Evaluating Large Language Models for Budget-Related Question Answering

Shivam Singh Rawat
School of Information Sciences
University of Illinois at Urbana-Champaign
Champaign, USA
shivambitid007@gmail.com

Abstract

This study evaluates the effectiveness of Large Language Models (LLMs) in retrieving and interpreting information from complex U.S. federal budget documents to enhance public finance transparency and accessibility. A Retrieval-Augmented Generation (RAG) framework is implemented using SentenceTransformer embeddings and Pinecone vector database to support query-based semantic retrieval. Five LLMs—LLaMA 3.2B, Mistral 7B, GPT-3.5-turbo 20B, Gemma2 2B, and DeepSeek-R1 1.5B—are tested on 50 curated queries spanning fiscal years 2023 to 2025. Model-generated answers are evaluated against manually constructed reference responses using BLEU, ROUGE-1/2/L, METEOR, TER, BERTScore, and semantic similarity metrics. LLaMA 3.2B achieved the strongest overall performance, with a BLEU score of 0.2403, ROUGE-L of 0.4550, METEOR of 0.5545, and the lowest TER among top models. GPT-3.5-turbo followed closely, with BLEU of 0.2091 and METEOR of 0.4938. In contrast, DeepSeek-R1, the smallest model, scored lowest across most metrics, including a BLEU of 0.0164 and TER of 91.6. These results demonstrate the ability of larger models to generate accurate and fluent responses for complex, domain-specific financial texts when paired with effective retrieval mechanisms.

1 Introduction

This project focuses on evaluating the capability of Large Language Models (LLMs) to accurately retrieve and interpret information from complex U.S. federal budget documents. The primary objective is to predict how effectively different LLMs—such as LLaMA, Mistral, GPT-3.5, DeepSeek R1, and Gemma2 2—can answer structured, budget-related questions. The context for this task stems from the growing need for transparent and accessible analysis tools in public finance. Federal budget documents are often dense, technical, and difficult for non-experts to interpret. By leveraging LLMs for question answering, this study seeks to improve the accessibility of budgetary information for policymakers, researchers, journalists, and the public.

To implement this task, a Retrieval-Augmented Generation (RAG) approach was employed. The U.S. federal budget documents were first segmented into smaller, meaningful chunks and ingested into a Pinecone vector database. When a query is posed, the system retrieves the top five most relevant document chunks based on semantic similarity. These retrieved segments are then fed into various LLMs to generate responses. The final evaluation is performed by comparing these LLM-generated responses with a set of mutually agreed reference answers.

To systematically assess the quality of the generated responses, a comprehensive set of evaluation metrics is applied, including METEOR, BERTScore, Semantic Similarity, Translation Edit Rate (TER), ROUGE, and BLEU scores. These metrics evaluate linguistic quality, semantic alignment, and content fidelity, providing a holistic view of each model's performance. Evaluating LLM effectiveness in this domain will help determine their suitability for practical applications in promoting government transparency and supporting informed policy analysis.

2 Literature Review

The evaluation of Large Language Models (LLMs) in Retrieval-Augmented Generation (RAG) systems has gained increasing attention due to their potential to enhance factual accuracy and mitigate hallucinations. This literature review synthesizes recent research on the accuracy of five LLMs—Mistral:7b, Gemma2:2b, LLaMA3.2:3b, DeepSeek-R1:1.5B, and GPT-3.5—in the RAG framework. The discussion focuses on retrieval quality, response accuracy, and hallucination mitigation strategies.

In the research paper "Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases" by Li et al. (2024), they use a form of RAG to help improve the factual accuracy of LLM's. They created a Retrieval Augmented Generation (RAG) pipeline that uses several models including a retriever model and a generator model. The retriever model retrieves the most relevant information from the dataset and returns it as context for later generation. The generator model takes the returned information and adds it to the QA prompts as context to generate answers. They also use a Mixedbread.ai embedding model. An embedding model is a model that calculates vector embeddings for words or phrases. This is important, because the smaller the size of embedding, the better it will perform. So they chose this model for it's compact size and high performance. They also use a BgeRerank reranker model, which employs the BAAI/bge-reranker-large model. This model assesses pairs of queries and documents, assigning relevance scores that reflect each document's alignment with the query's intent. BgeRerank reorders the initial broad set of documents, prioritizing the top-N (top-5 out of 10 documents retrieved in our experiments) most pertinent ones to the user's query. For the core generation model, they use LLAMA-2. They chose this model because it is an open source model that allows for continuous training and experiments. They also state that it has demonstrated impressive performance across a wide range of natural language processing benchmarks. Pre-trained on 2 Trillion tokens and further fine-tuned on 100k human annotation data, LLaMA-2 can capture linguistic patterns and domain knowledge, enabling it to generate highly fluent and coherent text.

In the research paper "LRP4RAG: Detecting Hallucinations in Retrieval-Augmented Generation via Layer-wise Relevance Propagation" by (Hu et al., 2024), they highlight the persistent challenge of hallucinations in Retrieval-Augmented Generation (RAG) systems, despite RAG's role as a key method for reducing inaccuracies in large language models (LLMs). The authors propose a novel approach named LRP4RAG, which utilizes Layer-wise Relevance Propagation (LRP) to detect hallucinations by evaluating the relevance between inputs and outputs in the RAG generator. This method aims to address the gap in understanding and detecting the conditions under which hallucinations occur in RAG systems. LRP4RAG leverages a back-propagation technique to analyze the correlation between prompt inputs and LLM-generated responses, processing the relevance matrix through classifiers to predict the presence of hallucinations. The approach is evaluated using the RAGTruth dataset, demonstrating superior performance over existing baselines. This research not only introduces a new method for detecting hallucinations but also underscores the importance of analyzing internal LLM states to improve the reliability of RAG systems.

In the research paper "Honest AI: Fine-Tuning "Small" Language Models to Say "I Don't Know", and Reducing Hallucination in RAG" by Chen et al. (2024), they tackle the issue of hallucination in Large Language Models (LLMs) by proposing a fine-tuning strategy named "Honest AI." This method trains smaller LLMs to accurately respond with "I don't know" when faced with uncertain information, enhancing their reliability in accuracy-sensitive applications. The authors also explore and evaluate several Retrieval-Augmented Generation (RAG) approaches, including integrating search engine and knowledge graph results, and find that combining RAG with fine-tuning yields the best performance in the CRAG benchmark. Their hybrid approach not only improves accuracy but also effectively reduces hallucination by enabling models to better discern and admit uncertainty. This research underscores the importance of fine-tuning in conjunction with RAG to significantly enhance LLM performance, particularly in generating truthful and reliable responses. Although I won't be fine-tuning the LLMs, effective prompt engineering can ensure they respond with "I Don't Know" when asked questions beyond the scope of the U.S. Budget (2023-2025).

In "A Brief Survey of Vector Databases" by (Xie et al., 2023), the authors provide a comprehensive overview of vector databases, which are increasingly relevant in the context of RAG systems. The paper highlights the challenges of managing high-dimensional data and how vector databases offer efficient solutions for storing and retrieving vector embeddings, which are crucial for tasks like similarity search. The authors discuss various similarity search algorithms (e.g., K-Means, Locality Sensitive Hashing, Hierarchical Navigable Small Worlds) and similarity metrics (e.g., Euclidean Dis-

tance, Dot Product, Cosine Similarity) that are essential for optimizing retrieval processes in vector databases. These techniques are directly applicable to RAG systems, where the quality of retrieval directly impacts the accuracy of generated responses. The paper also compares popular vector database products like Pinecone, Chroma, and Milvus, emphasizing their scalability, reliability, and compatibility, which are critical factors for integrating them into RAG pipelines. This survey underscores the importance of vector databases in improving the efficiency and precision of retrieval processes, which is vital for improving the performance of LLMs in RAG frameworks. In my project I'll be using Pinecone as my vector database.
Additionally, I drew insights from the Learning semantic similarity for very short texts (De Boom et al., 2015) work on semantic similarity for very short texts, which explores the use of distributed word representations to capture semantic relationships in contexts with limited word overlap. This research informs our understanding of how embedding models can enhance semantic similarity in budget-related queries, particularly when dealing with concise or fragmented text.

3 Descriptive Statistics of the Data

The dataset used in this study comprises U.S. federal budget documents from fiscal years 2023 to 2025, totaling 528 pages—156 pages from 2023, 184 from 2024, and 188 from 2025. These documents are publicly available and provide detailed insights into federal financial allocations, revenue sources, and policy priorities across a wide range of sectors. In total, three budget documents were analyzed.

The dataset includes references to 21 unique federal departments, with frequent mentions of major agencies such as the Department of Agriculture, Department of Commerce, and Department of Homeland Security. This breadth of departmental coverage ensures a diverse and representative sample of U.S. federal spending and planning priorities.

To evaluate the performance of Large Language Models (LLMs), a curated set of 50 sample queries was developed, focusing on key budget topics from the 2023, 2024, and 2025 documents. These queries span issues such as funding allocations, policy initiatives, and strategic investments across different departments. They are designed to challenge the models' ability to retrieve, synthesize, and accurately report information from complex and lengthy budget texts. Table 1 showcases 10 of the 50 manual queries.

Table 1: Sample Budget Queries and Manual Reference Responses

#	Query	Manual Response
1	How much discretionary funding does the President's 2023 Budget request for USDA?	The Budget requests $28.5 billion in discretionary funding for USDA, a $4.2 billion or 17.1-percent increase from the 2021 enacted level, excluding Food for Peace Title II Grants.
2	What steps does the 2023 Budget take to address wildfires?	The Budget provides nearly $4.9 billion for Forest Service Wildland Fire Management, including $2.2 billion for the Wildfire Suppression Operations Reserve Fund. It also upholds the President's commitment that no Federal firefighter would make less than $15 an hour, increases the size of the Federal firefighting workforce, and provides critical technological support for wildfire detection and response, including FireGuard satellite imagery.
3	How does the 2023 Budget strengthen the Nation's supply chains?	The Budget provides $372 million, an increase of $206 million from the 2021 enacted level, for the National Institutes of Standards and Technology's (NIST) manufacturing programs. These resources would help launch two additional Manufacturing Innovation Institutes in 2023 and continue support for two institutes funded in 2022 as part of the Administration's growing Manufacturing USA network.
4	How does the 2023 Budget address the impacts of climate change and extreme weather?	The Budget invests $6.9 billion in the National Oceanic and Atmospheric Administration (NOAA), an increase of $1.4 billion from the 2021 enacted level, supporting programs that would catalyze wind energy, restore habitats, protect the oceans and coasts, and improve NOAA's ability to predict extreme weather associated with climate change.
5	How does the 2023 Budget advance U.S. leadership in emerging technologies?	The Budget includes a $187 million increase for research initiatives at NIST that would focus on developing standards to accelerate adoption of critical and emerging technologies with a focus on artificial intelligence, quantum science, and advanced biotechnologies.
6	What measures does the 2023 Budget propose to strengthen deterrence in the Indo-Pacific region?	The Budget prioritizes China as the Department's pacing challenge. DOD's 2023 Pacific Deterrence Initiative highlights some of the key investments the Department is making that are focused on strengthening deterrence in the Indo-Pacific region.
7	How does the 2023 Budget address cybersecurity threats?	The Budget invests in cybersecurity programs to protect the Nation from malicious cyber actors and cyber campaigns. These priorities include strengthening cyber protection standards for the defense industrial base and investing in the cybersecurity of DOD networks.
8	How does the 2023 Budget invest in both scientific research and pandemic preparedness?	The Budget proposes a major investment of $5 billion for ARPA-H, significantly increasing direct Federal research and development spending in health. The Budget makes transformative investments in pandemic preparedness and biodefense across HHS public health agencies—$81.7 billion available over five years—to enable an agile, coordinated, and comprehensive public health response to future threats.
9	What investments does the 2023 Budget make to end the HIV/AIDS epidemic?	The Budget includes $850 million across HHS to aggressively reduce new HIV cases, increase access to pre-exposure prophylaxis (PrEP), and ensure equitable access to services and supports for those living with HIV.
10	How does the 2023 Budget improve pay and workforce policies for TSA employees?	The Budget provides a total of $7.1 billion for TSA pay and benefits, an increase of $1.6 billion from the 2021 enacted level, to compensate TSA employees at rates comparable to their peers in the Federal workforce and expand access to labor benefits.

4 Preprocessing and Transformation Steps

The preprocessing and transformation of the U.S. federal budget documents (2023–2025) involved several key steps to prepare the data for efficient storage and retrieval using Pinecone, a vector database. Below is a detailed explanation of the steps:

4.1 Loading the Documents

The budget documents were loaded using the `PyPDFLoader` from the `langchain_community.document_loaders` library. This loader extracts text content from PDF files, making it accessible for further processing.

Three budget documents were loaded:

- `budget_fy2025.pdf`
- `BUDGET-2023-BUD.pdf`
- `BUDGET-2024-BUD.pdf`

4.2 Text Splitting

To handle the large size of the budget documents, the text was split into smaller, manageable chunks using the `RecursiveCharacterTextSplitter` from the `langchain` library. This splitter ensures that the text is divided into meaningful fragments while preserving context.

The following parameters were used:

- `chunk_size=1000`: Each chunk contains up to 1000 characters.
- `chunk_overlap=100`: A 100-character overlap between chunks ensures continuity and prevents loss of context at chunk boundaries.

4.3 Generating Embeddings

The text fragments were converted into vector embeddings using the `SentenceTransformerEmbeddings` model (`all-mpnet-base-v2`). This model generates high-quality embeddings that capture the semantic meaning of the text, enabling efficient similarity search and retrieval.

5 Model parameters, and evaluation results

This study evaluates the performance of five Large Language Models (LLMs) in answering budget-related queries using a retrieval-augmented generation (RAG) framework. The models assessed include LLaMA 3.2 (3B), Mistral (7B), Gemma2 (2B), DeepSeek-R1 (1.5B), and GPT-3.5-turbo (20B). Each model was tested on 50 carefully curated budget-related questions, designed to reflect a wide range of government finance topics such as spending breakdowns, policy initiatives, and departmental investments.

For each question, a manual reference answer was created based on the budget documents. These reference responses served as the ground truth for evaluation, enabling consistent, high-quality comparison of generated model outputs. Each query was processed using a Retrieval-Augmented Generation (RAG) pipeline, where the top-5 semantically relevant document chunks were retrieved from a Pinecone vector database and passed into the LLMs for response generation.

We used the following prompt to ensure relevance and control over the model's output:

```
Using the context provided below,
answer the user's question while
ensuring the response is related
to the U.S. budget. While
writing the answer, only provide
details relevant to the user
question. If the question
or context falls outside this
scope, politely inform the user
that no relevant information was
found and suggest rephrasing or
changing the question.
Context: {context}
User question: {user_query}
```

The temperature was set to 0 to minimize hallucinations and ensure deterministic, fact-based responses.

To comprehensively assess the quality of these responses, we applied several widely used natural language evaluation metrics: BERTScore, BLEU, ROUGE-1, ROUGE-2, ROUGE-L, METEOR, Translation Edit Rate (TER), and semantic similarity scores. These metrics measure various dimensions of performance including semantic alignment, lexical overlap, fluency, and information preservation.

Among all models, LLaMA 3.2B exhibited the best overall performance, achieving the highest scores in BLEU (0.2403), ROUGE-1 (0.4704), ROUGE-2 (0.3220), ROUGE-L (0.4550), METEOR (0.5545), and semantic similarity (0.7776). These results highlight its strong capability to generate fluent, accurate, and semantically rich responses. Representative outputs from LLaMA 3.2B are presented in Table 2.

GPT-3.5-turbo (20B) closely followed, achieving the highest BERTScore (0.910) and very competitive results across all other metrics, including a semantic similarity of 0.767, showing its consistency and accuracy in aligning with reference responses. Representative outputs from GPT-3.5 are presented in Table 3.

Mistral produced moderate results with strong METEOR (0.397) and BERTScore (0.883), but

Table 2: Sample Queries and Responses by LLaMA

#	Query	LLaMA Response
1	How much discretionary funding does the President's 2023 Budget request for USDA?	The President's 2023 Budget requests $28.5 billion in discretionary funding for USDA, which is a $4.2 billion or 17.1-percent increase from the 2021 enacted level, excluding Food for Peace Title II Grants.
2	What steps does the 2023 Budget take to address wildfires?	The 2023 Budget provides nearly $4.9 billion for Forest Service Wildland Fire Management, including $2.2 billion for the Wildfire Suppression Operations Reserve Fund. It increases the size of the Federal firefighting workforce and supports wildfire detection and response technologies like FireGuard satellite imagery. It also invests $325 million in Hazardous Fuels Management and Burned Area Rehabilitation to restore lands affected by fire, and complements Bipartisan Infrastructure Law efforts to mitigate wildfire risks.
3	How does the 2023 Budget strengthen the Nation's supply chains?	The 2023 Budget strengthens supply chains by allocating $372 million to NIST's manufacturing programs to launch two new Manufacturing Innovation Institutes and support two existing ones. It also provides a $125 million increase for the Manufacturing Extension Partnership to enhance competitiveness of small and medium-sized manufacturers.

Table 3: Sample Queries and Responses by GPT-3.5

#	Query	GPT-3.5 Response
1	How much discretionary funding does the President's 2023 Budget request for USDA?	The President's 2023 Budget requests $28.5 billion in discretionary funding for USDA, which represents a $4.2 billion or 17.1-percent increase from the 2021 enacted level.
2	What steps does the 2023 Budget take to address wildfires?	The Budget provides nearly $4.9 billion for Forest Service Wildland Fire Management, including $2.2 billion for the Wildfire Suppression Operations Reserve Fund. It ensures a minimum wage of $15/hour for federal firefighters, expands the firefighting workforce, and invests in technologies like FireGuard satellite imagery. It also includes $325 million for Hazardous Fuels Management and Burned Area Rehabilitation to reduce wildfire risk and restore affected lands.
3	How does the 2023 Budget strengthen the Nation's supply chains?	The Budget allocates $372 million to NIST's manufacturing programs to launch new Manufacturing Innovation Institutes and support existing ones, boosting American manufacturing. It also provides a $125 million increase for the Manufacturing Extension Partnership to enhance the competitiveness of small and medium-sized manufacturers.

Table 4: Sample Queries and Responses by Mistral

#	Query	Mistral Response
1	How much discretionary funding does the President's 2023 Budget request for USDA?	The President's 2023 Budget requests $28.5 billion in discretionary funding for the U.S. Department of Agriculture (USDA).
2	What steps does the 2023 Budget take to address wildfires?	The Budget increases the Federal firefighting workforce, sustains a pilot program using satellite imagery for early detection, and invests in Hazardous Fuels Management and Burned Area Rehabilitation programs. It also supports a new drought and fire science initiative at the U.S. Geological Survey and promotes equitable access to parks for underserved communities.
3	How does the 2023 Budget strengthen the Nation's supply chains?	The Budget strengthens supply chains by funding NIST's manufacturing initiatives, allocating $230 million for port infrastructure, and supporting local supply chain resilience. It enhances cybersecurity within the ICT supply chain and increases funding for the Bureau of Industry and Security to enforce export controls and monitor supply chain trends in critical technologies.

Table 5: Sample Queries and Responses by Gemma

#	Query	Gemma Response
1	How much discretionary funding does the President's 2023 Budget request for USDA?	The President's 2023 Budget request for the U.S. Department of Agriculture (USDA) is $28.5 billion.
2	What steps does the 2023 Budget take to address wildfires?	The context does not provide specific information about the 2023 Budget's steps to address wildfires, so I cannot answer this question from the provided context.
3	How does the 2023 Budget strengthen the Nation's supply chains?	The context provides a detailed explanation of how the 2023 Budget strengthens the Nation's supply chains. It highlights how the budget provides $372 million for the National Institutes of Standards and Technology's manufacturing programs to launch two additional manufacturing innovation institutes in 2023 and continue support for two institutes funded in 2022. Additionally, it provides a $125 million increase for the Manufacturing extension Partnership to make America's small and medium manufacturers more competitive.

a lower BLEU score (0.102) and similarity score of 0.730, suggesting more generic or paraphrased outputs. Representative outputs from Mistral are presented in Table 4

Gemma2 displayed balanced performance, achieving a BLEU of 0.120, METEOR of 0.320, and similarity score of 0.710, reflecting moderate lexical and semantic alignment. Representative outputs from Gemma are presented in Table 5

DeepSeek-R1, being the smallest model, showed the weakest performance overall with the lowest BLEU (0.016), highest TER, and a similarity score of 0.643, indicating limitations in handling complex and domain-specific content. Representative outputs from deepseek-r1 are presented in Table 6

These results suggest that larger models, such as LLaMA 3.2B and GPT-3.5-turbo (20B), are better suited for tasks requiring nuanced comprehension of complex financial documents. In contrast, smaller models may require further fine-tuning or domain-specific pretraining to perform effectively in specialized applications like budget analysis. Detailed results are presented in Table 7 and Table 8.

6 Error Analysis, Insights, and Interpretation

While the overall evaluation provided valuable quantitative metrics, a deeper qualitative error analysis revealed key limitations and insights regarding model behavior and performance.

6.1 Error Patterns and Examples

A common theme observed across multiple LLMs was their inability to generate a response when the retrieved context from the vector database lacked directly relevant information. For example, in response to the query *"What is the total discretionary funding requested for HUD in the 2023 Budget?"*, several models replied with statements like *"The context does not provide information about the total discretionary funding requested for HUD in the 2023 Budget, so I cannot answer this question*

Table 6: DeepSeek-R1 Full Responses to Sample Budget Queries

#	Query	DeepSeek-R1 Response
1	How much discretionary funding does the President's 2023 Budget request for USDA?	<think> Okay, so I need to figure out how much discretionary funding the President's 2023 Budget requests for the Department of Agriculture (USDA)... Therefore, the answer is that the President's 2023 Budget requests $28.5 billion in discretionary funding for the Department of Agriculture (USDA). </think> The President's 2023 Budget requests $28.5 billion in discretionary funding for the Department of Agriculture (USDA).
2	What steps does the 2023 Budget take to address wildfires?	<think> Okay, so I need to figure out what steps the 2023 U.S. budget takes to address wildfires based on the provided context... These actions collectively aim to reduce wildfire risk and improve community safety. </think> The 2023 U.S. budget addresses wildfires through several key steps: 1. Supporting Wildfire Management Programs... 4. Providing Preparation Resources.
3	How does the 2023 Budget strengthen the Nation's supply chains?	<think> Okay, so I need to figure out how the 2023 U.S. budget strengthens the nation's supply chains... This likely strengthens the nation's supply chains by supporting businesses to produce more efficiently and meet demand effectively. </think> The 2023 U.S. budget strengthens the nation's supply chains by increasing funding for domestic manufacturing through specific programs like NIST's initiatives...

Table 7: Evaluation Metrics for LLMs on Budget QA Task (Semantic Similarity)

Model	BERT	Similarity
DeepSeek-R1	0.8121	0.6425
Gemma2 2B	0.8874	0.7097
GPT-3.5	0.9097	0.7668
LLaMA 3.2B	0.9056	0.7776
Mistral 7B	0.8835	0.7305

from the provided context." This behavior highlights a limitation in the retrieval stage of the RAG pipeline, suggesting a need for improved retrieval strategies to surface more contextually relevant passages. Nonetheless, the model's refusal to answer questions outside the scope of its retrieved knowledge is a positive and expected outcome, reflecting responsible handling of uncertainty.

In other cases, LLMs provided more detailed or expanded answers compared to the manual references. For example, in response to the query *"How does the 2023 Budget address the impacts of climate change and extreme weather?"*, the manual reference focused specifically on NOAA's $6.9 billion investment, whereas GPT-3.5 generated a broader, accurate summary mentioning $18 billion allocated across several federal agencies. Though semantically rich, these differences led to lower ROUGE and BLEU scores due to lexical mismatch, despite strong semantic overlap.

6.2 Metric Behavior and Model Artifacts

Interestingly, models such as LLaMA 3.2B and GPT-3.5-turbo achieved over 90% on BERTScore and high semantic similarity, indicating strong

Table 8: Evaluation Metrics for LLMs on Budget QA Task (Lexical and Structural)

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	TER
DeepSeek-R1	0.0164	0.1614	0.0534	0.1549	0.2225	91.5993
Gemma2 2B	0.1196	0.3379	0.1654	0.3068	0.3205	1.0043
GPT-3.5	0.2091	0.4552	0.2793	0.4391	0.4938	1.1281
LLaMA 3.2B	0.2403	0.4704	0.3220	0.4550	0.5545	1.4929
Mistral 7B	0.1016	0.3323	0.1653	0.3133	0.3966	2.3737

alignment with the intended meaning of manual responses, even when the exact phrasing differed. However, BLEU and ROUGE scores were occasionally low, especially when models avoided jargon or used more general language, making the answers more accessible but lexically divergent.

The DeepSeek-R1 model consistently underperformed in traditional metrics. A notable artifact in its responses was the inclusion of structured segments such as <think> and <answer>, where the <think> portion contained internal reasoning. While this structure reflects a form of chain-of-thought reasoning, it negatively impacted metrics like BLEU and ROUGE, which are sensitive to verbosity and token overlap. Cleaning such artifacts before evaluation may yield more accurate metric reflections.

6.3 Recommendations and Future Directions

This analysis highlights the need for more sophisticated evaluation methods that combine both semantic and contextual relevance. Sole reliance on surface-level metrics like ROUGE or BLEU may undervalue meaningful yet lexically diverse responses. Incorporating human judgment or feedback from domain experts is essential for more accurate quality assessments in future work.

Furthermore, the relatively weak performance of DeepSeek-R1 suggests that models with fewer parameters may struggle in specialized domains such as government budgeting. Future iterations should explore either models with larger parameter counts or those fine-tuned on public finance datasets.

Finally, implementing advanced or hybrid retrieval strategies—such as reranking, query expansion, or feedback-based refinement—may enhance the RAG pipeline's ability to deliver highly relevant contexts, ultimately improving the quality and reliability of generated answers.

7 Limitations

While this study offers valuable insights into the application of LLMs for budget-related question answering, several limitations must be acknowledged.

7.1 Dataset Scope and Coverage

The dataset comprises only three U.S. federal budget documents from fiscal years 2023 to 2025. While these documents are representative of federal budget planning, they do not cover state or local government budgets, nor do they reflect the full diversity of financial documentation encountered in public sector analysis. Furthermore, the curated set of 50 queries, though diverse, is centered primarily on the 2023 budget, potentially limiting generalizability across fiscal years and policy shifts.

7.2 Evaluation Constraints

The evaluation relies heavily on automated metrics such as BLEU, ROUGE, METEOR, BERTScore, TER, and semantic similarity. While these metrics offer useful signals of model performance, they are not fully aligned with human judgment of answer quality, especially when answers are lexically different but semantically equivalent. For example, LLMs that paraphrased or generalized their responses tended to receive lower scores on overlap-based metrics despite producing factually correct outputs. This suggests a need for hybrid evaluation frameworks that integrate both automated scoring and human review, particularly in domains where precision and policy nuance are critical.

7.3 Model Limitations and Artifacts

Models varied significantly in their response behavior. DeepSeek-R1, for instance, embedded intermediate reasoning using <think> sections, which—while potentially beneficial for explainability—degraded performance on string-matching metrics like ROUGE and BLEU. This artifact also introduced verbosity that was not present in other model outputs. Smaller parameter models, such as DeepSeek-R1 (1.5B), also struggled more than larger counterparts like GPT-3.5-turbo (20B) or LLaMA 3.2B when interpreting nuanced or complex budgetary queries. This disparity highlights the need for further model-specific tuning and preprocessing adaptations.

7.4 Retrieval Limitations in RAG Pipeline

The retrieval process, though effective in many cases, occasionally failed to supply context relevant to the user query. In such instances, even high-performing LLMs returned fallback responses such as "The context does not provide information...", despite the presence of relevant information elsewhere in the corpus. This points to a limitation in top-k retrieval and embedding coverage, underscoring the potential benefit of incorporating retrieval enhancement techniques such as reranking, multi-hop search, or hybrid retrieval using keyword and vector-based methods.

7.5 Human Supervision and Ground Truth Quality

While manual reference answers were used for evaluation, they were not reviewed or verified by policy experts or federal budget analysts. As a result, there may be cases where multiple correct answers exist but only one was considered for comparison. This limits the interpretive power of quantitative metrics and highlights the importance of involving domain experts in future benchmarking efforts.

References

Xinxi Chen, Li Wang, Wei Wu, Qi Tang, and Yiyao Liu. 2024. Honest ai: Fine-tuning" small" language models to say" i don't know", and reducing hallucination in rag. *arXiv preprint arXiv:2410.09699*.

Cedric De Boom, Steven Van Canneyt, Steven Bohez, Thomas Demeester, and Bart Dhoedt. 2015. Learning semantic similarity for very short texts. In *2015 ieee international conference on data mining workshop (icdmw)*, pages 1229–1234. IEEE.

Haichuan Hu, Yuhan Sun, and Quanjun Zhang. 2024. Lrp4rag: Detecting hallucinations in retrieval-augmented generation via layer-wise relevance propagation. *arXiv preprint arXiv:2408.15533*.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*.

Xingrui Xie, Han Liu, Wenzhe Hou, and Hongbin Huang. 2023. A brief survey of vector databases. In *2023 9th International Conference on Big Data and Information Analytics (BigDIA)*, pages 364–371. IEEE.