CS 6120: Natural Language Processing | Final Project Report
Retrieval-Augmented Generation for United States Congressional Bills and Laws

Andrew Cai	Joynae Whitehurst	Lili Xiang
cai.and@northeastern.edu	whitehurst.j@northeastern.edu	xiang.l@northeastern.edu

PROJECT GIT REPOSITORY

1. Executive Summary and Abstract

We developed a Retrieval-Augmented Generation (RAG) system that uses a Facebook AI Similarity Search (FAISS) index housing semantic embeddings (generated with all-MiniLM-L6-v2) of United States Congressional bill text to answer legislation-related questions. We evaluated several free LLMs from OpenRouter, such as lite versions of DeepSeek and Gemini, to generate digestible summaries that include citations to specific bills. The deployed user interface is built using Streamlit and hosted on Google Cloud Platform (GCP). Our project aims to minimize hallucination, ensure answers are verifiable, and have results arrive in a reasonable time to provide accurate and transparent answers to the public.

2. Motivation and Impact

Understanding legislative changes is critical for policymakers, researchers, and the public. However, congressional bills are long, complex, and frequently updated. This makes it difficult to track changes and understand implications. This project aims to leverage RAG-based models to provide concise, accurate, and cited bill summaries and information. Our project will enable more transparent legislative analysis, helping users quickly retrieve relevant bill details, amendments, and key voting records. Furthermore, taking information directly from bills will allow for more objective information rather than bias from news sources or other secondary sources. Our system supports the reduction of misinformation by answering from only primary source material.

3. Background and Related Work

Recent advances in large language models (LLMs) have transformed how unstructured documents are processed and queried, especially in knowledge-intensive domains like law and public policy. Among these, RAG has emerged as a dominant paradigm for combining retrieval precision with the generative flexibility of LLMs. RAG systems typically embed documents into a vector space, retrieve relevant chunks based on query similarity, and pass them to a language model for generation [1]. Inspired by success in medical question and answer and legal natural language processing [2], we applied RAG to U.S. Congressional legislation. Vector embeddings and implementation guided how the pre-processed data we generated will be fed into our model [3]. We referenced studies on text chunking for long documents and zero-shot classification to categorize bills into topics like military, budget, and education [4]. While commercial applications often use closed-source APIs our project intentionally uses free and open-source alternatives that further offer transparency and cost-efficiency.

In our project domain, congressional bills originate in either the House of Representatives or the Senate. Then they pass through multiple stages before becoming law. The United States Government Publishing Office provides data on bills. The raw data consists of legislative documents in XML formats from the GovInfo Congressional Bills Collection (https://www.govinfo.gov/bulkdata/BILLSUM). Due to the bill/XML formatting consistencies across all bills, we were able to easily parse the XML documents for the relevant metadata and text to create a structured data table metadata such as measure ID, measure type, date, chamber, and full text. The dataset was processed using text chunking techniques to generate over 100,000 rows from documents spanning the 115th to the current 119th Congress.

4. Modeling Methodology

Though we initially planned to explore Latent Dirichlet Allocation (LDA) and Named Entity Recognition (NER) for topic extraction, we focused on optimizing retrieval and summarization performance instead. Our dataset is sourced from the GovInfo Congressional Bills Collection, encompassing text from the 115th to 119th Congress. Furthermore, we had to assess risks such as hallucinations, API rate limits or outages, and cost of inference if scaled broadly. Specifically with costs we had to operate with a $50 GCP budget and available CPU/GPU memory in our instances.

Our system uses a modular RAG pipeline that integrates open-source LLMs, efficient vector search, and controlled generation. The end goal is to deliver accurate, citation-supported answers to user queries about U.S. legislative bills while staying within tight budget and compute constraints.

The major components of our pipeline are:

Preprocessing: Bills are retrieved in XML format and cleaned using a custom parser (e.g. cleaned up HTLM encoding issues). We use a tokenizer-compatible chunking strategy to break large documents into ~500-token segments for embedding efficiency. We specifically used LangChain for chunking because it has built-in text splitters that split text at natural language boundaries, token-aware splitting, and compatibility with FAISS and embedding models from HuggingFace. These features allowed for improved retrieval and generation quality [5].

Embedding: Our project uses Sentence Transformers for generating high-quality semantic embedding. We used the all-MiniLM-L6-v2 model from SentenceTransformers for embedding which is a distilled transformer known for its strong performance on semantic similarity tasks despite a compact 22M parameter size. This aligns with growing trends in LLM research favoring lightweight models with less memory-intensive/compute costs [6].

Indexing: For retrieval, we employed FAISS, an open-source library designed for efficient similarity search. Our use of flat indexing ensures deterministic top-k recall and enables CPU or GPU deployment, making the system easy to implement in resource constrained environments like low-tier GCP instances [7].

Controlled Generation: Unlike many RAG setups that use large commercial APIs, our system relies entirely on open-access LLMs. We used accessible models that had over 1 billion parameters to generate 2–3 sentence summaries that cite relevant legislative chunks. The choice of Gemini (gemini-2.0-flash) balances accuracy with controllability with its instruction-tuned design. This architecture allows for precise generation that rely on prompt structure and mitigates hallucinations [8]. Furthermore, this model performed the best in our evaluations.

README: https://github.com/sruxll/rag_us_congressional_bills/blob/main/README.md

5. Data and Data Analysis

All our data can be found at GovInfo ("Govinfo | U.S. Government Publishing Office." *Govinfo.gov*, 2000, www.govinfo.gov/.) which had bill lengths vary significantly, from one paragraph resolutions to multi-line detailed bills [9]. The dataset contains structured metadata in **Table 1** below. In our processing, we ensured that no values were null. Chunks were examined for content density and language consistency.

Column	Non-Null Count	Dtype
congress	121238	Int64
measure_type	121238	String
measure_number	121238	Int64
measure_id	121238	String
origin_chamber	121238	String
current_chamber	121238	String
orig_publish_date	121238	Date
update_date	121238	Date
title	121238	String
action_date	121238	Date
action_desc	121238	String
topic_tags	121238	List
text_chunk	121238	String
embedding	121238	List

Table 1: Column Information (Post-Processing)

As one can see from below in **Tables 2-6**, class Imbalance is present in a sense (congress, bill type, originated chamber, and topics) even though our task is not classification. Although we are not doing classification directly, our dataset of bill chunks is imbalanced by policy area. This could result in bias retrieval if embedding similarity alone favors heavily represented topics. We acknowledge this form of implicit class imbalance in our project since we did not use the other metadata columns besides the text chunks and embeddings as inputs to our RAG model. We did not explicitly address this during preprocessing. However, our retrieval pipeline uses semantic similarity (SentenceTransformer embeddings and FAISS) and introduces diversity by grouping results by bill title and limiting the number of returned summaries. This helps mitigate the dominance of any single bill or topic in the final output.

Congress	Bill Counts
115	33312
116	31081
117	33064
118	21941
119	1840

Table 2: Congress vs Bill Counts

Congress	115	116	117	118	119
House Bill	20873	17814	19537	12897	1214
House Concurrent Resolution	260	141	155	104	24
House Joint Resolution	228	670	139	177	65
House Simple Resolution	1560	1459	1761	1381	110
Senate Bill	9190	9954	10338	6489	334
Senate Concurrent Resolution	129	67	75	30	14
Senate Joint Resolution	108	102	88	95	26
Senate Simple Resolution	964	874	971	768	53

Table 3: Congress and Bill Type Counts

Originated Chamber	Bill Count
HOUSE	80569
SENATE	40669

Table 4: Chamber vs Bill Count

Measure Type	Bill Count
House Bill	72335
House Concurrent Resolution	684
House Joint Resolution	1279
House Simple Resolution	6271
Senate Bill	36305
Senate Concurrent Resolution	315
Senate Joint Resolution	419
Senate Simple Resolution	3630

Table 5: Measure Type vs Bill Count

Index	Topic	Count
0	Congress	82736
1	Social Welfare	48729
2	Health	46979
3	Commemorations	46776
4	Taxation	45823
...	...	...
5369	Africa and Pacific Conflict	1
5370	Africa and Oceans	1
5371	Afghanistan	1
5372	Aeronautica	1
5373	International Affairs	1

Table 6: Topic Tag Counts

6. Results and Evaluation

We compare and evaluate multiple LLMs (Gemini, DeepSeek, Qwen) using:

- **BERTScore** (semantic similarity to reference answers): RAG systems generate free-form text, which can't be evaluated by exact match or BLEU scores [10]. It measures semantic similarity between a generated summary and a human-written reference answer. Using contextual embeddings from BERT, we evaluate meaning, not just word overlap which is ideal for tasks like summarization or question and answer [11]. In our evaluation, with range 0 to 1, 0.85 or higher means strong similarity, 0.75 or higher means okay similarity, and lower than 0.70 is weak similarity.

- **Cosine similarity** (between query and retrieved chunks): High-quality generation depends on retrieving relevant context. It measures embedding similarity between the user's query and the top-k retrieved document chunks. It ensures that the retrieval system (FAISS/embeddings) is surfacing contextually appropriate information [12]. In our evaluation, with range -1 to 1, 0.8 or higher means very similar, 0.6 or higher means somewhat similar, anything lower means it is a poor result.

- **ROUGE-L** (Recall-Oriented Understudy for Gisting Evaluation – Longest Common Subsequence): This evaluates the overlap between the generated answer and a human-written reference answer by measuring the longest common subsequence (LCS). It captures both precision and recall in how much the output text resembles the reference. It focuses on word order where the longer the sequence of words in the correct order, the higher the score [13]. In our evaluation, with range 0 to 1, 0.4 or higher weans strong word and sequence overlap.

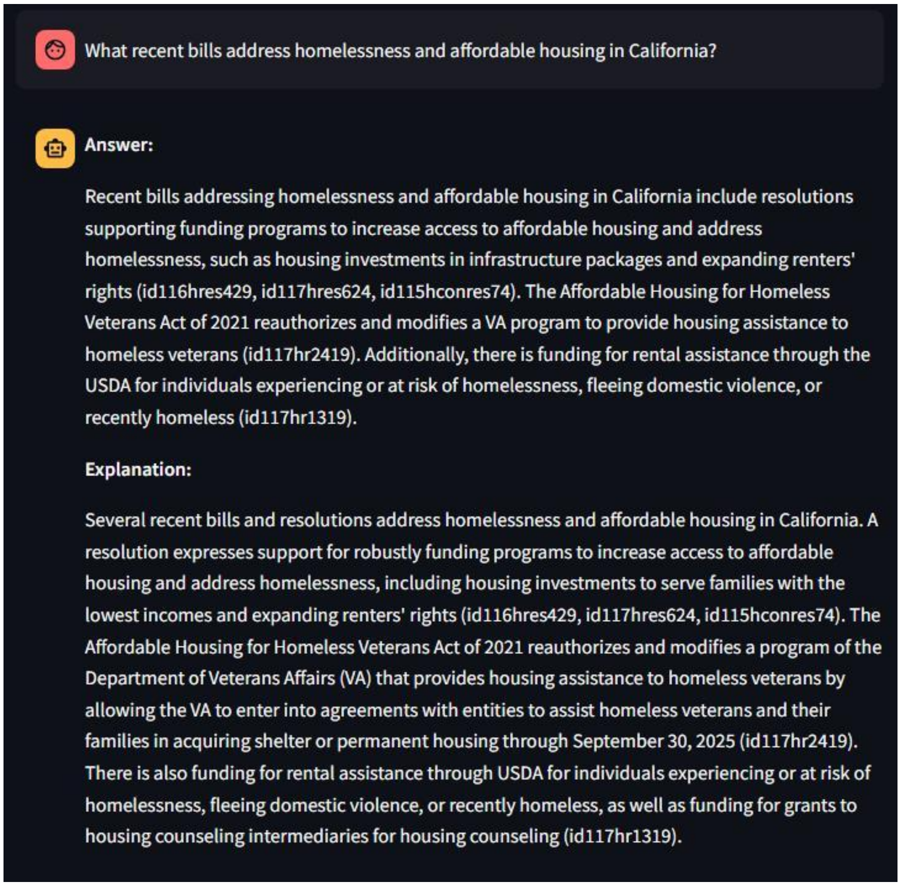
Table 7: Model Evaluations [14]

query	model	Tokens (B)	response_time (sec)	bert_score	cosine_similarity	rouge_l
What bills address climate change and renewable energy?	deepseek/deepseek-r1:free	40.9	94.86	0.837	0.719	0.291
	google/gemini-2.0-flash-exp:free	32.1	2.79	0.872	0.741	0.387
	qwen/qwq-32b:free	1.15	27.12	0.847	0.747	0.304
	deepseek/deepseek-chat-v3-0324:free	94.8	7.3	0.858	0.733	0.383
What legislation exists for healthcare access and affordability?	deepseek/deepseek-r1:free	40.9	47.77	0.812	0.719	0.176
	google/gemini-2.0-flash-exp:free	32.1	2.51	0.845	0.744	0.243
	qwen/qwq-32b:free	1.15	9.65	0.829	0.744	0.245
	deepseek/deepseek-chat-v3-0324:free	94.8	100.76	0	0	0
What policies exist regarding border security and immigration reform?	deepseek/deepseek-r1:free	40.9	18.26	0.844	0.723	0.256
	google/gemini-2.0-flash-exp:free	32.1	2.69	0.895	0.757	0.453
	qwen/qwq-32b:free	1.15	15.36	0.859	0.745	0.281
	deepseek/deepseek-chat-v3-0324:free	94.8	7.46	0.86	0.726	0.397
How is Congress addressing AI regulation and data privacy?	deepseek/deepseek-r1:free	40.9	12.27	0.832	0.713	0.209
	google/gemini-2.0-flash-exp:free	32.1	3.17	0.87	0.804	0.416
	qwen/qwq-32b:free	1.15	57.74	0	0	0
	deepseek/deepseek-chat-v3-0324:free	94.8	7.71	0.831	0.7	0.22
What bills exist for student loan forgiveness and education funding?	deepseek/deepseek-r1:free	40.9	56.15	0.833	0.748	0.272
	google/gemini-2.0-flash-exp:free	32.1	2.92	0.878	0.786	0.45
	qwen/qwq-32b:free	1.15	38.22	0.868	0.791	0.375
	deepseek/deepseek-chat-v3-0324:free	94.8	49.44	0.868	0.768	0.394

Model	Response Time (sec)	BERTScore	Cosine Similarity	ROUGE-L
deepseek/deepseek-r1:free	45.9	0.83	0.7	0.2
google/gemini-2.0-flash-exp:free	2.8	0.87	0.8	0.4
qwen/qwq-32b:free	29.6	0.68	0.6	0.2
deepseek/deepseek-chat-v3-0324:free	34.5	0.68	0.6	0.3

Table 8: Averaged Metrics Across Models

Based on **Table 7 and 8**, Gemini (gemini-2.0-flash) consistently returned answers the quickest (less than 5 seconds), high BERTScores (>0.85), Cosine similarities (0.8), and ROUGE-Ls (0.4) on average. This model was chosen for our project and **Figures 1-3** demonstrates the model answering the questions well with citations to check. Looking at the results in GovInfo, these bills were relevant and overall, accurately summarizing the listed bills. With our metrics and current outputs in our chatbot live on GCP, we have high confidence that our project can answer U.S. legislative questions sufficiently.



What recent bills address homelessness and affordable housing in California?

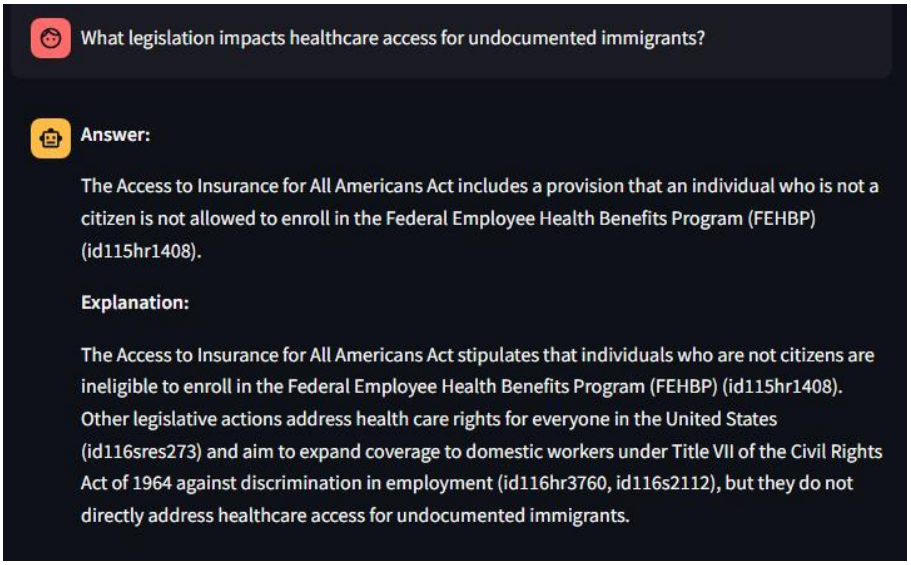
Answer:

Recent bills addressing homelessness and affordable housing in California include resolutions supporting funding programs to increase access to affordable housing and address homelessness, such as housing investments in infrastructure packages and expanding renters' rights (id116hres429, id117hres624, id115hconres74). The Affordable Housing for Homeless Veterans Act of 2021 reauthorizes and modifies a VA program to provide housing assistance to homeless veterans (id117hr2419). Additionally, there is funding for rental assistance through the USDA for individuals experiencing or at risk of homelessness, fleeing domestic violence, or recently homeless (id117hr1319).

Explanation:

Several recent bills and resolutions address homelessness and affordable housing in California. A resolution expresses support for robustly funding programs to increase access to affordable housing and address homelessness, including housing investments to serve families with the lowest incomes and expanding renters' rights (id116hres429, id117hres624, id115hconres74). The Affordable Housing for Homeless Veterans Act of 2021 reauthorizes and modifies a program of the Department of Veterans Affairs (VA) that provides housing assistance to homeless veterans by allowing the VA to enter into agreements with entities to assist homeless veterans and their families in acquiring shelter or permanent housing through September 30, 2025 (id117hr2419). There is also funding for rental assistance through USDA for individuals experiencing or at risk of homelessness, fleeing domestic violence, or recently homeless, as well as funding for grants to housing counseling intermediaries for housing counseling (id117hr1319).

Figure 1: Example Query about Homelessness and Housing in California



What legislation impacts healthcare access for undocumented immigrants?

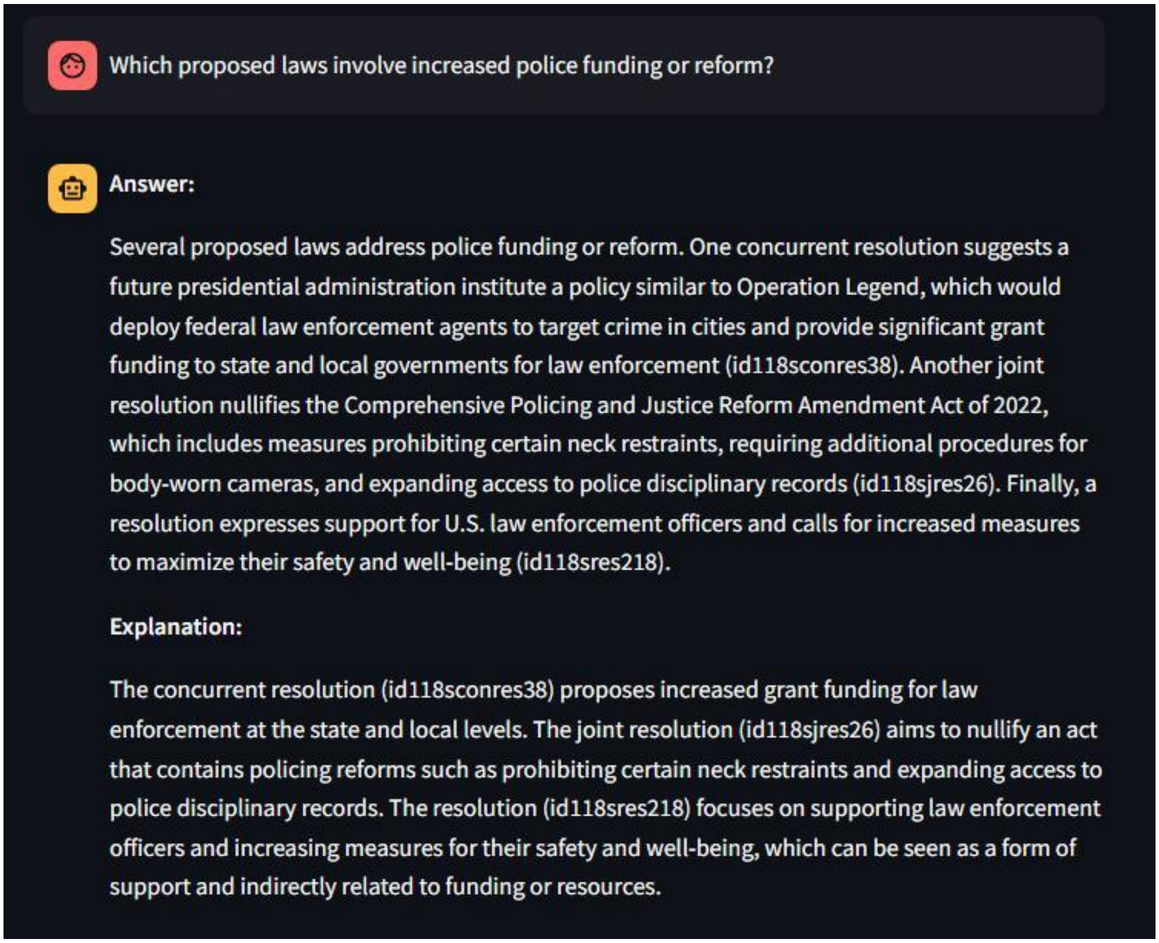
Answer:

The Access to Insurance for All Americans Act includes a provision that an individual who is not a citizen is not allowed to enroll in the Federal Employee Health Benefits Program (FEHBP) (id115hr1408).

Explanation:

The Access to Insurance for All Americans Act stipulates that individuals who are not citizens are ineligible to enroll in the Federal Employee Health Benefits Program (FEHBP) (id115hr1408). Other legislative actions address health care rights for everyone in the United States (id116sres273) and aim to expand coverage to domestic workers under Title VII of the Civil Rights Act of 1964 against discrimination in employment (id116hr3760, id116s2112), but they do not directly address healthcare access for undocumented immigrants.

Figure 2: Example Query about Healthcare and Undocumented Immigrants



Which proposed laws involve increased police funding or reform?

Answer:

Several proposed laws address police funding or reform. One concurrent resolution suggests a future presidential administration institute a policy similar to Operation Legend, which would deploy federal law enforcement agents to target crime in cities and provide significant grant funding to state and local governments for law enforcement (id118sconres38). Another joint resolution nullifies the Comprehensive Policing and Justice Reform Amendment Act of 2022, which includes measures prohibiting certain neck restraints, requiring additional procedures for body-worn cameras, and expanding access to police disciplinary records (id118sjres26). Finally, a resolution expresses support for U.S. law enforcement officers and calls for increased measures to maximize their safety and well-being (id118sres218).

Explanation:

The concurrent resolution (id118sconres38) proposes increased grant funding for law enforcement at the state and local levels. The joint resolution (id118sjres26) aims to nullify an act that contains policing reforms such as prohibiting certain neck restraints and expanding access to police disciplinary records. The resolution (id118sres218) focuses on supporting law enforcement officers and increasing measures for their safety and well-being, which can be seen as a form of support and indirectly related to funding or resources.

Figure 3: Example Query about Police Funding and Reform

7. Conclusions

This project gave us hands-on experience designing, implementing, and evaluating a modern Retrieval-Augmented Generation (RAG) system for legislation summarization. We learned how to balance retrieval quality and generation by combining lightweight embeddings (all-MiniLM-L6-v2), scalable indexing (FAISS), and modular pipelines (LangChain). In doing so, we addressed a practical need to help users navigate large collections of bills with clear, trustworthy summaries and citations to truly understand how their lives are being affected or handled. We learned also how to build a somewhat full-stack project from data sourcing to compiling it into an application with a user interface. Beyond RAG basics, we explored issues such as hallucination detection, confidence estimation, and embedding-query alignment. These insights are transferable to any domain where objective facts are critical such as healthcare and education.

If given more time, we would fine-tune the LLM on domain-specific summaries for better output by applying pre-retrieval logic such as metadata filtering (like chamber or year). We want to explore topic-level tagging or filtering to ensure balanced topics. Another way is to handle our implicit class imbalance through post-retrieval topic-aware reranking [15]. After retrieving the top-k chunks, we would rerank/resample to boost representation of underrepresented topic tags. Also, we would like to add a feedback loop where users rate summaries to improve future retrieval/generation which could incorporate reinforcement learning. Furthermore, combined with reinforcement learning, we can automate hallucination detection through zero-resource and black-box methods that are based on self-contradiction [16].

8. Works Cited

[1] Lewis, Patrick, et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” *ArXiv.org*, 12 Apr. 2021, arxiv.org/abs/2005.11401.

[2] “Exploring the Nexus of Large Language Models and Legal Systems: A Short Survey.” *Arxiv.org*, 2018, arxiv.org/html/2404.00990v1.

[3] Karpukhin, Vladimir, et al. “Dense Passage Retrieval for Open-Domain Question Answering.” *ArXiv:2004.04906 [Cs]*, 30 Sept. 2020, arxiv.org/abs/2004.04906.

[4] Jelodar, Hamed, et al. “Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey.” *Multimedia Tools and Applications*, vol. 78, no. 11, 28 Nov. 2018, pp. 15169–15211, link.springer.com/article/10.1007/s11042-018-6894-4, https://doi.org/10.1007/s11042-018-6894-4.

[5] LangChain. “LangChain.” *Www.langchain.com*, www.langchain.com/.

[6] Wang, Wenhui, et al. “MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers.” *ArXiv (Cornell University)*, 1 Jan. 2020, https://doi.org/10.48550/arxiv.2002.10957.

[7] Meta. “FAISS.” *Ai.meta.com*, ai.meta.com/tools/faiss/.

[8] “Gemini Developer API Pricing.” *Google AI for Developers*, 2025, ai.google.dev/gemini-api/docs/pricing.

[9] “Govinfo | U.S. Government Publishing Office.” *Govinfo.gov*, 2000, www.govinfo.gov/.

[10] “NLP - BLEU Score for Evaluating Neural Machine Translation - Python.” *GeeksforGeeks*, 23 Oct. 2022, www.geeksforgeeks.org/nlp-bleu-score-for-evaluating-neural-machine-translation-python/.

[11] “BERT Score - a Hugging Face Space by Evaluate-Metric.” *Huggingface.co*, huggingface.co/spaces/evaluate-metric/bertscore.

[12] “Cosine Similarity.” *GeeksforGeeks*, 2 Oct. 2020, www.geeksforgeeks.org/cosine-similarity/.

[13] Wikipedia Contributors. “ROUGE (Metric).” *Wikipedia*, Wikimedia Foundation, 23 July 2019, en.wikipedia.org/wiki/ROUGE_(metric).

[14] “OpenRouter.” *OpenRouter*, 2023, openrouter.ai/.

[15] “Unlocking RAG’s Potential: Enhancing Retrieval through Reranking.” *Lftechnology.com*, 2025, www.lftechnology.com/blogs/unlocking-rag-potential-retrieval-through-reranking. Accessed 21 Apr. 2025.

[16] Cao, Zouying, et al. “AutoHall: Automated Hallucination Dataset Generation for Large Language Models.” *ArXiv.org*, 2023, arxiv.org/abs/2310.00259. Accessed 21 Apr. 2025.