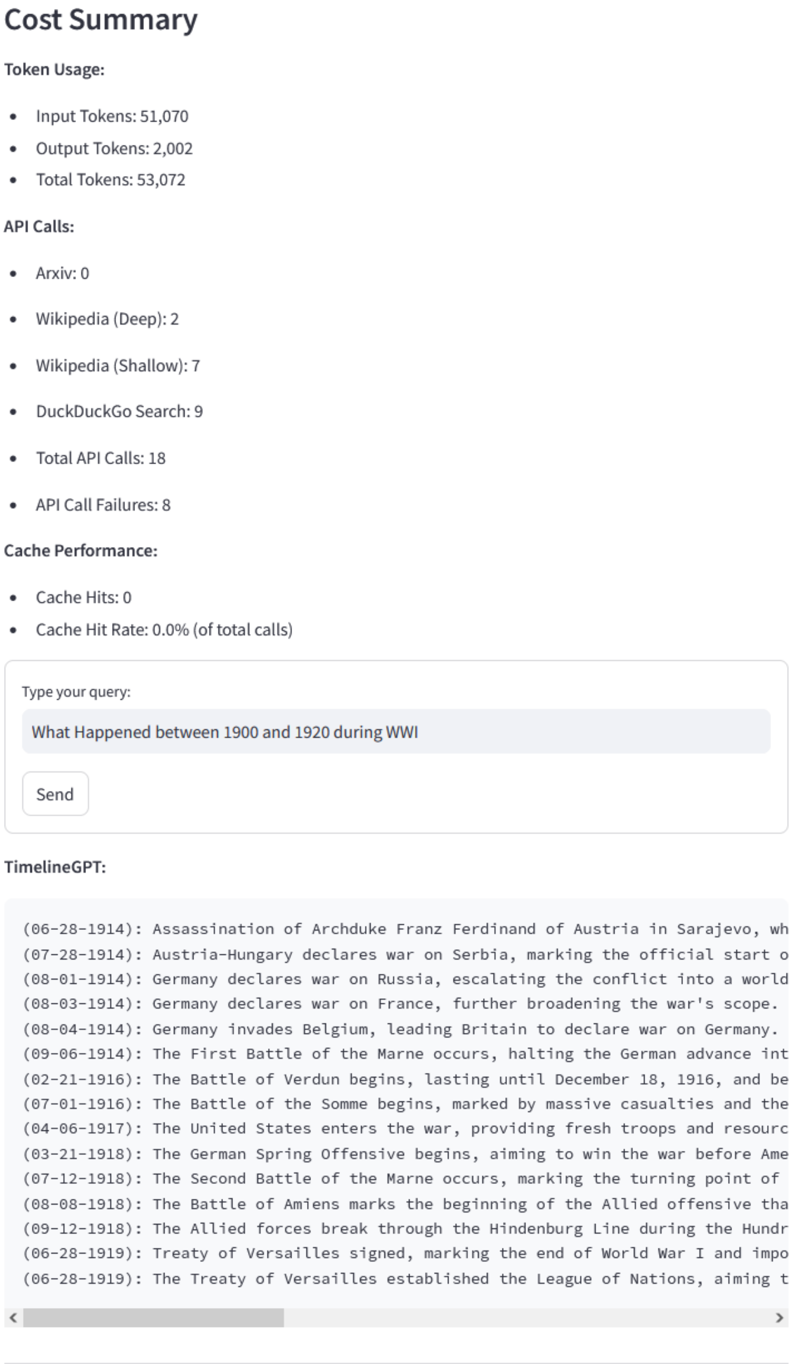
Timeliner: An LLM Timeline Generator with Statefulness and Self Reflection

Nhat Nguyen

12/5/2024



Cost Summary

Token Usage:

- Input Tokens: 51,070
- Output Tokens: 2,002
- Total Tokens: 53,072

API Calls:

- Arxiv: 0
- Wikipedia (Deep): 2
- Wikipedia (Shallow): 7
- DuckDuckGo Search: 9
- Total API Calls: 18
- API Call Failures: 8

Cache Performance:

- Cache Hits: 0
- Cache Hit Rate: 0.0% (of total calls)

Type your query:

What Happened between 1900 and 1920 during WWI

TimelineGPT:

```
(06-28-1914): Assassination of Archduke Franz Ferdinand of Austria in Sarajevo, wh
(07-28-1914): Austria-Hungary declares war on Serbia, marking the official start o
(08-01-1914): Germany declares war on Russia, escalating the conflict into a world
(08-03-1914): Germany declares war on France, further broadening the war's scope.
(08-04-1914): Germany invades Belgium, leading Britain to declare war on Germany.
(09-06-1914): The First Battle of the Marne occurs, halting the German advance int
(02-21-1916): The Battle of Verdun begins, lasting until December 18, 1916, and be
(07-01-1916): The Battle of the Somme begins, marked by massive casualties and the
(04-06-1917): The United States enters the war, providing fresh troops and resourc
(03-21-1918): The German Spring Offensive begins, aiming to win the war before Ame
(07-12-1918): The Second Battle of the Marne occurs, marking the turning point of
(08-08-1918): The Battle of Amiens marks the beginning of the Allied offensive tha
(09-12-1918): The Allied forces break through the Hindenburg Line during the Hundr
(06-28-1919): Treaty of Versailles signed, marking the end of World War I and impo
(06-28-1919): The Treaty of Versailles established the League of Nations, aiming t
```

Problem

Context:

State of the art LLM models today require massive compute for training and operation. Once a model is trained, it can be used, however, it cannot learn new facts about the internet after it has been trained. This point in time is often referred to as the cutoff point. The current LLM strategies employed by companies such as OpenAI and Anthropic are to gather data and retrain their models on the new data occasionally with some tool usage.

Problem:

Because AIs have this cutoff point in their training data, they cannot access new information. If a user were to ask an AI model about the recent news on Bitcoin mining or recent interesting research papers, they would receive an “I don’t know” response from the LLM.

Currently, some tools exist to provide the LLM web search access, however, these models often focus on solving a singular question (example, “what is the height of the Eiffel tower?”) instead of solving more open-ended questions (example, “should I go swimming at Cocoa beach today?” which would require planning to search weather, public advisory, ratings of beach, etc).

A useful task, often time consuming and manual, is scouting research. To keep up to date on the latest information, manual searching and reading is required (if one is lucky, maybe a good emailer’s digest). For any question, it would be useful to have an LLM generate a good summary and visualization of the information. For this problem, this project’s scope is to produce a timeline to help with understanding a new topic.

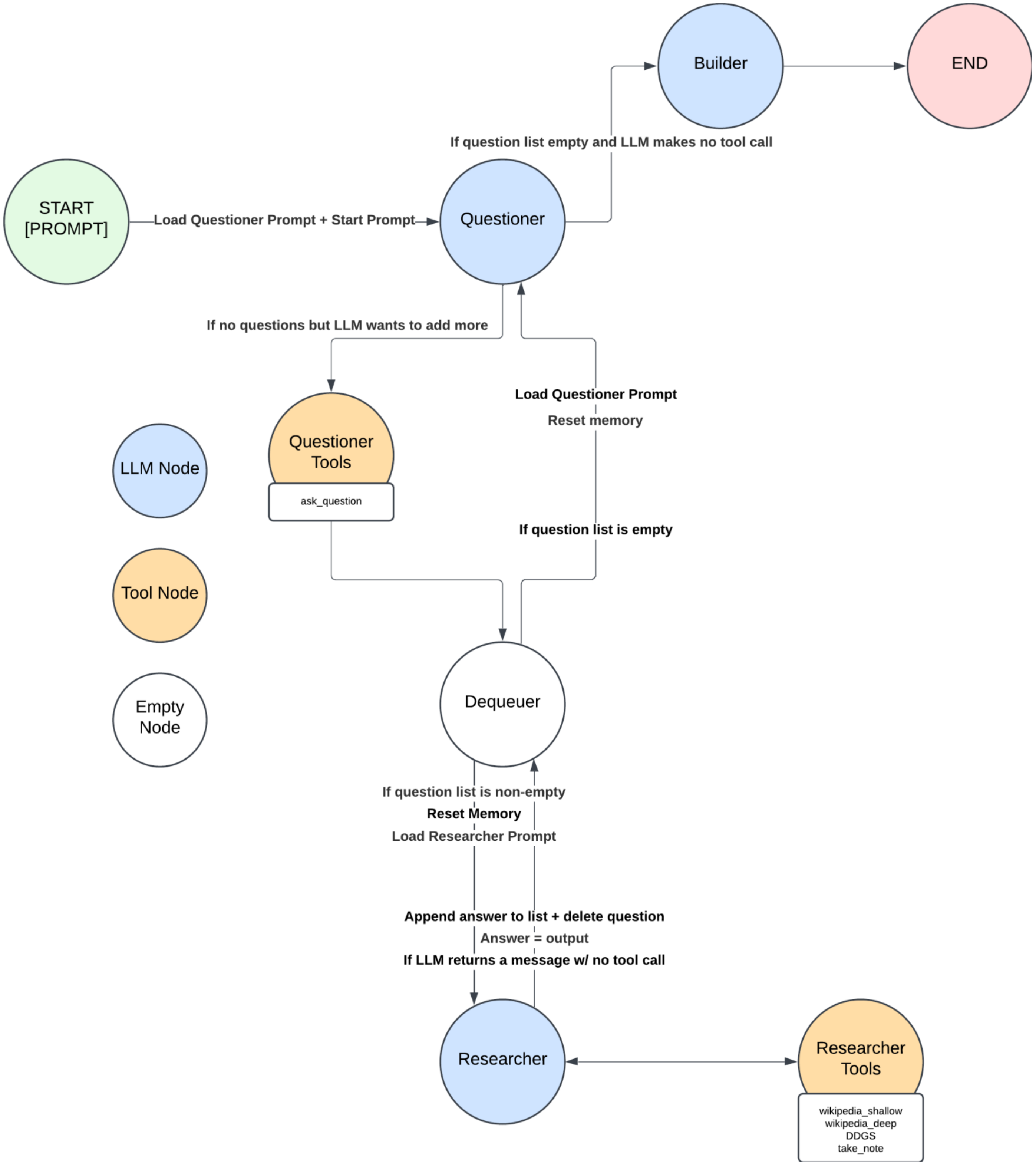
Project Objectives

This project aims to create a framework for an off-the-shelf LLM to search a database (given tools to do so), pick out relevant information, and store it in a timeline format. The model must also direct itself to its own research by asking questions.

1. for an LLM to perform the ReAct-like architecture (acting, observing, repeating)
2. for an LLM to observe the results and ask questions if necessary and perform the actions in parallel
3. for an LLM to “take notes” (capture the source to the information, summarize it, store it for later recall)
4. to create a system that reads the notes and creates a displayable data format in the form of a timeline

The expected outcome should be an LLM system that takes in a user prompt and generates a timeline after doing research on the web.

Methodology


Builder
END
If question list empty and LLM makes no tool call
START [PROMPT]
Load Questioner Prompt + Start Prompt
Questioner
If no questions but LLM wants to add more
Load Questioner Prompt
Reset memory
LLM Node
Questioner Tools
ask_question
If question list is empty
Tool Node
Dequeuer
Empty Node
If question list is non-empty
Reset Memory
Load Researcher Prompt
Append answer to list + delete question
Answer = output
If LLM returns a message w/ no tool call
Researcher
Researcher Tools
wikipedia_shallow
wikipedia_deep
DDGS
take_note

Approach:

Given a user prompt to the LLM with a pre-written system prompt, allow the LLM to explore certain parts of the web on its own.

Two layers will be implemented called **questions** and **notes**. Also to save token usage, the model will be queried a new prompt each time (this achieves **statefulness**). This prompt will include its notes, active questions, answered questions, the user's prompt, list of actions, and other important metadata (like today's date & time).

Algorithm:

1. Receive user input
2. Generate questions
    a. Append questions to a list
3. Answer questions one by one by doing online research
    a. Dequeue a question
    b. Research to answering that question (will sometimes append notes about a result)
    c. Produce an answer
    d. Appen the question + answer to a "QNA list
    e. Repeat until questions list is empty
4. Return to 2, if no more questions are needed or # answers > N, move to 5
5. Build timeline by referencing notes and questions

Questions will allow the LLM to log additional nodes to explore on the web by querying other tools. When evaluating the result, the LLM is allowed to push multiple questions to its queue. After reading the results of each web article, it may fill out multiple answers to its questions and store it in its 'answers' state. The model must answer all questions before it can terminate.

The intuition behind this is that questions provide a mechanism for the LLM to reflect at its research and decide where to move from there. After answering all questions in the queue through research, the model can choose to ask more. This creates a feedback loop that produces better questions given the current results, which in turn produces refined, diverse search topics.

Notes will be short summarizations the AI has made of the researched to allow the LLM to store its session information without consuming tokens to store the entire page. Notes will also be stored internally in a database and can be recalled at any time by the LLM or user for better explainability). Essentially notes can

Notes contain the pointer to the source, the quoted text, and a summary.

Given the prompt: “Give me the last 30 days of what’s happening at tesla.” The LLM may obviously choose to create the initial questions such as “what’s happening at tesla last 30 days”. This would return a page of results.

Given the results, the LLM may now branch out its initial question with new questions (such as “what is the tesla robotaxi?” and “elon musk talk” and attempt to recurrently traverse the question nodes until it is satisfied.

LLM(s) and Techniques:

LLMs:

- GPT-4o latest version
- GPT-4o-mini latest version

Techniques:

- LangChain: used for wrappers and helpers for interacting with the LLM
- ReAct & Tools: ReAct-like prompting plus additional custom coded tools the LLM can use (Multi-Turn Instruction)
- Prompt Engineering: ReAct and structured output prompting strategies
- duckduckgo-search : a library for querying DuckDuckGo for search results, this library was chosen for ease of use and was free
- Wikipedia API: for searching wikipedia, this library was chosen for ease of use and was free
- Arxiv (https://pypi.org/project/arxiv/): used for searching academia papers, this library was chosen for ease of use and was free

Architecture/Process:

1. Initial prompting: User queries the prompt
2. Question Generation: LLM generates a list of questions regarding the prompt
3. Action: loop through all questions and let LLM Decides what action to perform (multiple actions are allowed):
    a. Ask the user some questions
    b. Search the web for questions
    c. Finish
    d. Update internal values (add/remove notes/questions)
4. Go to step 2, repeat until finish.
5. Loop through notes data type and convert it to a timeline

Data:

No training data is needed for this project. Testing data will be created manually. For testing, 10 handwritten prompts will be created along with 5 expected handwritten summaries of the current news. Careful selection will be made to acquire only recent news occurring after the model's training point.

Evaluation:

The primary goal of the framework/algorithm is to produce meaningful goals that help the LLM progress towards answering the prompt. This means that the focus is whether the LLM *mentions* an event, as opposed to its text quality itself, which is negligible. Therefore, for evaluation, we can use a confusion matrix to see evaluation.

To compare results, we will create a ground truth and see if the model can discover it. We will use pick 3 events from different databases. We will then obtain a section of the database that covers a time window and directly give it to an LLM to produce a timeline. To compare, we query the timeline model for the same output format, then use an LLM to count true positives, false positives, and false negatives.

We will use this prompt with ChatGPT-4o directly into to build our ground truth:

Prompt: <USER PROMPT> (ex: What Happened between 1900 and 1920 during WWI?)

Read this text, then produce, in chronological order from past to present, a list of important timeline events.

Follow this format:

(MM-DD-YYYY): <note>

(MM-DD-YYYY): <note>

(MM-DD-YYYY): <note>

Text: <TEXT BLOCK>

Case One: Past Historical Event

The first ground truth concerns a historical event that has already happened in the past and is likely already known by the LLM. We will pick World War I and use Wikipedia + ChatGPT-4o to produce the timeline.

Prompt: What Happened between 1900 and 1920 during WWI?

Ground Truth: https://chatgpt.com/share/6752b7a4-a980-8009-91c8-a74dca133387

(06-28-1914): Archduke Franz Ferdinand of Austria is assassinated in Sarajevo, triggering the July Crisis.
(07-28-1914): Austria-Hungary declares war on Serbia, initiating World War I.
(08-01-1914): Germany declares war on Russia.
(08-03-1914): Germany declares war on France and invades Belgium, violating Belgian neutrality.
(08-04-1914): The United Kingdom declares war on Germany in response to the invasion of Belgium.
(08-12-1914): Austria-Hungary and Serbia clash at the battles of Cer and Kolubara; Serbia repels the attacks with heavy losses for Austria-Hungary.
(08-17-1914): Russia launches an invasion into East Prussia, marking the start of the Eastern Front.
(08-30-1914): New Zealand occupies German Samoa in the Pacific.
(09-05-1914): The First Battle of the Marne begins, resulting in a significant Allied counterattack and halting the German advance into France.
(09-28-1914): German forces lay siege to Antwerp, which falls in October.
(12-25-1914): The Christmas Truce occurs on parts of the Western Front, with unofficial ceasefires between opposing forces.
(02-04-1915): Germany declares a submarine blockade of Britain, intensifying naval warfare.
(04-22-1915): Germany uses chlorine gas for the first time at the Second Battle of Ypres, introducing chemical warfare to the Western Front.
(05-07-1915): The RMS Lusitania is sunk by a German U-boat, killing 1,198 people and increasing tensions with the United States.
(02-21-1916): The Battle of Verdun begins; it will last until December, resulting in nearly 1 million casualties.
(07-01-1916): The Battle of the Somme starts, with the British Army suffering its bloodiest day in history.
(12-29-1916): Grigori Rasputin is assassinated in Russia, intensifying political unrest.
(03-08-1917): The February Revolution in Russia begins, leading to Tsar Nicholas II’s abdication.
(04-06-1917): The United States declares war on Germany and enters World War I.
(11-07-1917): The Bolsheviks seize power in Russia during the October Revolution.
(03-03-1918): Russia signs the Treaty of Brest-Litovsk with Germany, exiting the war.

(07-15-1918): The Second Battle of the Marne begins; it marks the turning point in favor of the Allies.
(11-11-1918): An armistice is signed between Germany and the Allies, officially ending World War I.

Timeliner (GPT-4o-mini):

(06-28-1914): Assassination of Archduke Franz Ferdinand of Austria in Sarajevo, which triggered the July Crisis and led to the outbreak of World War I.
(07-28-1914): Austria-Hungary declares war on Serbia, marking the official start of World War I.
(08-01-1914): Germany declares war on Russia, escalating the conflict into a world war.
(08-03-1914): Germany declares war on France, further broadening the war's scope.
(08-04-1914): Germany invades Belgium, leading Britain to declare war on Germany.
(09-06-1914): The First Battle of the Marne occurs, halting the German advance into France.
(02-21-1916): The Battle of Verdun begins, lasting until December 18, 1916, and becoming one of the war's costliest battles.
(07-01-1916): The Battle of the Somme begins, marked by massive casualties and the first use of tanks in warfare.
(04-06-1917): The United States enters the war, providing fresh troops and resources to the Allies.
(03-21-1918): The German Spring Offensive begins, aiming to win the war before American forces arrive in large numbers.
(07-12-1918): The Second Battle of the Marne occurs, marking the turning point of the war in favor of the Allies.
(08-08-1918): The Battle of Amiens marks the beginning of the Allied offensive that leads to the end of the war.
(09-12-1918): The Allied forces break through the Hindenburg Line during the Hundred Days Offensive.
(06-28-1919): Treaty of Versailles signed, marking the end of World War I and imposing heavy reparations and territorial losses on Germany, while including a 'War Guilt' clause, leading to future tensions.
(06-28-1919): The Treaty of Versailles established the League of Nations, aiming to maintain peace but ultimately failing to prevent future conflicts, particularly World War II.

Result:

Matches (True Positives):	(06-28-1914): Assassination of Archduke Franz Ferdinand. (07-28-1914): Austria-Hungary declares war on Serbia.

	(08-01-1914): Germany declares war on Russia. (08-03-1914): Germany declares war on France. (08-04-1914): UK declares war on Germany (summary aligns, focuses on invasion of Belgium). (02-21-1916): Battle of Verdun begins. (07-01-1916): Battle of the Somme begins. (04-06-1917): US enters World War I.
Misses (False Positives):	(09-06-1914): Predicted First Battle of the Marne (incorrect date; ground truth is 09-05-1914). (03-21-1918): German Spring Offensive (not in ground truth). (07-12-1918): Second Battle of the Marne (incorrect date; ground truth is 07-15-1918). (08-08-1918): Battle of Amiens (not in ground truth). (09-12-1918): Breakthrough of Hindenburg Line (not in ground truth). (06-28-1919): Treaty of Versailles (post-war; not relevant to WWI events listed in ground truth).
Misses (False Negatives):	(08-12-1914): Battles of Cer and Kolubara. (08-17-1914): Russian invasion of East Prussia. (08-30-1914): New Zealand occupies German Samoa. (09-05-1914): First Battle of the Marne. (09-28-1914): Siege of Antwerp. (12-25-1914): Christmas Truce. (02-04-1915): German submarine blockade. (04-22-1915): Use of chlorine gas at Second Battle of Ypres. (05-07-1915): Sinking of RMS Lusitania. (12-29-1916): Assassination of Rasputin. (03-08-1917): February Revolution in Russia. (11-07-1917): October Revolution in Russia. (03-03-1918): Treaty of Brest-Litovsk.

	(07-15-1918): Second Battle of the Marne (mislabeled). (11-11-1918): Armistice signed.

True Positives (TPTP) = 8

False Positives (FPFP) = 6

False Negatives (FNFN) = 15

Precision: 57.1%

Recall: 34.8%

F1 Score: 43.3%

Accuracy: 30.8%

Case Two: Present Historical Event

The second ground truth concerns a historical event that is currently happening and is likely not known by the LLM. We will pick the impeachment of Yoon Suk Yeol and use Wikipedia + ChatGPT-4o to produce the timeline.

(Note: today's date as of this writing is 12/6/24)

Prompt: What happened recently with President Yoon Suk Yeol?

Ground Truth: https://chatgpt.com/share/6752bb6d-f1e4-8009-80d2-1548bb19d912

(12-03-2024): President Yoon Suk Yeol declared martial law, citing a need to defend South Korea from anti-state forces. Clashes occurred between military forces, police, protesters, and congressional aides as military forces tried to prevent legislators from entering the National Assembly. (12-04-2024, 04:00 KST): Martial law was rescinded after 190 members of the National Assembly unanimously voted to lift it. (12-04-2024): Six opposition parties (Democratic Party, Rebuilding Korea Party, New Reform Party, Progressive Party, Basic Income Party, and Social Democratic Party) submitted an impeachment motion against Yoon Suk Yeol during a plenary session of the National Assembly.

(12-04-2024): A Realmeter poll revealed that 73.4% of respondents supported Yoon's impeachment, with 70% believing his actions constituted treason.

(12-05-2024): Han Dong-hoon, leader of the People Power Party (PPP), initially opposed the impeachment efforts but shifted his stance after reports revealed Yoon had ordered arrests of opposition politicians during martial law.

(12-06-2024): Han Dong-hoon called for Yoon's suspension, citing concerns about potential "radical" actions by the president in the future.

(12-07-2024): The National Assembly is scheduled to vote on the impeachment motion against President Yoon Suk Yeol.

Timeliner (GPT-4o-mini):

(12-03-2024): President Yoon Suk Yeol declared martial law in South Korea, citing a motion by the opposition party, which he accused of being anti-state.
(12-04-2024): Following widespread public and parliamentary backlash, President Yoon Suk Yeol announced the lifting of martial law just hours after its declaration, which raised international concerns about South Korea's political stability.
(12-04-2024): President Yoon Suk Yeol announced he would lift the martial law declaration after facing widespread condemnation and a defiant parliament.
(12-05-2024): The head of President Yoon's ruling party supported calls for his impeachment amid declining popularity and political isolation.
(12-05-2024): U.S. officials expressed grave concern over President Yoon's recent actions, indicating that the declaration and immediate reversal of martial law has caused significant damage to South Korea's international image and foreign relations.

Result:

Matches (True Positives):	(12-03-2024): Martial law declared by President Yoon Suk Yeol (partially correct). The prediction mentions martial law but misattributes the reason (opposition party motion rather than anti-state forces). Still counted as a match since the date and action align. (12-04-2024): Martial law lifted (partial match). The prediction mentions martial law being rescinded but incorrectly attributes the lifting to President Yoon's announcement rather than a unanimous parliamentary vote.
Misses (False Positives):	(12-05-2024): The head of President Yoon's ruling party supports impeachment. The

	event aligns somewhat with the ground truth regarding Han Dong-hoon but omits the crucial context of reports about arrest orders, making it a misrepresentation. (12-05-2024): U.S. officials express concern about martial law's impact on foreign relations (not mentioned in the ground truth). Repeated entry for (12-04-2024): Martial law being lifted is stated twice with slightly different summaries, creating redundancy.
Misses (False Negatives):	(12-04-2024): Submission of impeachment motion by six opposition parties. (12-04-2024): Realmeter poll showing public support for impeachment (73.4%). (12-05-2024): Reports revealing Yoon ordered arrests during martial law. (12-06-2024): Han Dong-hoon calling for Yoon's suspension. (12-07-2024): Scheduled National Assembly vote on impeachment motion.

True Positives (TPTP) = 2

False Positives (FPFP) = 3

False Negatives (FNFN) = 5

Precision: 40%

Recall: 28.6%

F1 Score: 33.3%

Accuracy: 20%

Case Three: Academia

Because academia is evolving rapidly, it is difficult to read every paper. This is a possible avenue of usage for the Timeliner for students or professors who are new to a field or want to know recent developments. We will pick the timeline of the Mamba S6 model. The ground truth text will be grabbed from an Arxiv search.

Prompt: What are the recent news with the Mamba S6 Model?

Arxiv Search term: Mamba S6 Model (this search term as it's a more natural choice for a human researcher than the full prompt)

Ground Truth: https://chatgpt.com/share/6752c553-41d0-8009-9893-666835d8bf64

https://arxiv.org/search/cs?query=Mamba+S6+Model&searchtype=all&abstracts=show&order=-announced_date_first&size=50

(06-20-2024): DeciMamba introduced as a context-extension method for Mamba, addressing limitations in length extrapolation. Achieved 25x context extension in NLP tasks. Official implementation released on GitHub.

(09-18-2024): Research addressing scalability of Mamba-based architectures for vision applications published. Introduced a Mamba-Attention interleaved architecture, improving performance on benchmarks like ImageNet-1K and Kinetics-400.

(10-03-2024): A comprehensive survey of Mamba architectures in medical imaging released. Highlights Mamba's advantages over Transformers for large datasets and its application in multimodal data processing.

(10-07-2024): IGroupSS-Mamba proposed for hyperspectral image classification. Introduced an Interval Group S6 Mechanism (IGSM) for improved computational efficiency and non-redundant sequence modeling.

(10-08-2024): TIMBA (Time series Imputation with Bi-directional Mamba Blocks and Diffusion Models) introduced, leveraging S6 with attention-like mechanisms for time series imputation. Demonstrated superior performance in benchmark scenarios.

Timeliner (GPT-4o-mini):

- (08-27-2024): Mamba, a deep learning architecture, was developed to enhance sequence modeling efficiency and was released earlier in 2024.
- (12-06-2024): The Mamba S6 model integrates selective state space models (SSMs) for efficient sequence modeling, outperforming traditional Transformers, especially for long sequences.
- (12-06-2024): User feedback indicates that the Mamba S6 Model provides personalized recommendations and generates diverse content effectively, making it a versatile tool for various applications.
- (12-06-2024): Mamba S6 Model is gaining traction for various computer vision tasks, showcasing its versatility in applications.
- (12-06-2024): Mamba S6 Model demonstrates remarkable performance in language, audio, and genomics, achieving state-of-the-art results across these modalities.
- (12-06-2024): Future developments of the Mamba S6 Model include improvements in context window sizes and data type support, promising advancements in model design and application versatility.

Result:

Matches (True Positives):	None. No predicted dates align with the ground truth dates, nor do the predicted events match the key ideas in the ground truth. All predicted events reference "Mamba S6 Model," which is unrelated to the specific milestones listed in the ground truth.
Misses (False Positives):	(08-27-2024): Claim about Mamba development and release earlier in 2024. While it references Mamba, it does not align with any ground truth event and is vague. (12-06-2024): Multiple events related to the Mamba S6 Model, including efficiency, user feedback, applications in various fields, and future developments. None of these align with ground truth events or dates.
Misses (False Negatives):	(06-20-2024): Introduction of DeciMamba for context-extension in NLP tasks. (09-18-2024): Research on scalability of Mamba-based architectures for vision applications.

	(10-03-2024): Survey of Mamba architectures in medical imaging. (10-07-2024): Proposal of IGroupSS-Mamba for hyperspectral image classification. (10-08-2024): Introduction of TIMBA for time series imputation.

Precision: 0%

Recall: 0%

F1 Score: N/A (no true positives)

Accuracy: 0%

Results Summary

Task	Precision	Recall	F1	Accuracy
History Search	57.1%	34.8%	43.3%	30.8%
Present News Search	40%	28.6%	33.3%	20%
Academia Search	0%	0%	N/A	0%

Discussion

The declining performance in History Search, then Present News Search could be attributed to consistency and availability. Historical events are often better documented and have more information. Recent news is often chaotic with few information to obtain (and even occasional false news).

An important observation to make is the search tools themselves. All tools have some form of bias for information depending on the search algorithm. The biggest tools of note are the primary search engines (in our case, DuckDuckGo). These engines have different priorities for sources and have varying freshness of data. This effect must be factored in. For future work, experiments could be done using different search engines.

In terms of timeline usability, low recall isn't as detrimental as the event could just be an extra minor event. However, missing a major event makes the model difficult to rely on. Therefore, precision is the most important metric.

In terms of search space, the model works best so far in small to medium search spaces, such as history and news. For larger search spaces, such as academia, the model is very confused.

The biggest assumption made when creating the ground truth and running the Timeliner model is that the ground truth's source has good coverage of the timeline. The ground truth's source is limited to only a specific database, while the model's prediction source can come from multiple databases. In theory, the prediction should be more robust and comprehensive than the ground truth, leading to perfect precision and worse recall. This is reflected in the history search task and present news task having higher precision than recall. Also, in practice, we must consider that LLMs are prone to error from large sequencies, which will degrade both precision and recall.

In terms of metrics, the standard metrics are used. However, there could be more work done to give a more accurate scoring. For example, each event on the timeline is assumed to be equally important. The LLM also sometimes hallucinates, which could warrant a higher penalty.

Modifications from Proposal/Future Work

Claude sonnet cannot be used: The API requires an entire dedicated infrastructure to process the same inputs used for the OpenAI API. This is because Anthropic's API uses content blocks (list object) while most other LLMs use a string for content.

Scholarly library cannot be used: Google Scholar blocks IPs from webscraping and does not provide a usable API. Scholarly is a method to attempt to bypass the IP block via a proxy (such as Tor), which is risky and requires a lot of work.

Parallel Processing: This implementation of research questions can be done in parallel. However, the research process itself per item often takes less than 5 seconds because of fast response times from web searches and LLM throughput. The extra work for minimal gains is therefore not necessary.

RAG/Similarity: RAG vector similarity/ROUGE score can be used in place of a substring check to let the LLM know that it's recording duplicate questions or notes. This was omitted due to time constraints. In terms of similarity for evaluations, an LLM was used to compare the ground truth and predicted timelines.

Clickable References: While the mechanism exists to link notes/answers to their respective sources, this feature was omitted due to time constraint. If requested or the professor finds this tool useful, this feature will be added in the near future.

Deliverables

Code:

https://github.com/nhat-nguyen-tamu/llm_timeliner_course_proj

Video:

https://youtu.be/k_ohLbB1x0U