LLM Based Efficient CSR Summarization using Structured Fact Extraction and Feedback

Kunwar Zaid, Amit Sangroya, Mayur Patidar, Lovekesh Vig
TCS Research and Innovation
New Delhi, India
{kunwar.zaid,amit.sangroya,lovekesh.vig}@tcs.com

Abstract

Summarizing clinical trial data poses a significant challenge due to the structured, voluminous, and domain-specific nature of clinical tables. While large language models (LLMs) such as ChatGPT, Llama, and DeepSeek demonstrate potential in table-to-text generation, they struggle with raw clinical tables that exceed context length, leading to incomplete, inconsistent, or imprecise summaries. These challenges stem from the structured nature of clinical tables, complex study designs, and the necessity for precise medical terminology. To address these limitations, we propose an end-to-end pipeline that enhances the summarization process by integrating fact selection, ensuring that only the most relevant data points are extracted for summary generation. Our approach also incorporates a feedback-driven refinement mechanism, allowing for iterative improvements based on domain-specific requirements and external expert input. By systematically filtering critical information and refining outputs, our method enhances the accuracy, completeness, and clinical reliability of generated summaries while reducing irrelevant or misleading content. This pipeline significantly improves the usability of LLM-generated summaries for medical professionals, regulators, and researchers, facilitating more efficient interpretation of clinical trial results. Our findings suggest that targeted preprocessing and iterative refinement strategies within the proposed pipeline can mitigate LLM limitations, offering a scalable solution for summarizing complex clinical trial tables.

1 Introduction

The growing scale of medical research, reflected in thousands of clinical trials conducted globally each year, has resulted in a vast amount of tabular data that requires effective interpretation. Clinical trial tables, which summarize key aspects such as patient demographics, treatment arms, and outcomes, play a critical role in the evaluation of medical interventions. However, these tables are often complex and dense, containing a mixture of statistical information and clinical findings that are not easily digestible without significant time and expertise. This creates a bottleneck in the dissemination and practical application of clinical findings, as stakeholders-ranging from healthcare professionals to policy makers struggle to extract meaningful insights quickly and accurately from trial reports.

The recent advances in natural language processing, particularly with large language models (LLMs) like ChatGPT, have unlocked new opportunities for automating the conversion of structured data into readable and informative summaries. LLMs have shown significant potential in table-to-text generation tasks (Hegselmann et al., 2023), where they can summarize data tables into coherent narratives by identifying key patterns and relationships. In fields such as business analytics (Nasseri et al., 2023), (Jiang et al., 2024), (Teubner et al., 2023) and scientific reporting (Telenti et al., 2024), (Sallam, 2023), LLMs have demonstrated their utility in transforming structured datasets into succinct summaries (Chen, 2022). However, when applied to the highly specialized domain of clinical trial data, these models face substantial limitations.

Clinical trial tables are often vast and intricately detailed, encompassing a wide array of variables such as multiple treatment arms, efficacy measures, adverse events, and participant characteristics. The complexity and scale of these tables overwhelm current LLM capabilities, leading to incomplete or overly generalized summaries when the tables are provided as direct input. Moreover, clinical data requires precision, as even minor inaccuracies in summarization can have significant implications for patient safety and medical decision making. The inherent challenge lies in ensuring that the generated summaries retain both the accuracy and

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 148–157

the contextual relevance of the underlying data, a requirement that LLMs struggle to meet without intervention.

To address these limitations, we propose an end-to-end pipeline designed to improve the summarization of clinical trial tables using LLMs. This pipeline incorporates a fact selection mechanism that preprocesses the tables by extracting the most relevant data points, ensuring that the input to the LLMs is both concise and focused. The pipeline further integrates a feedback loop, allowing users to refine and improve the generated summaries iteratively. This approach not only enhances the quality and reliability of the summaries but also offers flexibility, enabling the adaptation of summaries based on specific user requirements.

2 Related Work

Clinical Study Reports (CSRs) provide a detailed account of a clinical study's design, methodology, and outcomes, serving as crucial documents for regulatory approval, labeling, and commercialization. Unlike academic papers, CSRs offer a comprehensive, data-driven evaluation of a drug's therapeutic effectiveness. Earlier approaches to summary generation using tabular data devise complex template schemes in collaboration with domain experts to build a consistent set of data-to-word rules (Bao et al., 2018), (Chen et al., 2019a), (Chen et al., 2019b). This has been used in domains such as weather and medical report generation (Deng et al., 2013; Reiter et al., 2005; Varges et al., 2012). These works relied heavily on expert knowledge to bring out semantics from structured-data.

Most of the modern techniques for *Table-to-Text* summary generation can be divided into two independent components: (1) content selection: involves choosing a subset of relevant records in a table to include in the summary. (2) generating natural language descriptions for this subset. Multiple approaches have been proposed for the individual modules. For content selection, the approach by (Barzilay and Lapata, 2005) builds a content selection model by aligning records and sentences. Summary generation is often treated as a surface realization problem where text is generated from a given concept representation.

Authors in (Lebret et al., 2016), (Wiseman et al., 2017) have approached the table-to-text problem by formulating the input table as a sequence of records. They have developed table-to-text methods using the Seq2Seq framework, and in the process, they explored the modeling of table representation, as studied by (Geng et al., 2018) and (Gong et al., 2019) in their respective works. In the paper by (Li et al., 2023), a non-autoregressive model for table-to-text generation is introduced, named "Plan-then-Seam" (PTS). This model is designed to generate outputs in parallel through a single network.

The PTS approach consists of two distinct steps that are executed iteratively while sharing parameters. In the first step, the model creates and refines a content plan for the generated output. In the second step, the model uses this content plan as context to decode the description. In the work presented by (Gong et al., 2020), a method called TableGPT is introduced for table-to-text generation. The approach involves a multi-step process aimed at enhancing the alignment between structured tables and their corresponding natural language summaries.

The incorporation of auxiliary tasks to enhance the table representation is another paradigm for tackling the table-to-text problem, as demonstrated in the works of (Tian et al., 2019), (Li et al., 2021). In (Chen et al., 2023) have proposed an approach for table-to-text generation with a pre-trained language model. In the paper by (Lin et al., 2023) the authors introduce the "Inner Table Retriever,", a general-purpose approach to address the challenge of handling large tables in TableQA (Table Question Answering). This method involves extracting sub-tables from the original large table to retain the most pertinent and relevant information specifically related to a given question.

In the study conducted by (Gao et al., 2023) the authors investigate ChatGPT's capacity to perform human-like summarization evaluation. They assess the model's summarization outputs and compare them against commonly used automatic evaluation metrics. The findings reveal that ChatGPT exhibits superior performance compared to these conventional metrics, suggesting that it is capable of producing summaries that align more closely with human-like quality and judgment.

3 Approach

Traditionally, medical writing experts transform complex clinical data into structured narratives that meet regulatory requirements. However, advancements in AI-driven solutions are reshaping this process. Generative AI models can now interpret in-

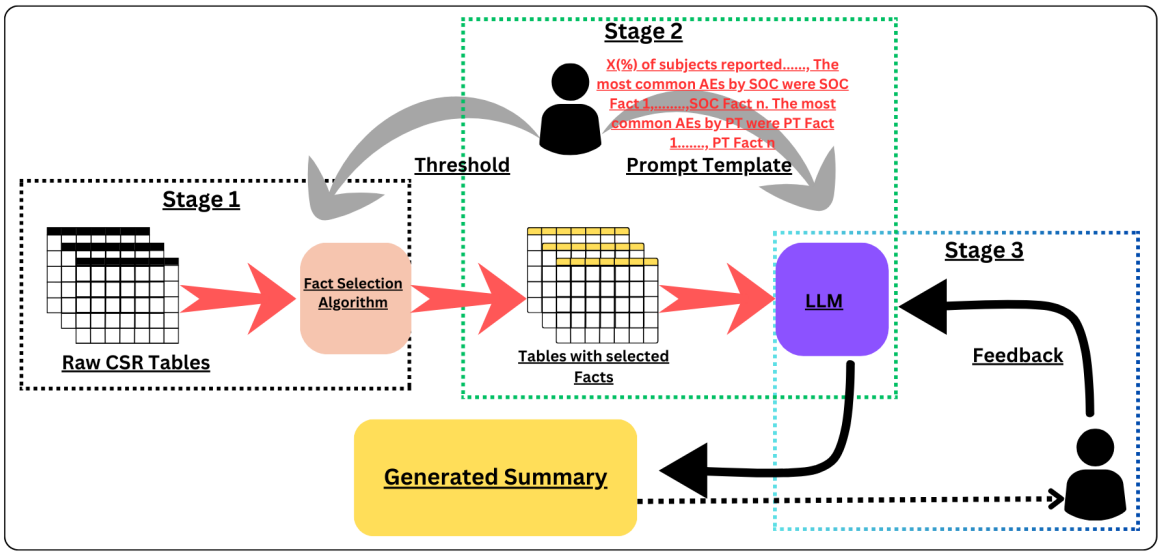

Figure 1: Overall Architecture of CSR Summary Generation

tricate CSR tables and produce reliable summaries. Our approach focuses on handling large and complex tables that existing table-to-text summarization methods struggle to process (see Figure 2).

3.1 Task Description

Given a clinical trial table (See example Table 1), the objective is to generate a concise and informative summary that captures all the factual information depicted in the table while avoiding hallucinations. The task can be broken down into the following key steps:

- **Table Linearization:** Convert the table into a linearized structure that is easy for an LLM to interpret. The linearized format is represented as: $|Cell_1|Cell_1|Cell_2|.....|Cell_n|$.
- **Input Preprocessing and Strategy Selection:** Depending on the size of the table and its compatibility with the model's input capacity, different strategies are employed to generate summaries. These include:
  - Zero-Shot Techniques: Directly prompting the LLM to summarize the linearized table without prior examples
  - Few-Shot Techniques: Providing the LLM with curated examples of correctly formatted summaries to guide its output.
  - Selection Algorithms: Applying algorithms to filter and prioritize the most relevant data points from, ensuring that the input to the LLM is both concise and contextually significant
- **Summary Generation:** Using the processed input, the LLM generates a summary that encompasses all relevant factual information while maintaining contextual coherence and precision.
- **User Feedback Integration:** Incorporate user feedback to refine and improve the generated summaries iteratively, ensuring alignment with specific use cases and requirements

3.2 Automatic Assessment of CSR Tables

A significant challenge in working with large and complex tables is their size. Most tables are very large, often exceeding the context length limitations of large language models (LLMs). The complexity is further compounded by hierarchical relationships between system organ classes (SOCs), and preferred terms (PTs), missing data, and the need to ensure accuracy and completeness when summarizing. Addressing these challenges required innovative strategies to preprocess and structure the data for effective summarization without losing critical information.

3.2.1 Handling Large Tables

The novelty of this study lies in the approach to handling large clinical trial tables. To ensure that no critical information is missed while fitting the data within the model's context length, we explored multiple approaches:

Primary system organ class	Preferred term MedDRA version 19.0	NA	Drug A DPI 28 on/off N=171 (100%)	Drug A DPI 14 on/off N=174 (100%)	Placebo 28 on/off N=86 (100%)	Placebo 14 on/off N=88 (100%)	Pooled Placebo N=174 (100%)	Total N=519 (100%)
Number (%) of subjects with at least one such adverse event		Number (%) of subjects with at least one such adverse event	94 (55.0%)	127 (73.0%)	63 (73.3%)	53 (60.2%)	116 (66.7%)	337 (64.9%)
Blood and lymphatic system disorders		Blood and lymphatic system disorders	2 (1.2%)	4 (2.3%)	1 (1.2%)	2 (2.3%)	3 (1.7%)	9 (1.7%)
Blood and lymphatic system disorders	Anaemia	Anaemia	2 (1.2%)	2 (1.1%)	0	0	0	4 (0.8%)
Blood and lymphatic system disorders	Coagulopathy	Coagulopathy	0	1 (0.6%)	0	0	0	1 (0.2%)

Figure 2: Example of a Clinical Trial Table

- **Dividing the Table into Smaller Chunks:** Large tables were segmented into smaller, logically coherent sections based on SOCs or study arms. However, this approach often led to a loss of context and missed critical cross-segment information.

- **Mean-Based Thresholding:** This method involved calculating the mean of the data values as a threshold for selecting facts from the tables. While this approach simplified the selection process, it did not consistently capture the most clinically relevant data points, particularly in cases where data distributions were highly skewed. Mean SOC and PT is defined as:

$$\mu_{SOC} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

where μ_{SOC} is the SOC threshold, x_i are the SOC values, and N is the total number of SOCs.

$$\mu_{PT} = \frac{1}{M} \sum_{i=1}^{M} y_i$$

where μ_{PT} is the PT threshold, y_i are the PT values, and M is the total number of PTs.

- **Percentile-Based Thresholding:** Ultimately, we adopted a percentile-based thresholding method, which proved most effective. By selecting data points based on predefined percentiles, this approach ensured that significant facts were consistently included while maintaining a manageable context length for the model. For the p-th percentile, where p is the desired percentile (e.g., 90 for the 90th percentile), threshold T_p is defined as:

$$T_p = x_{(\lceil \frac{p}{100} \cdot n \rceil)} + \left(\frac{p}{100} \cdot n - \left\lceil \frac{p}{100} \cdot n \right\rceil \right) \cdot \left(x_{(\lceil \frac{p}{100} \cdot n \rceil + 1)} - x_{(\lceil \frac{p}{100} \cdot n \rceil)} \right)$$

where:

  - T_p is the threshold corresponding to the p-th percentile,
  - $x_1, x_2, \ldots, x_n$ are the data points sorted in ascending order,
  - n is the number of data points,
  - p is the desired percentile (e.g., $p = 90$ for the 90th percentile),
  - $\lceil \cdot \rceil$ denotes the ceiling function.

Using above formula, threshold can be calculated for SOCs and PTs, based on desired percentile.

3.3 Automatic Extraction of Important Facts

Our fact selection algorithm aims to extract the most important facts from large and complex CSR tables. One example for extracting facts from adverse events table is shown in Algorithm 1. The fact selection algorithm plays a crucial role in our pipeline. It is capable of handling very large tables that usually fail to fit within the input constraints of the LLMs.It is designed to extract the most pertinent facts from the tables, significantly reducing the size of large tables. The algorithm handles all the table types, regardless of their complexity or size. It determines the threshold based on the percentile and selects the relevant facts accordingly. By focusing on relevant facts, the algorithm enhances both the efficiency and reliability of the summarization process.

Algorithm 1: Fact selection Algorithm for Adverse Events

```
For each table type T = 1,2,3,....N
  while T < N do
    For each table t = 1,2,3,....M
      while t < M do
        Identify SOCs and PTs ;
        Remove the empty values;
        Extract SOC and PT values;
        Apply percentile-based thresholding;

        Select the SOCs and PTs using
          threshold;
        Reconstruct the table using selected
          SOCs and PTs;
    end
  end
```

4 Experiments

4.1 Dataset

We could not find any publicly available datasets for this specific task, nor could we identify prior work that addresses the summarization of clinical trial tables using LLMs. While some clinical trial reports are available on public portals (NLM), (GSK) the data they provide is limited. The clinical trial tables used in this study are proprietary data from a large pharmaceutical company. Due to confidentiality agreements, the name of the company and the dataset cannot be disclosed. Table 1 summarizes the number and types of tables used in the generation process. The table types are described as follows:

- ***Subject Disposition:*** Provides a summary of the participants included in each analysis group and the reasons for any exclusions.
- ***Subject Demography:*** Displays demographic and other relevant baseline characteristics of study participants, either categorized or by descriptive statistics.
- ***Medical History:*** Presents a summary of participants' medical history, ordered by frequency of occurrence.
- ***Overall Summary:*** Summarizes adverse events (AEs) across various categories.
- ***AEs by SOC and PT:*** Lists AEs by treatment group, categorized by system organ class (SOC, highest level) and preferred term (PT, second-highest level), ordered by frequency.
- ***AEs by Maximum Intensity:*** Categorizes AEs by treatment group, based on the maximum intensity of each event, in descending order of frequency.
- ***AEs by Worst Outcome:*** Categorizes AEs by treatment group, with classification based on the worst outcome, and further categorized by SOC and PT.
- ***AEs by Common % or more by SOC and PT:*** Lists AEs that exceed a predefined frequency threshold, organized by SOC and PT.

These tables present structured data on adverse events (AEs), system organ classes (SOCs), and preferred terms (PTs), along with numeric summaries like incidence rates and percentages for each study arm. The size of the tables varies, with some large enough to exceed the context length of large language models (LLMs). For example, the tables for Medical History, AEs by SOC/PT, AEs by Maximum Intensity, and AEs by Worst Outcome are especially large.

Table 1: CSR Table Types

Table Types	Number of Tables
Subject Disposition	14
Subject Demography	16
Medical History	14
Overall Summary	13
AEs by SOC and PT	7
AEs by Maximum intensity	7
AEs by Worst Outcome	4
AEs by Common % or more by SOC and PT	4

4.2 Experimental Setup

We conducted experiments with the following models:

- **GPT-4o-mini:** A state-of-the-art model known for its robust summarization capabilities and large token limit (Achiam et al., 2023).
- **DeepSeek** (Chat window): DeepSeek is a Chinese artificial intelligence company that develops open-source large language models (LLMs) (Liu et al., 2024). We used the latest advanced language model comprising 671 billion parameters.

- **Llama 3.1 70B Instruct:** An open source model fine-tuned for instruction following task (HuggingFace, a).

- **Nous Hermes 2 Mixtral 8x7B DPO:** A model further fine-tuned on Mixtral 8x7B MOE with reinforcement learning via direct preference optimization (DPO), also featuring a 32k token limit (HuggingFace, b).

Due to cost constraints, we could not do experiments with some of the latest LLMs with higher capabilities. However, a variety of architectural and assessment capabilities are offered by the chosen models. Some clinical trial tables in our dataset exceeded the context length of the largest models tested such as GPT-4o-mini, due to which models were unable to process the entire table leading to incomplete outputs. This limitation further highlights the importance of fact selection algorithm for handling large tables effectively.

4.3 Quantitative Evaluation

To evaluate the quality of generated summaries, we used the following metrics.

4.3.1 Claim Recall and Claim Precision

- This framework, introduced by (Xie et al., 2024) as *DOCLENS: Multi-aspect Fine-grained Evaluation for Medical Text Generation*, is specifically tailored to assess medical text generation tasks.

- ***Claim Recall:*** This metric evaluates the completeness of the generated text. The reference summary is segmented into individual sub-claims or facts using GPT-4, with each sub-claim representing a single fact. The generated text is then analyzed by an evaluator model to determine whether it entails each sub-claim from the reference summary.

- ***Claim Precision:*** This metric evaluates the conciseness of the generated text. The generated summary is divided into sub-claims. The reference summary is then analyzed to determine if it entails each sub-claim from the generated summary.

We utilized GPT-4o to create the sub-claims for both the reference text and the generated text. Additionally, we employed the same model as an evaluator.

4.4 Human Evaluation

For human evaluation, we sought assistance from our organization's internal medical writers. They devised a set of rules tailored to the evaluation of summaries generated for clinical trial tables. The rules guaranteed a consistent and clinically suitable evaluation of the generated outputs. for example, for adverse events table type (under the safety evaluation section)some rules are:

- *Threshold for SOCs and PTs:* A proper cutoff should be decided for both System Organ Classes (SOCs) and Preferred Terms (PTs). Above this threshold, all SOCs and PTs should be selected and included in the summary to maintain relevance and completeness.

- *Template Adherence:* Summaries should follow a consistent and predefined template, ensuring clarity and alignment with organizational or regulatory standards. For example, as shown in Table 3

Medical writer manually evaluated all the generated summaries to verify that the summaries adhere to the following criteria.

- ***Rule Compliance:*** Whether the summary follows the rules and templates specific to the table type.

- ***Accuracy:*** Ensuring there are no hallucinations, incorrect interpretations, missing data, or data mismatches.

- ***Conciseness:*** Exclusion of irrelevant or redundant details.

- ***Fluency:*** Readability and coherence of the summary.

5 Results and Discussion

The performance of various models was assessed using the metrics outlined in the Evaluation section, including Claim Recall, Claim Precision, and manual evaluation. The detailed results are shown in Table 4 and Table 5, which emphasizes the effectiveness of our fact selection algorithm and the overall quality of the generated summaries.

Among the models tested, without the fact selection algorithm, GPT-4o-mini in a 1-shot setting achieved a claim recall of 0.67 and claim precision

"Reference Summary": In total 37/63 children (58.7%) were reported with at least one TEAE. Most frequently reported primary system organ classes affected by TEAEs were: Gastrointestinal disorders (13/63, 20.6%), general disorders and administration site conditions (11/63, 17.5%), musculoskeletal and connective tissue disorders (8/63, 12.7%), nervous system disorders (9/63, 14.3%) as well as infections and infestations (7/63, 11.1%). Overall, there was a similar distribution of TEAEs across both age cohorts, treatments and formulations, with the exception of menorrhagia, which was more frequently reported in adolescents of the Drug A o.d. tablet group (n = 4) than in the comparator group (n = 0)."
"Claims":
Claim 2: The most frequently reported primary system organ class affected by TEAEs was gastrointestinal disorders (20.6%, 13/63).
Claim 4: Musculoskeletal and connective tissue disorders were reported in 12.7% (8/63) of children.

"GPT-4o-mini(1-shot)":{
"claim 2": "The most frequently reported primary system organ class affected by TEAEs was gastrointestinal disorders (20.6%, 13/63).",
"explanation": "The clinical summary provides that gastrointestinal disorders were the most common system organ class, and the percentage (20.6%) is also included. The specific number of subjects affected (13/63) is also consistent.",
"entailment_prediction": 1
},
{
"claim 4": "Musculoskeletal and connective tissue disorders were reported in 12.7% (8/63) of children.",
"explanation": "The clinical summary does not mention musculoskeletal and connective tissue disorders as a common adverse event. Therefore, this claim cannot be fully entailed.",
"entailment_prediction": 0
}

"Our Approach": {
"claim 2": "The most frequently reported primary system organ class affected by TEAEs was gastrointestinal disorders (20.6%, 13/63).",
"explanation": "The summary lists Gastrointestinal Disorders as 20.6% of patients, but it does not explicitly confirm the number of children affected (13/63). However, the percentage matches, so the claim can be inferred.",
"entailment_prediction": 1
},
{
"claim 4": "Musculoskeletal and connective tissue disorders were reported in 12.7% (8/63) of children.",
"explanation": "The clinical summary mentions Musculoskeletal and Connective Tissue Disorders as 12.7% of patients but does not give the exact number (8/63). However, the percentage supports the claim.",
"entailment_prediction": 1
}

Table 2: Example of Claim Recall evaluation for GPT-4o-mini and Nous-Research. **The summary generated by GPT-4o-mini fails to include an important fact, "Claim 4," resulting in lower claim recall.** In contrast, applying the fact selection algorithm to Nous-Research improves claim recall by ensuring all critical facts are present in the generated summary

"AEs By SOC and PT" : X (%) Number of subjects reported at least one such adverse event.....,. The most common adverse events (AEs) by System Organ Class (SOC) were SOC Fact 1 (in Drug A X% of patients, in Drug B Y%,.....so on), SOC Fact 2 (in Drug A X% of patients, in Drug B Y%,.....so on), and SOC Fact 3 (Z%)....., and SOC Fact n (n % of patients)..... The most common AEs by Preferred Term (PT) were PT Fact 1 (a% of patients), PT Fact 2 (in Drug A X% of patients, in Drug B Y%,.....so on), PT Fact 3 (in Drug A X% of patients, in Drug B Y%,.....so on),.....and PT Fact n (n% of patients).

Table 3: An Example Template for AEs by SOC and PT

Table 4: Comparison of Claim Recall and Precision Across Different Models and Approaches

Model	**Claim Recall**	**Claim Precision**
Nous-Hermes-2-Mixtral-8x7B DPO (with fact selection algorithm) **Our Approach**	**0.72**	**0.44**
GPT 4o-mini (0-shot)	0.58	0.38
GPT 4o-mini (1-shot)	0.67	0.47
DeepSeek (0-shot)	0.5	0.36
DeepSeek (1-shot)	0.55	0.44
Llama-3.1-70B-Instruct (0-shot)	0.18	0.15
Llama-3.1-70B-Instruct (1-shot)	0.22	0.18
Nous-Hermes-2-Mixtral-8x7B DPO (0-shot)	0.27	0.22
Nous-Hermes-2-Mixtral-8x7B DPO (1-shot)	0.23	0.29

of 0.47. DeepSeek performed similarly to GPT-4o-mini, while Llama-3.1-70B-Instruct showed the weakest performance. We tested the fact-selection algorithm with Nous-Hermes-2-Mixtral, which attained the highest claim recall of **0.72**, though its claim precision was **0.44**. Additionally, Table 5 demonstrates that the summary generated using the fact selection algorithm outperformed the proprietary models in terms of informativeness, consistency, fluency, and conciseness. Unfortunately, we could not apply the fact selection algorithm to proprietary models due to API costs. However, the superior performance of the open-source models after applying the algorithm suggests that applying it to the proprietary models would yield even better results.

A medical expert from our internal team evaluated the generated summaries. They observed that the output from open-source models, such as Llama 3.1 Instruct 70B, is not acceptable. These models tend to hallucinate, exhibit data mismatches, and fail to adhere to the correct output template. In contrast, proprietary models like GPT-4o-mini produce significantly better results. While hallucinations are less frequent and the model largely presents accurate information from the tables, it still struggles with maintaining the proper output template and occasionally overlooks key facts. As shown in Table 2, GPT-4o-mini misses an important fact ('claim 4'). However, when a fact-selection algorithm is applied and a well-defined output format is provided, the performance of the LLM improves, producing outputs that closely resemble those of a human writer.

The reason for this improved performance lies in the fact that without a fact selection algorithm, the LLM is tasked with both selecting the relevant facts from the provided table and generating the summary. We observed that LLMs struggle with determining an appropriate threshold based on data trends and applying that threshold for fact selection. In contrast, when the fact selection algorithm is used, the generation task is divided into two distinct steps: first selecting the relevant facts, then generating the summary. With the fact selection algorithm in place, the LLM no longer needs to perform fact selection itself. Instead, the selected facts are provided to the LLM along with the necessary template, making it easier for the model to generate the output by simply filling in the blanks of the template. With this approach, both recall and precision can be improved by adjusting the threshold.

Table 5: Overall Evaluation

Type	Model	Informative	Conciseness	Fluency	Consistency	Score
1-shot	Llama-3.1-Instruct-70B	2.8	1.5	3.5	3.1	2.73
1-shot	Nous-research-Mixtral	3.1	2.2	3.8	3.4	3.13
1-shot	DeepSeek	4.2	3.8	4.6	4.5	4.28
1-shot	GPT-4o-mini	4.4	3.8	4.6	4.5	4.33
Algo	Nous-reseaarch-Mixtral	**4.7**	**4.5**	**4.7**	**4.5**	**4.6**

6 Conclusions and Future Work

In this work, we developed an end-to-end pipeline that automates the generation of clinical table summaries from large complex tables. Complexities may be there because of size, density and domain-specific knowledge, that make it difficult for LLMs to consistently generate accurate and relevant summaries. The proposed pipeline enables the LLMs to produce more concise and accurate summaries. Additionally, we incorporated a feedback mechanism within the pipeline, allowing users to refine the output and improve the quality of summaries iteratively.

7 Limitations

Due to some constraints, we could not perform extensive experiments in diverse domains. Our future work aims to address this by experimenting in other complex domains and at a larger data scale. Moreover, we can also perform a comparison with the latest LLMs, particularly those with larger context windows and improved summarization capabilities.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Junwei Bao, Duyu Tang, Nan Duan, Zhao Yan, Yuanhua Lv, Ming Zhou, and Tiejun Zhao. 2018. Table-to-text: Describing table region with natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Regina Barzilay and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 331–338, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Miao Chen, Xinjiang Lu, Tong Xu, Yanyan Li, Jingbo Zhou, Dejing Dou, and Hui Xiong. 2023. Towards table-to-text generation with pretrained language model: A table structure understanding and text deliberating approach. *arXiv preprint arXiv:2301.02071*.

Wenhu Chen. 2022. Large language models are few (1)-shot table reasoners. *arXiv preprint arXiv:2210.06710*.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019a. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2019b. Few-shot nlg with pre-trained language model. *arXiv preprint arXiv:1904.09521*.

Dong Deng, Yu Jiang, Guoliang Li, Jian Li, and Cong Yu. 2013. Scalable column concept determination for web tables using large knowledge bases. *Proc. VLDB Endow.*, 6(13):1606–1617.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Xinwei Geng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2018. Adaptive multi-pass decoder for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 523–532.

Heng Gong, Xiaocheng Feng, Bing Qin, and Ting Liu. 2019. Table-to-text generation with effective hierarchical encoder on three dimensions (row, column and time). *arXiv preprint arXiv:1909.02304*.

Heng Gong, Yawei Sun, Xiaocheng Feng, Bing Qin, Wei Bi, Xiaojiang Liu, and Ting Liu. 2020. Tablegpt: Few-shot table-to-text generation with table structure reconstruction and content matching. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1978–1988.

GSK. Clinical Study reports. https://www.gskstudyregister.com/en/.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag.

2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.

HuggingFace. a. Llama. https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct.

HuggingFace. b. NousResearch. https://huggingface.co/NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO.

Jie Jiang, Haining Xie, Yu Shen, Zihan Zhang, Meng Lei, Yifeng Zheng, Yide Fang, Chunyou Li, Danqing Huang, Wentao Zhang, et al. 2024. Siriusbi: Building end-to-end business intelligence enhanced by large language models. *arXiv preprint arXiv:2411.06102*.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*.

Liang Li, Ruiying Geng, Chengyang Fang, Bing Li, Can Ma, Binhua Li, and Yongbin Li. 2023. Plan-then-seam: Towards efficient table-to-text generation. *arXiv preprint arXiv:2302.05138*.

Liang Li, Can Ma, Yinliang Yue, and Dayong Hu. 2021. Improving encoder by auxiliary supervision tasks for table-to-text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5979–5989.

Weizhe Lin, Rexhina Blloshmi, Bill Byrne, Adrià de Gispert, and Gonzalo Iglesias. 2023. An inner table retriever for robust table question answering.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Mehran Nasseri, Patrick Brandtner, Robert Zimmermann, Taha Falatouri, Farzaneh Darbanian, and Tobechi Obinwanne. 2023. Applications of large language models (llms) in business analytics–exemplary use cases in data preparation tasks. In *International Conference on Human-Computer Interaction*, pages 182–198. Springer.

NLM. Clinical studies. https://clinicaltrials.gov/.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artif. Intell.*, 167(1–2):137–169.

Malik Sallam. 2023. The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *MedRxiv*, pages 2023–02.

Amalio Telenti, Michael Auli, Brian L Hie, Cyrus Maher, Suchi Saria, and John PA Ioannidis. 2024. Large language models for science and medicine. *European Journal of Clinical Investigation*, 54(6):e14183.

Timm Teubner, Christoph M Flath, Christof Weinhardt, Wil van der Aalst, and Oliver Hinz. 2023. Welcome to the era of chatgpt et al. the prospects of large language models. *Business & Information Systems Engineering*, 65(2):95–101.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv preprint arXiv:1910.08684*.

Sebastian Varges, Heike Bieler, Manfred Stede, Lukas C. Faulstich, Kristin Irsig, and Malik Atalla. 2012. SemScribe: Natural language generation for medical reports. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2674–2681, Istanbul, Turkey. European Language Resources Association (ELRA).

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*.

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. 2024. Doclens: Multi-aspect fine-grained evaluation for medical text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.