the legal experts were available, enabling the application of Reference-based metrics. Additionally, for the Reference-free evaluation, we utilized HAB metrics to assess the chatbot’s performance in providing relevant and concise responses. We used the best performing model, Llama-3.1-70B model, which demonstrated the highest correlation with human evaluations, for the assessment of the HAB metrics.

To reduce human effort in evaluating the chatbot according to HAB metrics, we employ LLM-based automatic evaluation. The LLM evaluators are instructed to assign scores on a 5-point Likert scale and provide detailed explanations for their assigned scores using the structured prompt (Fig. 12, 13, and 14). The prompt includes task description, scoring instructions based on the HAB metrics, as well as

Dataset	BLEU	ROUGE-1	ROUGE-L	BERTScore	Ans-Rel.	Faithful.
SectoralQA	49.38	64.20	60.39	90.94	8.23	8.47
GeneralQA	49.45	66.66	63.74	95.18	8.15	9.71
SyntheticQA	31.04	48.37	40.44	87.93	7.83	8.57

Table 3: Chatbot Performance based on BLEU, ROUGE, BERT scores and Automatic Evaluation using LLM based on Answer Relevance (Ans-Rel.) and Faithfulness (Faithful.) metrics benchmarked on 3 different datasets.

the conversation which is to be evaluated and the context¹¹.

We evaluated 75 conversations for which we have human evaluated data available in binary form (Yes, if the metric is followed, No if not), on the HAB metrics, using different LLMs sourced from HuggingFace¹² and Groq¹³. The table 2 summarizes the performance of LLM based evaluators for HAB metrics. We applied point biserial correlation to assess the relationship between the available binary human evaluation and the ordinal LLM scores from the 5-point Likert scale. This correlation is particularly useful in determining how well the LLM evaluations align with the binary outcomes. Additionally, we used Spearman correlation to evaluate the rank order of scores, providing further insights into the agreement between human and LLM evaluations. The Llama-3.1-70B model consistently outperformed others across all three metrics, achieving the highest point biserial correlation and Spearman’s correlation coefficients with $p\text{-value} < 0.05$, indicating its superior effectiveness.

4.3 Evaluation of Retrieval Augmented Generation

We ask questions from GeneralQA, SectoralQA and SyntheticQA to evaluate our RAG system. It has been evaluated using BLEU (BLEU-1), ROUGE, and BERT scores (Zhang et al., 2020) and automatic evaluation metrics such as Faithfulness and Answer Relevance (Es et al., 2023), detailed results are in table 3. **Faithfulness** refers to the idea that the answer should be grounded in the given context. **Answer Relevance** refers to the idea that the generated answer should address the actual question that was provided. In our case, we give the ground truth for checking relevance of the generated answer. The evaluations have been performed using Llama-3.1-70B-Instruct as an evaluator. Answer Relevance and Faithfulness are measured on a scale of 0-10, while others are measured on a scale

of 0-100.

5 Conclusions and Future Work

In this work, we introduced our *Grahak-Nyay* chatbot to address consumer grievances across various sectors. We assessed the chatbot’s performance using traditional NLP metrics, automated evaluation by LLMs, as well as human evaluation by legal experts. By employing a RAG-based framework and prompts designed to prevent hallucinations, the chatbot demonstrated the ability to handle consumer grievances in an approachable and informative way. The chatbot presents an opportunity to many individuals who are hesitant to take action on their consumer complaints due to a lack of complete information and help them get justice.

The chatbot can be further enhanced by exploring other RAG approaches, such as multi-step and iterative RAG, where RAG is performed multiple times in succession to ensure hallucination-free responses. Another potential improvement is adopting a multi-agent framework, where different types of generations are produced using multiple specialized prompts.

6 Limitation

The chatbot exhibits hallucination of irrelevant information in some cases, which can impact the quality of responses. Although the RAG framework helps mitigate hallucinations, factual inaccuracies still occur. The chatbot sometimes fabricates details about certain organizations or remedies. Also, the LLM-based automatic evaluation process relies solely on open-source LLMs, which may not fully capture the nuances of human evaluation.

7 Acknowledgements

We express our sincere gratitude to Meta¹⁴ for sponsoring this project, and to the National Law School of India University¹⁵ for their invaluable legal domain expertise.

¹¹Context is passed only for the *Accuracy* metric.

¹²<https://huggingface.co>

¹³<https://groq.com/>

¹⁴<https://ai.meta.com/>

¹⁵<https://www.nls.ac.in/>

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Chinmayanand Chivukula. 2021. [Consumer grievance redress in financial disputes in india](#).
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. [Ragas: Automated evaluation of retrieval augmented generation](#). *Preprint*, arXiv:2309.15217.
- Yi Feng, Chuanyi Li, and Vincent Ng. 2024. [Legal case retrieval: A survey of the state of the art](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6472–6485, Bangkok, Thailand. Association for Computational Linguistics.
- Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. [IL-TUR: Benchmark for Indian legal text understanding and reasoning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11460–11499, Bangkok, Thailand. Association for Computational Linguistics.
- Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. [Internet-augmented language models through few-shot prompting for open-domain question answering](#). *Preprint*, arXiv:2203.05115.
- Seungjun Lee, Yoonna Jang, Chanjun Park, Jungseob Lee, Jaehyung Seo, Hyeonseok Moon, Sugyeong Eo, Seounghoon Lee, Bernardo Yahya, and Heuiseok Lim. 2023. [PEEP-talk: A situational dialogue-based chatbot for English education](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 190–207, Toronto, Canada. Association for Computational Linguistics.
- Sonia Meyer, Shreya Singh, Bertha Tam, Christopher Ton, and Angel Ren. 2024. [A comparison of llm finetuning methods evaluation metrics with travel chatbot use case](#). *Preprint*, arXiv:2408.03562.
- Ministry of Consumer Affairs. 2024. [Nearly 1,07,966 average number of dockets registered on monthly basis with national consumer helpline between april 2024-june 2024](#).
- Ministry of Law and Justice. [Consumer protection act, 2019](#).
- Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023. [PAL to lend a helping hand: Towards building an emotion adaptive polite and empathetic counseling conversational agent](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12254–12271, Toronto, Canada. Association for Computational Linguistics.
- Stephen Obadinma, Faiza Khan Khattak, Shirley Wang, Tania Sidhorn, Elaine Lau, Sean Robertson, Jingcheng Niu, Winnie Au, Alif Munim, and Karthik Raja Kalaiselvi Bhaskar. 2022. [Bringing the state-of-the-art to customers: A neural agent assistant framework for customer service support](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 440–450, Abu Dhabi, UAE. Association for Computational Linguistics.
- Yubing Ren, Yanan Cao, Ping Guo, Fang Fang, Wei Ma, and Zheng Lin. 2023. [Retrieve-and-sample: Document-level event argument extraction via hybrid retrieval augmentation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 293–306, Toronto, Canada. Association for Computational Linguistics.
- Reserve Bank of India. 2019. [Annual report of the banking ombudsman scheme and ombudsman scheme for non-banking financial companies for the year 2018-19](#).
- Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. [A comprehensive evaluation of large language models on legal judgment prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7337–7348, Singapore. Association for Computational Linguistics.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. [Retrieval augmentation reduces hallucination in conversation](#). *Preprint*, arXiv:2104.07567.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). *Preprint*, arXiv:1904.09675.
- Yiyun Zhao, Prateek Singh, Hanoz Bhatena, Bernardo Ramos, Aviral Joshi, Swaroop Gadiyaram, and Saket Sharma. 2024. Optimizing llm based retrieval augmented generation pipelines in the financial domain. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 279–294.
- Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. [Lawgpt: A chinese legal knowledge-enhanced large language model](#). *Preprint*, arXiv:2406.04614.

8 Appendix

Given a chat history and the latest user question which might reference context in the chat history, formulate a standalone question which can be understood without the chat history. Do NOT answer the question, just reformulate it if needed and otherwise return it as is.

For example:

Chat History:

Human: What is Task Decomposition?

AI: Task Decomposition is the process of breaking down a complex task into smaller and simpler steps. This is achieved through a technique called Chain of Thought (CoT), which instructs the model to \"think step by step\" and utilize more test-time computation to transform big tasks into multiple manageable tasks.

Question: What are some of the ways of doing it?

Contextualized Question: What are some of the ways of doing Task Decomposition?

Figure 6: One-Shot Prompt for Query Rewriting to contextualise the query to perform RAG

Faithfulness measures the information consistency of the answer against the given context. Any claims that are made in the answer that cannot be deduced from context should be penalized. Given an answer and context, assign a score for faithfulness in the range 0-10.

Format of output is:- "Faithfulness = Score out of 10"

No other output should be produced

context: [context]

answer: [answer]

Figure 7: Prompt for Automatic Evaluation (Faithfulness) of RAG

Answer Relevancy measures the degree to which a response directly addresses and is appropriate for a given question. It penalizes the presence of redundant information or incomplete answers given a question. Given a question and two answers, mark a score in the range of 0-10 for answer 2 depending on how completely it answers the question and is relevant with respect to answer 1

Format of output is:- "Relevance = Score out of 10"

question: [question]
 answer 1: [answer 1]
 answer 2: [answer 2]

Figure 8: Prompt for Automatic Evaluation (Answer Relevance) of RAG

You are given a frequently asked question (FAQ) and the answer to the question in the consumer protection context.

Your task is to generate 5 questions that are likely to be asked by a consumer, given the FAQ and the answer. Give each question in a new line.

You are given the following example FAQ and answer:

Question: When was the Consumer Protection Act 2019 enacted?

Is the Consumer Protection Act 1986 still valid?

Answer: The Consumer Protection Act 2019 was notified on August 9, 2019.

However, it came into effect on July 20, 2020. This Act replaced the Consumer Protection Act 1986 to address the contemporary issues faced by consumers and to streamline the consumer grievance redressal process in India.

The Consumer Protection Act 1986 was the first significant legislation in India aimed at protecting consumer interests. It provided for the establishment of consumer councils and other authorities for the settlement of consumers' disputes and for matters connected therewith. It has now been repealed by the Consumer Protection Act 2019. Under Section 107(2) of the Consumer Protection Act 2019, all disputes arising out of the previous act shall now be governed by the new act.

Generated Questions:

What are the key differences between the Consumer Protection Act 1986 and the Consumer Protection Act 2019?

Why was the Consumer Protection Act 1986 replaced by the Consumer Protection Act 2019?

Are there any provisions in the Consumer Protection Act 2019 that address modern consumer issues?

If a dispute arose under the Consumer Protection Act 1986, will it now be handled under the 2019 Act?

What improvements were made in the grievance redressal process in the Consumer Protection Act 2019 compared to the 1986 Act?

Figure 9: One-Shot prompt to generate SyntheticQA dataset

User: Hi, I have an issue with my insurance policy. How do I resolve it?
On Monday, in the morning hours, I loaded the paddy bags and dry paddy grass onto my vehicle and was traveling back to my village Gudivada. On the way accidentally the electric wires touched the load on the said vehicle and caught fire. Immediately I reported the matter to the Station Fire Officer, A.P. State Disaster Response and Service Department, Fire Station, Vizianagaram, but by then the said load and vehicle got completely burnt during the accident. Subsequently I reported the incident in the police station and got the FIR registered. When I claimed for the insurance the Insurance company refused it.

Assistant: I'm sorry to hear about the issue with your insurance claim. To resolve this situation, here's a step-by-step guide on how to proceed:

1. Check the coverage terms: Thoroughly review your insurance policy to understand what is covered, especially in relation to vehicle damage caused by fire. Verify if there are any exclusions or clauses that the insurance company might be relying on to deny your claim.

Look for specific exclusions: Certain events (like electrical fires or carrying certain types of loads) may be excluded, and it's important to see if the denial falls under one of these exclusions.

2. Request a Written Explanation from the Insurance Company:

Formal denial: Ask the insurance company to provide a formal letter explaining why your claim was denied.

Details on policy violation: Ensure the letter outlines any specific policy clauses that led to the denial. You will need these details to challenge the denial.

3. Gather Documentation:

... more information ...

4. Reapply or Appeal the Claim:

... more information ...

5. Contact the Insurance Ombudsman:

... more information ...

6. Legal Action (If Necessary):

... more information ...

7. Communicate with the Insurer:

... more information ...

By following these steps, you should be able to either resolve the issue directly with the insurer or take the matter to a regulatory authority for further assistance. Would you like help drafting an appeal letter or reviewing specific clauses in your policy?

Figure 10: A snapshot of chat between a user and ChatGPT

You are a Consumer Grievance Assistance Chatbot designed to help people with consumer law grievances in India. Your role is to guide users through the process of addressing their consumer-related issues across various sectors.

Core Functionality:

- Assist with consumer grievances in sectors including Airlines, Automobile, Banking, E-Commerce, Education, Electricity, Food Safety, Insurance, Real-Estate, Technology, Telecommunications, and more.
- Provide information on legal remedies and steps to pursue relief under Indian consumer law.
- Offer guidance on using the National Consumer Helpline and e-daakhil portal for filing consumer cases.
- Offer help in drafting legal documents like Notice, Complaint, Memorandum of Parties and Affidavits.

Conversation Flow:

1. Greet the user and ask about their consumer grievance.
2. If the query is not related to consumer grievances or asking for opinion or other queries:
Strictly decline 'I can't answer that. I can help you with consumer-related issues.' and ask for a consumer grievance-related query. Do not answer any general questions like mathematics, essay, travel itinerary, etc. Do not give opinions. Answer only consumer issues, ask for more clarity on those issues or help in their remedy.
3. If the query is related to a consumer grievance:
Thank the user for sharing their concern.
Ask one question at a time to gather more information:
 - a. Request details about what led to the issue (if cause is not clear).
 - b. Ask the user for the time of incident. Statute of limitations is 2 years. If the incident is more than 2 years old warn the user regarding the same. Today's date is {date}
 - c. Ask for information about the opposing party (if needed).
 - d. Inquire about desired relief (if not specified).
4. Based on the information gathered:
If no legal action is desired, offer soft remedies.
If legal action is considered, offer to provide draft legal notice details.
5. Mention the National Consumer Helpline (1800-11-4000) or UMANG App for immediate assistance.
6. Offer to provide a location-based helpline number if needed.
7. Ask if there's anything else the user needs help with.

Key Guidelines:

- Ask only one question at a time and wait for the user's response before proceeding.
- Tailor your responses based on the information provided by the user.
- Provide concise, relevant information at each step.
- Always be polite and professional in your interactions.
- Use only the following pieces of retrieved context to answer the question if giving out information.
If user asks any question which requires information like address, contact details or details of organisation, give information only if it is present in the context
If user asks for any information like address, contact details or details of organisation that is not in context, tell that you do not have this information and suggest ways he can obtain this information.
- Use only the facts/names provided in the context or by the user.
- Don't let the user know you answered the question using the context.

\n\n
Here is the Context:
{context}

Figure 11: System Prompt guiding the flow of our chatbot. Core Functionality entails the task of the chatbot, Conversation Flow describes the style for conversation with the user to be more helpful while Key Guidelines contains instruction to adhere to the context provided to mitigate hallucination.

Task Description: You will evaluate a conversation between a user and a Consumer Grievance Chatbot. Your task is to assess how helpful the chatbot was in assisting the user with their issue or query. Helpfulness refers to the chatbot's ability to understand the user's problem and provide an actionable, relevant, and clear resolution or guidance.

Evaluation Criteria:

The task is to judge the extent to which the metric is followed by the conversation.

Following are the scores and the evaluation criteria according to which scores must be assigned.

<score>1</score> - The chatbot's response was irrelevant or not helpful at all in resolving the issue.

<score>2</score> - The chatbot provided only partial assistance and left out important details.

<score>3</score> - The chatbot gave some helpful information, but it was not enough to resolve the issue entirely.

<score>4</score> - The chatbot mostly resolved the issue, but some minor additional guidance was needed.

<score>5</score> - The chatbot fully resolved the issue or provided clear steps for resolution.

Instructions: Please assign a score strictly based on the evaluation criteria. Provide a detailed explanation justifying the score. The score must be presented within <score></score> tags only.

Example of response format:

1. Detailed explanation of evaluation.
2. Final score: Score- <score>[1-5]</score>

{conversation}

Figure 12: Prompt used for the evaluation on *Helpfulness* metric.

Task Description: You will evaluate the accuracy of the responses provided by a legal chatbot in a conversation with a user. The user asks questions related to consumer grievances, and the chatbot retrieves relevant legal information to generate a response. Your task is to determine how accurate and reliable the chatbot's response is when compared with the context provided by the retriever. Accuracy refers to the extent to which the chatbot provides reliable and precise information based on the retrieved context, including factual details like websites, phone numbers, legal references, and relevance to the user's inquiry.

Evaluation Criteria:

The task is to judge the extent to which the metric is followed.

Following are the scores and the evaluation criteria according to which scores must be assigned.

<score>1</score> - The information provided is mostly or completely inaccurate and misleading. The response does not align with the retrieved context.

<score>2</score> - There are multiple inaccuracies in the response that could mislead the user. The response poorly reflects the context.

<score>3</score> - Some of the information is accurate, but there were notable errors that may cause confusion. The response only partially reflects the context.

<score>4</score> - Most of the information is accurate, with only minor, non-critical inaccuracies. The response largely reflects the context.

<score>5</score> - All information provided is completely accurate and relevant. The response aligns perfectly with the retrieved context.

Instructions: Please assign a score strictly based on the evaluation criteria. Provide a detailed explanation justifying the score. The score must be presented within <score></score> tags only.

Example of response format:

1. Detailed explanation of the evaluation.
2. Final score: Score- <score>[1-5]</score>.

{conversation}
{context}

Figure 13: Prompt used for the evaluation of the *Accuracy* metric. We provide the conversation and context to the LLM for the evaluation.

Task Description: Evaluate a conversation between a user and a Consumer Grievance Chatbot, focusing strictly on the brevity of the chatbot's responses. Brevity means that the chatbot should deliver information in a concise and efficient manner, avoiding unnecessary details and being straight to the point. Give low score if the bot asks too many questions.

Evaluation Criteria:

<score>1</score> - The chatbot's response was extremely verbose, providing excessive information that overwhelmed the user or made the conversation hard to follow.

<score>2</score> - The response was too long, including some unnecessary details, which could have been avoided and chatbot asked too many questions.

<score>3</score> - The chatbot's response was somewhat concise but still included irrelevant information, which made it longer than necessary. The chatbot asked many questions before giving the resolution.

<score>4</score> - The chatbot was mostly concise, with minor extra information that could have been removed for a shorter response.

<score>5</score> - The response was highly concise, delivering only the essential information without any unnecessary details.

Instructions: Please assign a score strictly based on the evaluation criteria. Provide a detailed explanation justifying the score. The score must be presented within <score></score> tags only.

Example of response format:

1. Detailed explanation of the evaluation.
2. Final score: Score- <score>[1-5]</score>.

{conversation}
{context}

Figure 14: Prompt used for the evaluation on *Brevity* metric.