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

Understanding and analyzing government budgets is essential for policymakers, researchers, and the public. Given the complexity of U.S. federal budget documents, Large Language Models (LLMs) offer a promising approach for enhancing accessibility and comprehension. This study evaluates the effectiveness of five LLMs—Llama 3.2, Gemma, Mixtral, GPT-3.5 or later, and R1—in answering budget-related questions. Using a dataset of U.S. federal budget documents from 2023 to 2025, we preprocess the data into structured text and assess model performance across accuracy, relevance, consistency, and completeness. A structured 50-question questionnaire is used to compare responses, ensuring uniform evaluation. The findings will provide insights into the suitability of LLMs for budget analysis, aiding in improved financial data retrieval and interpretation.

1 Introduction

Understanding and analyzing government budgets is crucial for policymakers, researchers, and citizens. Given the complexity of budget documents, employing Large Language Models (LLMs) for information retrieval and question answering can significantly enhance accessibility and comprehension. This study aims to evaluate the effectiveness of various LLMs in extracting accurate responses from U.S. federal budget documents.

2 Research Questions

This study focuses on answering the following questions using text mining techniques:

1. How accurately can different LLMs respond to a structured questionnaire based on U.S. federal budgets?
2. Which LLM performs best in terms of factual accuracy and consistency?

3 Dataset Description

The dataset consists of U.S. federal budget documents from 2023, 2024, and 2025. These are comprehensive government reports detailing financial allocations, revenue sources, and policy priorities. Examples from the dataset include:

- *"The President's Budget proposes an investment of $1.2 trillion in infrastructure projects over the next decade."*
- *"The 2024 budget estimates a revenue increase of 5.6% compared to the previous fiscal year."*

4 Data Preprocessing

- Convert PDF documents into structured text.
- Segment the data by sections (e.g., expenditures, revenues, economic projections).
- Remove irrelevant metadata and extract numerical figures where applicable.
- Tokenize text for model input and fine-tuning.

5 Methodology and LLM Models

5.1 LLM Models

We will evaluate the following five LLMs:

Table 1: LLM Models, Versions, and Prices

Model	Version	Price (per 1k token)
Llama	3.2	$0.0
Gemma	To Be Decided	$0.0
Mixtral	To Be Decided	$0.0
GPT	3.5-turbo or later	$0.0080+
R1	1.0	$0.0

5.2 Workflow

1. Selecting appropriate versions of all five LLMs.
2. Implementing code to query all models for sample questions.
3. Designing a 50-question structured questionnaire covering various budget aspects.
4. Engineering a uniform prompt for consistency across models.
5. Generating and collecting responses from all LLMs.
6. Comparing responses based on accuracy, completeness, and consistency.

6 Evaluation Metrics

- **Accuracy**: Compare model responses against ground truth answers from budget documents.
- **Relevance**: Assess if responses correctly address the context of the question.
- **Consistency**: Evaluate how stable responses are across multiple queries.
- **Completeness**: Check if models provide full and precise answers.

7 Expected Outcomes

This study will offer insights into the performance of various LLMs in extracting factual budgetary information. The results will help identify the most suitable models for budget analysis and text mining tasks, ultimately improving accessibility to complex financial data.

The LaTeX and BibTeX

References

Tianchen Gao, Jiashun Jin, Zheng Tracy Ke, and Gabriel Moryoussef. 2025. A comparison of deepseek and other llms. Accessed: 2025-02-14.

Todor Ivanov and Valeri Penchev. 2024. Ai benchmarks and datasets for llm evaluation. Accessed: 2025-02-14.

OpenAI. 2025. Openai api pricing. Accessed: 2025-02-14.