Cloud-Based News Chatbot

Final Project Report

1. Abstract

In an age where overwhelming amounts of news are generated online and news articles are increasingly being written in massive numbers, it is redundant to read a news article to get any relevant information immediately. We decided to develop an informative question-answering agent that employs scalable serverless technologies and advanced machine learning to provide real-time, user-tailored information accessible through any modern web browser. Powered by cloud computing, providing a concise overview of a broad field of current affairs is what we looked to achieve.

2. Problem Statement

The exponential growth of online news makes it challenging for users to filter relevant information manually since it is time-intensive and impractical. This project aims to address this challenge by creating a system capable of efficiently answering and presenting news articles in a user-friendly manner. Our chatbot does this by automating the retrieval, processing, and summarization of news articles based on user queries.

For example, given a query like *"What's new with India in politics?"*, the system provides concise updates grouped by categories such as politics, finance, technology, or culture. Foundational language models are introduced to enhance the power of the accuracy and quality of answers to queries. Using cloud-based resources that efficiently integrate and communicate with one another to provide a solution to accessing public data and hosting a seamless user interface with easy-to-use server or serverless resource management, is the crux of this project.

3. Related Work

While incorporating powerful cloud computing resources with language models, we drew inspiration from recent advancements in large language models (LLMs) and retrieval frameworks. Meta's LLaMA, known for efficiency and scalability, excels in natural language generation, especially when paired with Retrieval-Augmented Generation (RAG). Lewis et al. (2020) demonstrated RAG's ability to enhance factual accuracy by integrating external data sources, making it ideal for conversational AI and summarization tasks.

News aggregation systems also benefit from APIs like NewsAPI, which enable real-time article retrieval. Zhang et al. (2018) showcased the value of integrating APIs with NLP models like BERT for personalized content delivery. Building on this foundation, our project combines NewsAPI with a RAG-enabled LLaMA model to deliver relevant, dynamic, and concise news responses.

4. Proposed Solution and Significance

(i) Solution Overview

Our solution integrates various AWS services to deliver a seamless and scalable news chatbot.

- **AWS Lambda**

  Used for serverless computation and API Gateway integration.

- **S3**

  Provided secure and scalable article storage.

- **AWS Bedrock**

  Deployed and executed pre-trained large-scale machine learning models.

- **API Gateway**

  Lightweight HTTP service for monitoring, integrating and communicating with AWS Lambda and News API(public)

The system extracts relevant data based on user queries, processes it using machine learning models, and provides a concise summary in a single paragraph.

(ii) Project Flow

Our application is composed of basic HTML, CSS and JavaScript, making for a simple design of a chatting interface. Whenever the user asks a query and prompts it in the chat box, it is sent to the API gateway that triggers our AWS Lambda function. Brute force code is defined in the python script attached to it that uses RegEx(Regular Expressions) to extract keywords and a potential category that the query is related to or comes under. When this is done, the lambda function triggers the News API. Here the API gateway is constructed as a light HTTP protocol API.

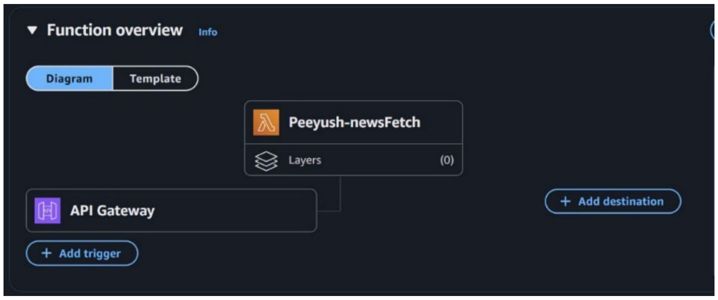


Fig 1. Tree Diagram of Lambda Function and Trigger

The extracted keywords that our program is looking for are “country” and “category” in the query. Based on these parameters, the News API responds using the python requests module and hands it back to the event handler in JSON format. This is promptly then saved to a designated S3 bucket with the date and time for reference.

The lambda function then sends a request to the inference endpoint to process a Retrieval Augmented Generative answer (RAG) with a language model(llama) from Amazon Bedrock. This is done by firstly, retrieving the data via S3 input bucket key, which is then loaded into a fresh variable. This json variable holding the data is then checked to see if it holds a valid article returned by the API that can be added as context for the language model. This context and original query is then passed to the Bedrock model “us.meta.llama3-2-90b-instruct-v1:0.”, a variant of the Llama model released by Meta AI. With the context, our model performs a text-based question-answering task and provides a response that is given back to the lambda function which in turn passes it back to the API gateway back to our front-end interface.

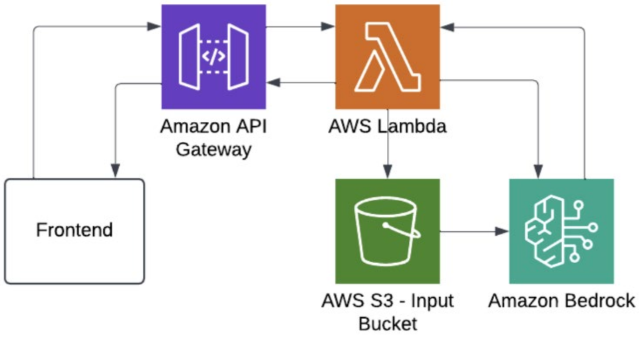


Fig 2. Project Workflow

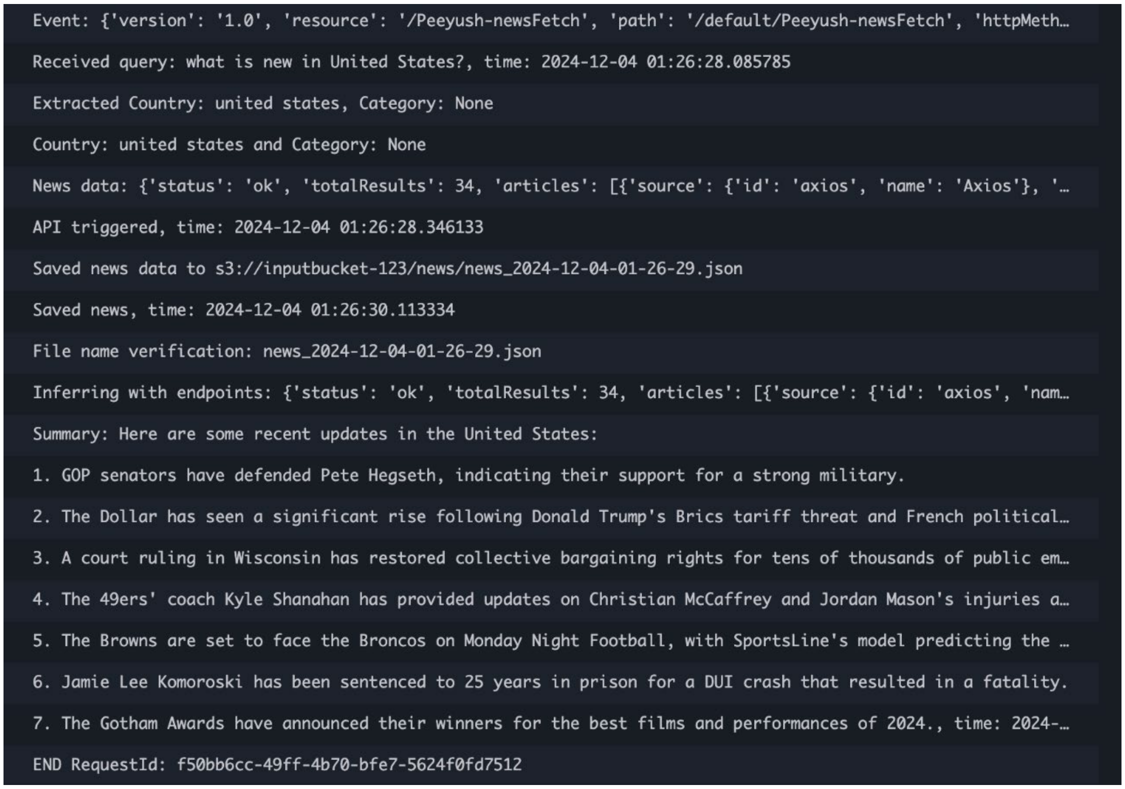
(iii) Significance

- **Serverless Architecture:** Reduces operational overhead while ensuring scalability.
- **ML-Powered Summarization:** Offers tailored responses via state-of-the-art models.
- **Accessibility:** Deployed on a web interface, accessible globally.
- Many people can access at the same time.

5. Evaluation Results

Technical Highlights

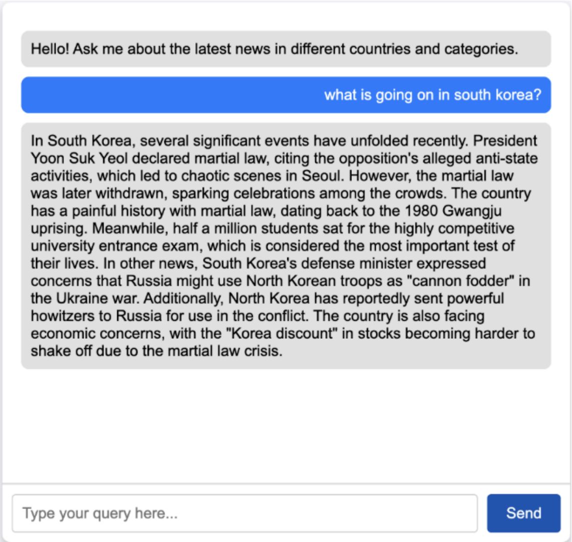
- **News API Integration:** Successfully fetched and processed real-time news data.
- **AWS Bedrock Model:** Delivered concise summaries with the "llama3-2-90b-instruct-v1:0" model.
- **Latency and Scalability:** Optimized query-response cycle to minimize delays.



```
Event: {'version': '1.0', 'resource': '/Peeyush-newsFetch', 'path': '/default/Peeyush-newsFetch', 'httpMeth…
Received query: what is new in United States?, time: 2024-12-04 01:26:28.085785
Extracted Country: united states, Category: None
Country: united states and Category: None
News data: {'status': 'ok', 'totalResults': 34, 'articles': [{'source': {'id': 'axios', 'name': 'Axios'}, '…
API triggered, time: 2024-12-04 01:26:28.346133
Saved news data to s3://inputbucket-123/news/news_2024-12-04-01-26-29.json
Saved news, time: 2024-12-04 01:26:30.113334
File name verification: news_2024-12-04-01-26-29.json
Inferring with endpoints: {'status': 'ok', 'totalResults': 34, 'articles': [{'source': {'id': 'axios', 'nam…
Summary: Here are some recent updates in the United States:
1. GOP senators have defended Pete Hegseth, indicating their support for a strong military.
2. The Dollar has seen a significant rise following Donald Trump's Brics tariff threat and French political…
3. A court ruling in Wisconsin has restored collective bargaining rights for tens of thousands of public em…
4. The 49ers' coach Kyle Shanahan has provided updates on Christian McCaffrey and Jordan Mason's injuries a…
5. The Browns are set to face the Broncos on Monday Night Football, with SportsLine's model predicting the …
6. Jamie Lee Komoroski has been sentenced to 25 years in prison for a DUI crash that resulted in a fatality.
7. The Gotham Awards have announced their winners for the best films and performances of 2024., time: 2024-…
END RequestId: f50bb6cc-49ff-4b70-bfe7-5624f0fd7512
```

Fig 3. Cloud Log of a Successful Inference Integration with Bed Rock

Metrics of accuracy are quite high given that we are using a benchmarked foundational model. Given that the query is reasonably simple and understandable, and the context is successfully extracted, a maximum of 150 words is produced with ease and the number of words can be changed in the AWS Lambda.



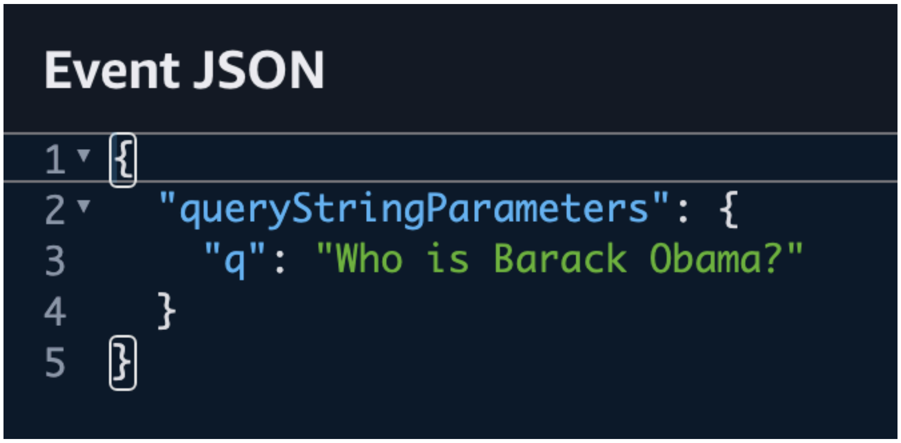
Hello! Ask me about the latest news in different countries and categories.

what is going on in south korea?

In South Korea, several significant events have unfolded recently. President Yoon Suk Yeol declared martial law, citing the opposition's alleged anti-state activities, which led to chaotic scenes in Seoul. However, the martial law was later withdrawn, sparking celebrations among the crowds. The country has a painful history with martial law, dating back to the 1980 Gwangju uprising. Meanwhile, half a million students sat for the highly competitive university entrance exam, which is considered the most important test of their lives. In other news, South Korea's defense minister expressed concerns that Russia might use North Korean troops as "cannon fodder" in the Ukraine war. Additionally, North Korea has reportedly sent powerful howitzers to Russia for use in the conflict. The country is also facing economic concerns, with the "Korea discount" in stocks becoming harder to shake off due to the martial law crisis.

Fig 4. Frontend Interface of a Chatbot Session

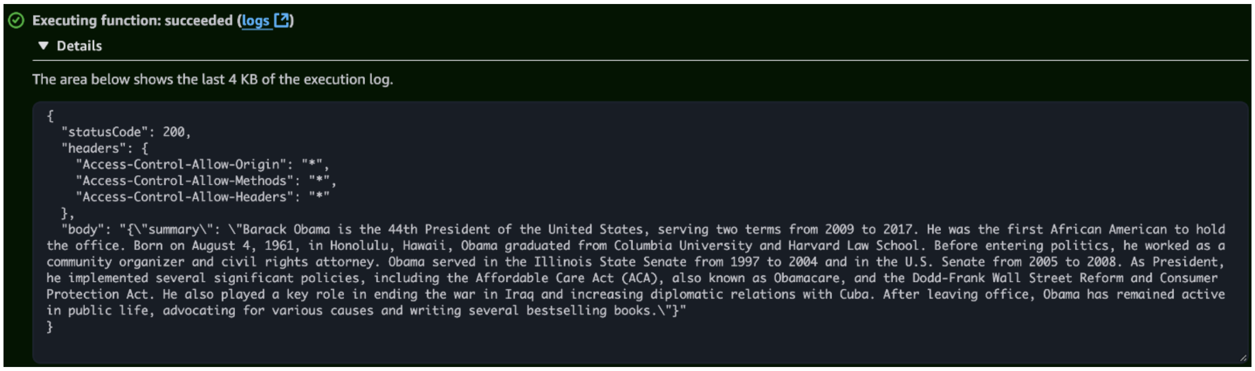
It takes roughly 20-25 seconds for the API gateway, Lambda trigger, and Bedrock to integrate and submit an answer, and for the UI/UX to output it as well. The benefit of a serverless application helps keep a lightweight framework when managing the given resources in AWS.



Event JSON

```
{
  "queryStringParameters": {
    "q": "Who is Barack Obama?"
  }
}
```

Fig 5. Test Query in AWS Lambda



Executing function: succeeded (logs)

Details

The area below shows the last 4 KB of the execution log.

```
{
  "statusCode": 200,
  "headers": {
    "Access-Control-Allow-Origin": "*",
    "Access-Control-Allow-Methods": "*",
    "Access-Control-Allow-Headers": "*"
  },
  "body": "{\"summary\": \"Barack Obama is the 44th President of the United States, serving two terms from 2009 to 2017. He was the first African American to hold
the office. Born on August 4, 1961, in Honolulu, Hawaii, Obama graduated from Columbia University and Harvard Law School. Before entering politics, he worked as a
community organizer and civil rights attorney. Obama served in the Illinois State Senate from 1997 to 2004 and in the U.S. Senate from 2005 to 2008. As President,
he implemented several significant policies, including the Affordable Care Act (ACA), also known as Obamacare, and the Dodd-Frank Wall Street Reform and Consumer
Protection Act. He also played a key role in ending the war in Iraq and increasing diplomatic relations with Cuba. After leaving office, Obama has remained active
in public life, advocating for various causes and writing several bestselling books.\"}"
}
```

Fig 6. Sample Response in AWS Lambda

6. Future Prospects

To enhance the overall system, we aim to integrate additional data sources for richer summaries and optimize the pipeline architecture to minimize costs. Implementing logging with DynamoDB will help track and monitor important data, while exploring platform integration options, such as deploying the system on WhatsApp or other messaging platforms, will extend its reach. To ensure the quality and accuracy of generated content, we focus on mitigating risks associated with AI-generated hallucinations by validating models effectively. Additionally, the goal is to test and deploy more advanced models using GPU-optimized services to improve performance and scalability.

7. Submission Information

- **Team Members:**

  - Krishna Taduri – Initially created a pipeline with SpaCy for extracting country and category information for NLP. He continued updating the Lambda function and then created a simple HTML, CSS, and JavaScript interface. He explored coding an endpoint for SageMaker and looked into integrating SNS, SQS, and DynamoDB with Lambda, but faced challenges. As a result, he tried using Ngrok for local hosting.
  - Peeyush Dyavarashetty – Started by creating a sample Lambda function and proceeded to develop a base DistilBERT model. He rectified errors in the DistilBERT model's learning and answers, then explored ways to speed up the tokenization process in SageMaker. Despite facing delays, he eventually shifted focus to models in Bedrock.
  - Rahul Velaga – Worked with a base DistilBERT model and Squad 2, also experimenting with SNS and DynamoDB. He tested the machine learning models on local host computing environments and validated the Lambda function.
  - Subha Venkat Milind Manda: Assisted in drafting the final project report.
  - Srijinesh Alanka: Worked on creating a React Interface, looked into potential language models, studied and analyzed literature relevant to the project, structured the final report in the aspects of project flow and overview.

- **Implementation Tools Used**
  - **Frontend:** HTML, CSS, and JavaScript
  - **Backend:** AWS Lambda.
  - **Cloud Infrastructure:** AWS Lambda, S3, API Gateway, and Bedrock.
  - Programming Language: Python
  - Coding Interface: Visual Studio Code, AWS Lambda Coding Interface

- **Repository URL:** https://github.com/Peeyush4/AWS-News-Conversational-Chat-Bot

8. References

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. Advances in Neural Information Processing Systems, 33, 9459-9474.
- Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., & Dolan, W. B. (2018). *Personalized dialogue generation: A deep reinforcement learning approach*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2200–2210.