Adaptive Agent-Based Personalized Contextual Table Summarization

Hithaishi Surendra hsurendr@asu.edu
Shiven Agarwal sagar147@asu.edu
Poorvi Raddi praddi@asu.edu
Cheng Guo chenggu2@asu.edu
LS Ganesh Rathnam glalgudi@asu.edu

Abstract

This proposal outlines "Adaptive Agent-Based Personalized Contextual Table Summarization." The rapid development of extensive language models has significantly impacted natural language processing and text generation, particularly in the contextual summarization of tabular data. The field presents unexploited opportunities for innovative methods to improve precision and efficiency, notwithstanding existing advancements. This study presents a novel framework designed for tabular data, employing adaptive agent roles and customized summarization layers. The system employs Personalized Multi-Modal Representation to tailor table summarizing according on user profiles and contextual inputs, yielding more pertinent and significant outputs. Additionally, the Novel Summarization Model employs agent-based fine-tuning to adapt to user corrections and dataset variability, thereby enhancing accuracy and customization over time. The team will focus on the Summarization Model.

1 Introduction & Problem Statement

The rapid progress of large-scale language models has transformed multiple fields, particularly natural language processing and text generation. These models have achieved impressive results in areas such as content creation, answering questions, and summarizing information, often exhibiting comprehension skills comparable to humans (Peng et al., 2024)(Matarazzo et al., 2024). One domain where these models have shown significant potential is summarizing tabular data with contextual awareness—an inherently difficult task requiring an understanding of both the data's structure and the relationships between its elements. Recent research has investigated how language models can be leveraged for reasoning and summarization of tables, demonstrating their capability to process both structured and unstructured information. Additionally, studies suggest that these models could enable personalized summaries and help mitigate biases. Despite these advancements, the field remains largely untapped, presenting opportunities for innovative techniques to further improve the precision and efficiency of tabular data summarization.(Chang et al., 2023)(Laskar et al., 2024)

This project aims to introduce a novel framework that leverages adaptive agent roles and personalized summarization layers specifically optimized for table data, enhancing both dynamism and user-centricity. By incorporating a Personalized Multi-Modal Representation, our system tailors table summarization based on user profiles and contextual cues, ensuring more relevant and meaningful outputs. Additionally, our Novel Summarization Model employs agent-based fine-tuning, allowing the system to adaptively learn from user corrections and dataset variability, thereby improving accuracy and personalization over time. For our team, the focus would be on the Summarization Model.

2 Research Objectives & Questions

The key research objectives of this project are:

- Improve the performance of table-based question answering by building and training a router layer that classifies queries as arithmetic, logical, or specialized tasks, then dispatches them to the appropriate adaptive agents.

- Train the adaptive agents and the personalization layer in a coherent manner, ensuring each agent's specialized reasoning integrates smoothly with user-oriented style or formatting requirements.

- Address potential bias, hallucination, and interpretability issues in model outputs by incorporating chain-of-thought (CoT) prompting techniques that enforce step-by-step reasoning.

- Evaluate the effectiveness of the proposed multi-agent approach using standard metrics (such as F1 and classification accuracy), focusing on how accurately the router, specialized agents, and personalization layer perform across TATQA, FINQA, and eventually other datasets.

The main research questions are:

- How does the proposed multi-agent routing method improve on existing single-model baselines?

- What are the key limitations of current QA systems in handling numeric- or logic-intensive tasks, and how can specialized agents address those limitations through tailored training?

- Can this proposed method generalize to multiple datasets, scaling beyond the initial datasets under consideration?

3 Methodology

3.1 Dataset

The project will utilize the following existing datasets:

- TAT-QA: Tabular And Textual dataset for Question Answering

- FinQA: Question-Answering pairs over Financial reports

3.2 Dataset Description

TAT-QA contains 16,552 questions based on 2,757 hybrid contexts from real-world financial reports.

- Contexts consist of a semi-structured table and at least two relevant paragraphs describing or analyzing the table.

- The dataset requires complex numerical reasoning operations such as addition, subtraction, multiplication, division, counting, comparison, sorting, and their compositions.

- Answers include single-span, multi-span, and free-form responses, along with derivations and scales.

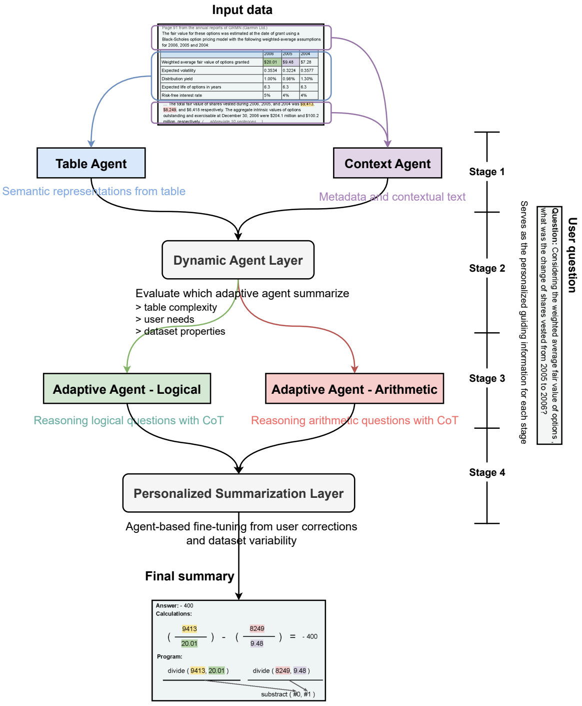
Input data
Page 91 from the annual reports of GRMN (Garmin Ltd.)
The fair value for these options was estimated at the date of grant using a Black-Scholes option pricing model with the following weighted-average assumptions for 2006, 2005 and 2004:
2006
2005
2004
Weighted average fair value of options granted
$20.01
$9.48
$7.28
Expected volatility
0.3534
0.3224
0.3577
Distribution yield
1.00%
0.98%
1.30%
Expected life of options in years
6.3
6.3
6.3
Risk-free interest rate
5%
4%
4%
... The total fair value of shares vested during 2006, 2005, and 2004 was $9,413, $8,249, and $6,418 respectively. The aggregate intrinsic values of options outstanding and exercisable at December 30, 2006 were $204.1 million and $100.2 million, respectively. (... abbreviate 10 sentences ...)
Table Agent
Context Agent
Semantic representations from table
Metadata and contextual text
Dynamic Agent Layer
Evaluate which adaptive agent summarize
> table complexity
> user needs
> dataset properties
Adaptive Agent - Logical
Adaptive Agent - Arithmetic
Reasoning logical questions with CoT
Reasoning arithmetic questions with CoT
Personalized Summarization Layer
Agent-based fine-tuning from user corrections and dataset variability
Final summary
Answer: - 400
Calculations:
(9413 / 20.01) - (8249 / 9.48) = - 400
Program:
divide (9413, 20.01)
divide (8249, 9.48)
substract (#0, #1)
Stage 1
Stage 2
Stage 3
Stage 4
User question
Question: Considering the weighted average fair value of options , what was the change of shares vested from 2005 to 2006?
Serves as the personalized guiding information for each stage

Figure 1: Workflow

FinQA contains 8,281 financial QA examples written by financial experts.

- Includes fully annotated numerical reasoning programs to support financial question answering.

- It is split into training (75%), validation (10%), and test (15%) sets.

3.3 Model & Approach

The proposed NLP system will be built using transformer-based architectures due to their state-of-the-art performance in agent-based personalized tabular context summarization. Below are the key components as seen in Figure 1:

- Table and Context Agents: Extract relevant information from the existing tabular datasets.

- Dynamic Agents Layer: Routes prompts to the appropriate adaptive agents based on table complexity, user needs, and dataset properties.

- Adaptive Agents: Two Adaptive agents will be used for Arithmetic and Logical Reasoning.

- Personalized Summarization Layer: Dynamically adjusts summaries based on user feedback and evolving preferences.

3.3.1 Baseline Models for Comparison

To benchmark our approach, we will compare against the following models:

- GPT-3.5 Turbo: A general-purpose large language model developed by OpenAI, known for its strong reasoning capabilities across various NLP tasks. Although GPT-3.5 Turbo performs well in table-based question answering, it is not specifically optimized for structured numerical reasoning, which may lead to errors in complex tabular data interpretation.

- TAT-LLM (7B): A domain-specific model fine-tuned for question-answering on tabular data, leveraging both arithmetic and logical reasoning. Unlike GPT-3.5 Turbo, which relies on general NLP capabilities, TAT-LLM is trained on financial and tabular datasets, leading to improved accuracy in structured data tasks. Despite impressive performance in both datasets, FinQA and TAT-QA (Zhu et al., 2021), TAT-LLM is optimized for documents that contain tabular and textual data with rich numerical values. However, it may not provide significant advantages for purely textual documents and could face challenges when processing extremely lengthy documents (e.g., those exceeding 100 pages) due to input sequence length constraints.

We compare against GPT-3.5 Turbo and TAT-LLM (7B) because they represent two different paradigms in tabular data processing. GPT-3.5 Turbo serves as a strong general-purpose model, helping to assess how well an LLM without domain-specific fine-tuning handles table-based reasoning. Meanwhile, TAT-LLM (7B) is specifically trained for tabular question answering, making it a relevant benchmark for evaluating our Adaptive Multi-Agent Framework (AMAF), which also focuses on structured numerical and logical reasoning. Compared with these models, we can measure the effectiveness of our adaptive agent-based approach in improving performance on personalized tabular summarization and reasoning tasks.

Each baseline model has strengths, but struggles with fundamental challenges in hybrid financial QA, and none yet has reached human-level performance. Key difficulties include multi-step numerical reasoning, where models often make errors in arithmetic operations or unit conversions, and integrating heterogeneous data from tables and text, leading to missed facts or misaligned information. Additionally, these models lack the ability to adapt to query nuances and user preferences, limiting their flexibility. These challenges highlight the need for improvements in reasoning accuracy, structured data interpretation, and personalized response generation, which our approach aims to address through specialized agents and enhanced adaptability.

3.3.2 Proposed Model Architecture

Our approach employs a combination of open-source and closed-source models to benchmark performance in our dataset, specifically using Google's closed-source Gemini 1.5 Flash and OpenAI's GPT-3.5 Turbo for most of our modeling approaches. To evaluate model performance across different levels of contextual understanding, we will incorporate a combination of zero-shot, one-shot, and few-shot prompting with CoT. Our approach will use adaptive agents for personalization to replace generic responses with more relevant, user-specific interactions. We are also considering variants of CoT - like EEDP (Srivastava et al., 2024) within our specialized agents so each agent provides more transparent, domain-targeted reasoning. In doing so, we aim to reduce hallucinations and give clearer justifications for the final answers.

3.4 Training & Evaluation

The models will be developed using LLM agents and evaluated based on the following metrics:

- F1-score (initial benchmark)

- BLEU (Bilingual Evaluation Understudy): Calculates the precision of n-grams in generated text compared to reference text, mainly used for evaluating machine translation and text generation quality.(Berger et al., 2000)

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): measures the overlap of n-grams between generated text and reference summaries, commonly used for evaluating text summarization.(Roth and tau Yih, 2004)

- Human evaluation (for qualitative assessment)

4 Project Plan & Timeline

The project is structured as follows:

Phase	Tasks	Duration
Phase 1	Literature Review, Dataset Selection and Exploration	2 weeks
Phase 2	Models Selection, Baseline Training	2 weeks
Phase 3	Model Fine-Tuning, Hyperparameter Optimization	3 weeks
Phase 4	Evaluation, Error Analysis	2 weeks
Phase 5	Refinement, Final Report	2 weeks

Table 1: Project Timeline

5 Expected Contributions & Impact

This project will contribute:

- A novel approach to improving how models handle arithmetic and logical reasoning in tabular data through our Adaptive Multi-Agent Framework (AMAF). By routing different types of queries to specialized agents, this approach aims to reduce errors commonly seen in general models like GPT-3.5 Turbo.

- A modular and adaptive architecture that overcomes the monolithic nature of models such as TAT-LLM (7B). Unlike TAT-LLM, which applies a single model approach to all queries, our framework will dynamically select specialized agents, leading to more efficient and accurate processing of tabular data.

- A Personalized Summarization Layer, which we plan to develop to move beyond one-size-fits-all summaries. This layer will adapt to user feedback and dataset variations, ensuring more relevant and user-centric output.

6 Related Work

TAT-LLM (Zhu, 2024) is a fine-tuned LLaMA2-based model that follows a structured Extractor–Reasoner–Executor pipeline for hybrid table-text reasoning. TAT-LLM generates step-wise intermediate results in a tabular format, each row corresponding to a reasoning step. TAT-LLM demonstrated state-of-the-art performance on financial reasoning benchmarks FinQA, TAT-QA, and TAT-DQA, with even its 7B parameter variant outperforming GPT-4 by 0.7–5 EM points across these datasets. In fact, TAT-LLM (7B) surpassed all prior fine-tuned models and GPT-3.5/GPT-4, highlighting that a specialized pipeline can rival much larger general-purpose LLMs. This underscores the benefit of structured reasoning and task-specific training over end-to-end black-box prompting.

Earlier hybrid QA models also adopted structured reasoning to handle tables + text. TagOP (Zhu, 2021) is a mixture-of-experts model with pre-defined discrete operations as specialized predictors. It first tags relevant table and text tokens as evidence, then selects an operation-specific predictor to compute the answer. This approach achieved strong results on TAT-QA, but relies on correct operation selection and can mis-order operands. FinQANet (Chen, 2021) advanced financial QA by using a retriever plus program generator: it retrieves supporting facts from financial reports and then generates a logical program (in a domain-specific language) to derive the answer. Learning an explicit program structure yielded significant accuracy gains over direct answer generation, improving numerical reasoning transparency. UniRPG (Zhou, 2022) unified table-text reasoning as sequenced program generation using a seq2seq model. By decoding a linearized program with constrained grammar, UniRPG achieved a new state-of-the-art on TAT-QA. However, these models follow fixed pipelines or require annotated programs, lacking the adaptivity to adjust reasoning procedures based on query complexity or user needs.

Beyond QA, recent work has explored adaptive summarization and agent-based reasoning relevant to our solution. Researchers propose an interactive Adaptive Summaries system for personalized summarization, where users iteratively accept/reject concepts to include, thereby guiding the summary to better match their preferences (Ghodratnama, 2021). This highlights the value of tailoring outputs to user-specific relevance rather than one-size-fits-all summaries. Separately, large LLMs augmented with multi-step reasoning have shown promise in handling complex data: for instance, chain-of-thought prompting (Wei et al., 2022) enables step-by-step numeric reasoning, though it operates within a single model's static prompt and lacks an interactive adaptation loop. To inject more control, a tri-agent generation pipeline (generator → instructor → editor) that refines an LLM's output according to inferred user instructions was used and it markedly improved the personal relevance of summaries (Xiao, 2024) . Moreover, new tasks like QTSumm (Zhao, 2023) explicitly require reasoning over tables to produce query-focused summaries, emphasizing the need for contextual analysis in generation. These approaches inform our design by showing how iterative feedback, contextual cues, and multi-agent collaboration can yield more personalized and context-aware table summaries.

ReAct (Yao et al., 2022) is a notable paradigm where a single agent interleaves reasoning traces with actions (e.g. tool or API calls) in a unified prompt, enabling the model to gather external evidence mid-thought. This synergy of reasoning and acting mitigates hallucinations and error propagation. However, ReAct still relies on one monolithic agent; decisions about when and how to act are learned implicitly and may falter if the prompt-design fails. ManyModalQA (Hannan, 2020) offers a more explicit multi-module design: it introduces a modality classifier that first decides which knowledge source is likely to contain the answer, then routes the question to a corresponding expert QA module. This agent delegation approach handles heterogeneous inputs and was shown to close some gap between uni-modal systems and human performance on mixed-modality queries. Yet, a hard routing can suffer if the modality disambiguator errs, leading to cascading failures. We extend on this by employing multiple coordinated agents. By dynamically orchestrating these expert agents, our approach leverages modular decision-making while maintaining end-to-end coherence, ultimately pushing multi-agent reasoning into the realm of contextual table summarization.

7 Potential Future Extensions

Future directions include:

- Extending the model to implement Reinforcement Learning.
- Expanding modality by incorporating visual and image data alongside tabular and text data.
- Deployment in real-world applications.

References

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 2000. A maximum entropy approach to natural language processing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 39–47. Association for Computational Linguistics.

Jonathan D. Chang, Kiante Brantley, Rajkumar Ramamurthy, Dipendra Misra, Wen Sun, and et al. 2023. Learning to generate better than your llm.

et al. Chen. 2021. Finqanet: A program-based question answering model for financial text. *Proceedings of EMNLP*.

et al. Ghodratnama. 2021. Adaptive summaries: Interactive personalization in text summarization. *arXiv preprint arXiv:2104.12345*.

et al. Hannan. 2020. Manymodalqa: Multimodal question answering across text, tables, and images. *Proceedings of ICLR*.

Md Tahmid Rahman Laskar, Elena Khasanova, Xue-Yong Fu, Cheng Chen, Shashi Bhushan TN, and et al. 2024. Query-opt: Optimizing inference of large language models via multi-query instructions in meeting summarization. In *Proceedings of the 2024 EMNLP Industry*.

Andrea Matarazzo, Riccardo Torlone, and et al. 2024. A survey on large language models with some insights on their capabilities and limitations.

B. Peng, Y. Zhu, Y. Liu, and et al. 2024. Graph retrieval-augmented generation: A survey.

Dan Roth and Wen tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the 8th Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 1–8. Association for Computational Linguistics.

Pragya Srivastava, Manuj Malik, Vivek Gupta, Tanuja Ganu, and Dan Roth. 2024. Evaluating llms' mathematical reasoning in financial document question answering. *arXiv preprint arXiv:2402.11194*.

Wei, Jason, Tay, Yi, Bommasani, Rishi, Raffel, Colin, Zoph, Barret, Borgeaud, Sebastian, Yu, Adams Wei, Lester, Brian, Du, Nan, Dai, Andrew M., Xie, and Qizhe et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

et al. Xiao. 2024. Tri-agent generation: Generator-instructor-editor framework for personalized summarization. *arXiv preprint arXiv:2402.67890*.

Yao, Shunyu, Yu, Zhengxuan, Narasimhan, Karthik, Cao, Yuan, Lei, and Lili et al. 2022. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

et al. Zhao. 2023. Qtsumm: Query-focused table summarization with contextual reasoning. *Proceedings of NeurIPS*.

et al. Zhou. 2022. Unirpg: Unified reasoning as program generation for table-text qa. *Proceedings of ACL*.

Zhu, Fengbin, Lei, Wenqiang, Huang, Youcheng, Wang, Chao, Zhang, Shuo, Lv, Jiancheng, Feng, Fuli, Chua, and Tat-Seng et al. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

et al. Zhu. 2021. Tagop: A table-and-text based qa model for complex financial reasoning. *Proceedings of ACL*.

et al. Zhu. 2024. Tat-llm: Structured reasoning for table-and-text question answering. *arXiv preprint arXiv:2401.13223*.