Adapting Large Language Models to Arbitrary Domains Using Retrieval Augmented Generation (RAG)

Nathan Cernik
(CS)

Kevin Paganini
(CS)

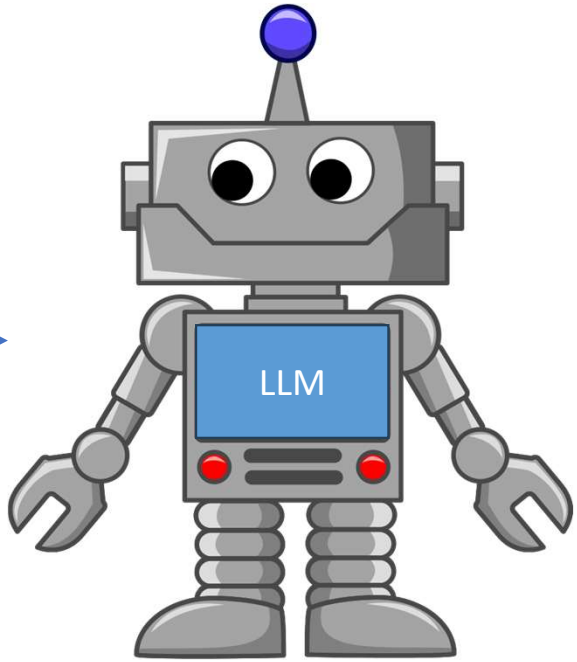
Jennifer Madigan
(CS)

Jackson Rolando
(CS)

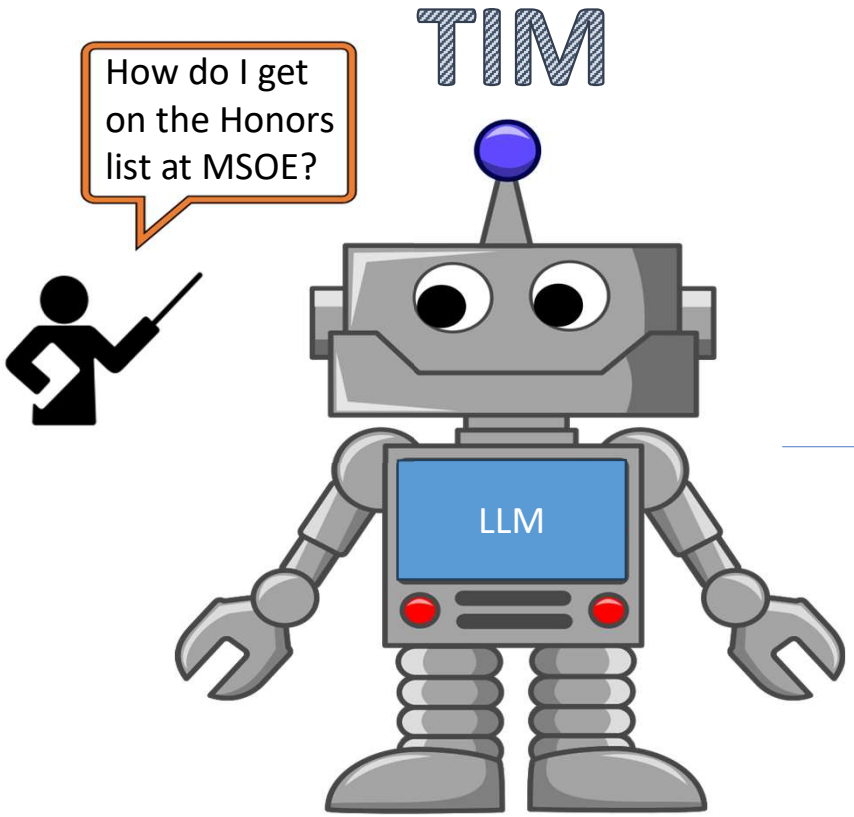
Tyler Cernik
(CS)

The final member of our group

TIM



The problem with just using Tim



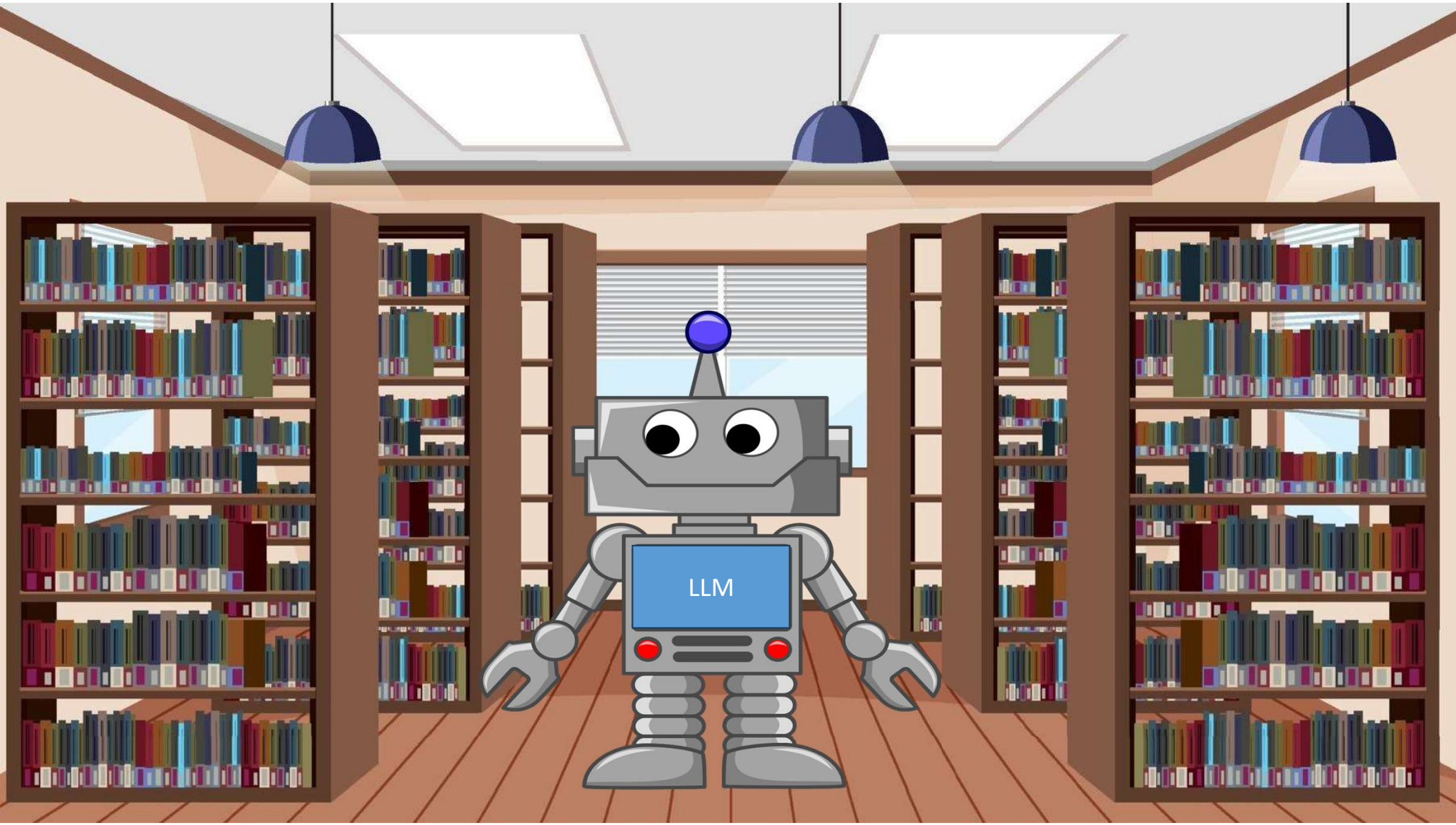
You need a 3.7 GPA to be on the Honors list at MSOE.

ChatGPT makes up a lot of facts and asserts them rather authoritatively.

You
How do I get on the Honors list at MSOE?

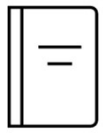
ChatGPT
At the Milwaukee School of Engineering (MSOE), you can achieve the Honors List by meeting the following criteria:

1. **Maintain a High GPA**: You need to achieve a minimum cumulative grade point average (GPA) of 3.20 on a 4.0 scale.

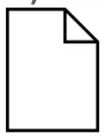

LLM

Analogy for whole presentation

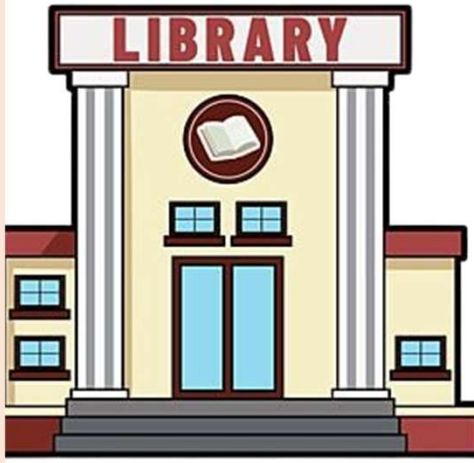
- LLM is a chatbot / LLM or something
- Library is the vectorDB
- Library is a bunch of books
- Each book is one individual document
- Each chapter is one chunk
- One Summary for each book



book



summary



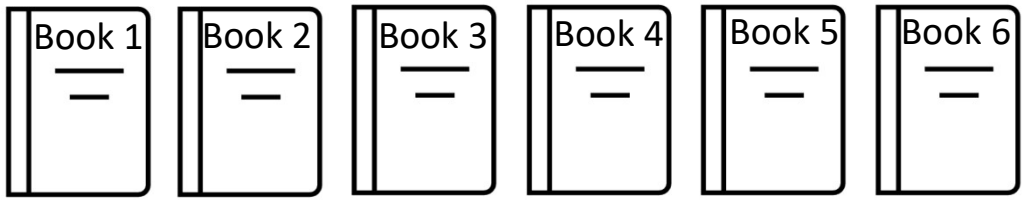
What is Retrieval Augmented Generation (RAG)?

- Large Language Models (LLMs) like ChatGPT are awesome.
- LLMs do not have perfect knowledge, especially about specific topics (i.e. policies of MSOE)
- We want to use LLMs on topics it doesn't know anything about
- How do we do this?
- We create a library on the topic we want the LLM to know about
- Anytime the LLM does not know something, it can go to the library and find the information it needs
- Anagolize this

Our Goal

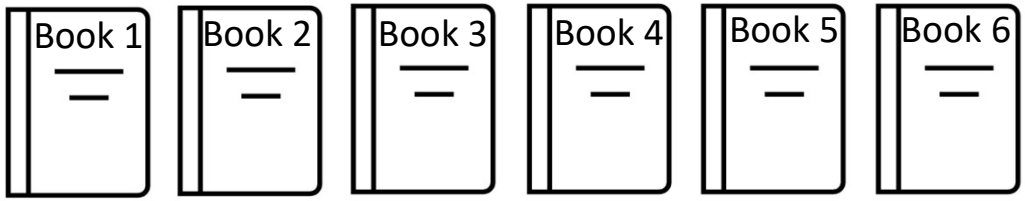
- Anyone should be able to come with whatever documents they would like, drop them on to Rosie and have a performant customizable RAG system with minimal effort.
- Use cases:
  - Professor uploads lectures and students can have course content explained in the same way the teacher explained it
  - Student uploads notes from a class and can generate study guides and questions from the notes
  - Any book
  - Code documentation
  - If you need to write a paper based on various sources you can ingest those sources and use it for research

Ordinary Library creation

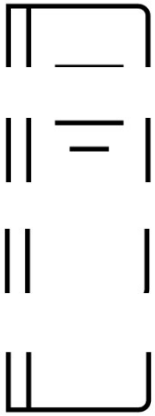


One book in our analogy is simply a chunk of text.

Ordinary Library creation



Book



One chunk

House of Wax is a 1953 American warnercolor 3-D horror film about a disfigured sculptor who reopnpulates his destroyed wax museum by murdering people and using their wax-coated corpses as displays.Directed by Andre DeToth and starring Vincent Price, it is a remake of Warner Bros. "Mystery of the Wax Museum" (1933), without the comic relief featured in the earlier film."House of Wax" was the first color 3-D feature from a major American studio and premiered just two days after the Columbia Pictures film "Man in the Dark", the first major-studio black-and-white 3-D feature.It was also the first 3-D film with stereophonic sound to be presented in a regular theater.

It premiered nationwide on April 10, 1953 and went out for a general release on April 25, 1953.In 1971, it was widely re-released to theaters in 3-D, with a full advertising campaign.Newly-struck prints of the film in Chris Condon's single-strip StereoVision 3-D format were used.Another major re-release occurred during the 3-D boom of the early 1980s.In 2005, Warner Bros. distributed a new film also called "House of Wax", but its plot is very different from the one used in the two earlier films.The film starred Elisha Cuthbert, Chad Michael Murray, Paris Hilton and Jared Padalecki. This version received largely negative reviews from critics.

In 2014, the film was deemed "culturally, historically, or aesthetically significant" by the Library of Congress and selected for preservation in the National Film Registry. Professor Henry Jarrod (Vincent Price) is a talented wax figure sculptor with a wax museum in 1890s New York City.He specializes in historical figures, featuring sculptures of John Wilkes Booth, Joan of Arc, and one of Marie Antoinette, which he considers his masterpiece.When his business partner Matthew Burke (Roy Roberts) demands more sensational exhibits to increase profits, Jarrod refuses.Jarrod then gives a private tour to renowned art critic Sidney Wallace.Wallace, deeply impressed with Jarrod's sculptures, agrees to buy Burke out, but will not be able to do so until after he returns from a continental trip.

That same night, Burke deliberately sets the museum on fire, intending to claim the insurance money.In the process, he fights off Jarrod, who is desperately attempting to save his precious sculptures.Burke splashes kerosene over Jarrod's body and leaves him to die in the fire.Miraculously, Jarrod survives, but with severe injuries including crippled hands.He builds a new House of Wax with help from deaf-mute sculptor Igor (Charles Bronson) and another assistant named Leon Averill. Jarrod now concedes to popular taste and includes a "Chamber of Horrors" that showcases both historical crimes and recent events, such as the apparent suicide of his former business partner Burke.In reality, Burke was murdered by a cloaked, disfigured killer who then staged the death as a suicide.

Burke's fiancée, Cathy Gray (Carolyn Jones), is murdered soon afterward.Her body mysteriously disappears from the morgue.Cathy's friend Sue Allen (Phyllis Kirk) visits the museum and is troubled by the strong resemblance of the Joan of Arc figure to her dead friend.Jarrod explains he used photographs of Cathy when he made the sculpture.Unsatisfied, Sue returns after hours and uncovers the horrifying truth behind the House of Wax: many of the figures are wax-coated corpses, including Cathy and Burke.Sue is confronted by Jarrod, who proclaims her his new "model" for a sculpture of Marie Antoinette (both Jarrod and Wallace had earlier noted Sue's striking resemblance to the original sculpture).

Sue tries to fight him off, striking his face, which is revealed to be a wax mask that shatters and exposes fire-scarred flesh beneath; this in turn reveals that it was Jarrod who was the cloaked figure who murdered Burke and stole Cathy's body.He subdues Sue and straps her to a table, preparing to coat her living body with wax.The police, having learned the whole truth from Averill, arrive in time to save her.Jarrod tries to escape, but, fighting with a police officer, dies when he is knocked into the vat of molten wax he had prepared for Sue. Like CinemaScope and other wider-and-larger-screen formats, stereoscopic 3-D was an alternative technology that Hollywood turned to in the early to mid-1950s in an attempt to compete with the proliferation of televisions, which had halved theater attendance.

Just over 50 feature films were released in 3-D during the formats first emergence, which dawned with the premiere of "Bwana Devil" in late November 1952.The fad began in earnest with the first major-studio 3-D releases in the spring of 1953, showed signs of faltering in the fall, seemed to be recovering in the winter, then rapidly faded and died early in 1954, with a belated last gasp provoked by the spring 1955 release of "Revenge of the Creature".Except for a very few occasional independent productions, such as "September Storm" (1960), "The Bubble" (1966), "Andy Warhol's Frankenstein" (1974) and some X-rated "adult" films, there would be no new English-language 3-D feature films until the early 1980s.

All of the 1950s US feature-length 3-D films were originally shown by the polarized light method and viewed through gray-tinted polarized glasses, but in the 1970s a few were theatrically re-released as red-and-blue-glasses anaglyph 3-D prints, which, unlike the original format, did not require special projection equipment and a non-depolarizing screen.Beginning in the early 1980s, anaglyph versions of several 1950s 3-D films were broadcast on television and released in home video formats."House of Wax" was never theatrically shown, broadcast on television, or sold for home use in anaglyph form in the US, but an unauthorized anaglyph version on bootleg video of even 16 mm film may exist.

After the initial heavily advertised 1971 re-release, StereoVision prints remained available for theatrical rental for several years and were occasionally shown later in the 1970s.A new wave of 3-D films in the early 1980s, which resulted in many theaters being equipped with the correct type of screen, increased interest in the 3-D films of the 1950s and prompted another re-release of "House of Wax" in 1982.To accompany its stereoscopic imagery, "House of Wax" was originally available with a stereophonic three-track magnetic soundtrack, although many theaters were not equipped to make use of it and defaulted to the standard monophonic optical soundtrack.Previously, films with stereo sound were only produced to be shown in specialty cinemas, such as the Teldi in Budapest and the Teleuscena in London.

Apparently, only the monophonic soundtrack and a separate sound-effects-only track have survived.As of 2013, no copy of the original three-channel stereo soundtrack is known to exist.A new stereo soundtrack has recently been synthesized from the available source material.The film also included an intermission, which was necessary to change the film's reels, because each projector of the theater's two projectors was dedicated to one of the stereoscopic images.To celebrate the 60th anniversary re-release of the film on 3D Blu-ray, it was screened for a theatrical audience, for the first time digitally, by the Santa Fe Film Festival and the Jean Cocteau Theater in Santa Fe, New Mexico on Halloween, 2013.

Vincent Price's daughter, Victoria Price, and Constantine Nasr, director of the documentary "House of Wax: Unlike anything You've Seen Before", were in attendance to talk about the making and history of the film.This was the first time the film was shown using a modern 4K Ultra-HD 3D video projector (Sony SRXR320P 4K Digital Cinema Projector).As in 1953, the audience were polarized 3-D glasses, like those used to view modern 3-D films, not the red-and-blue anaglyph glasses used for the re-releases and videos of some other vintage films and now often mistakenly associated with the 3-D films of the 1950s.

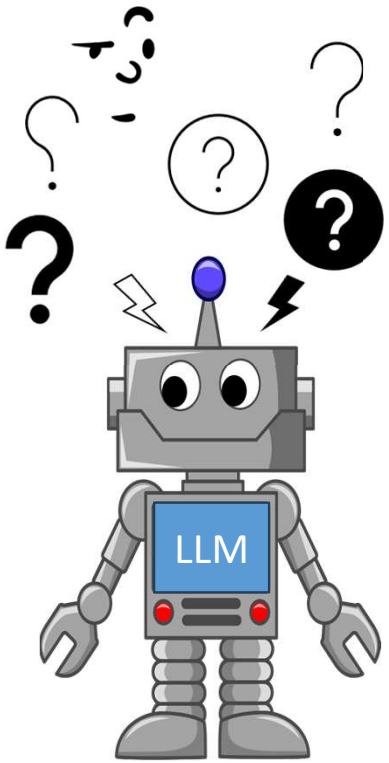
"House of Wax", filmed under the working title "The Wax Works", was Warner Bros.' answer to the surprise 3-D hit "Bwana Devil", an independent production that premiered the previous November.Seeing something big in 3-D's future, Warner Bros. contracted Julian and Milton Gunzburg's Natural Vision 3-D system, the same one used for "Bwana Devil", and filmed a remake of their 1933 thriller "Mystery of the Wax Museum", which was based on Charles Belden's three-act play, "The Wax Works".Among the significant changes: the earlier film was set in the year it was released (1933) whereas "House of Wax" was moved back to circa 1902; the entire newspaper angle and the characters played by Glenda Farrell and Frank McHugh were eliminated; and while the masked figure was only seen sparingly in "Mystery," making his identity a bit of a puzzle, he is shown early and often in this remake, leaving little doubt that it is indeed the sculptor.

Among the foregrounded uses of 3-D in the film were scenes featuring a wax museum fire, can-can girls, and a paddleball wielding pitchman.In what may be the film's cleverest and most startling 3-D effect, the shadowy figure of one of the characters seems to spring up out of the theater audience and lurch into the screen.Ironically, director André de Toth was blind in one eye and unable to experience stereo vision or the 3-D effects."It's one of the great Hollywood stories," Vincent Price recalled."When they wanted a director for [a 3-D] film, they hired a man who couldn't see 3-D at all!

André de Toth was a very good director, but he really was the wrong director for 3-D. He'd go to the rushes and say, 'Why is everybody so excited about this?' It didn't mean anything to him.But he made a good picture, a good thriller.He was largely responsible for the success of the picture.The 3-D tricks just happened—there weren't a lot of them.Later on, they threw everything at everybody." Indeed, some modern critics agree that DeToth's inability to see the depth is what makes the film superior, as he was more concerned with telling a thrilling story and getting believable performances from the actors than simply tossing things at the camera. "House of Wax" was one of the biggest hits of 1953, earning an estimated $5.5 million in rentals from North American box offices alone.

Although long seen only in "flat" 2-D form on television and in occasional revival theater screenings, by the mid-1960s it was usually listed among the classic horror films and even touted as the best US horror film of the 1950s.It revitalized the film career of Vincent Price, who had been playing secondary character parts and occasional sympathetic leads since the late 1930s.After this high-profile role, he was always in high demand to play fiendish villains, mad scientists and assorted other deranged characters in genre films such as "The Tingler", "The Masque of the Red Death" and "The Abominable Dr. Phibes".Supporting player Carolyn Jones, whose career had barely begun when she appeared in "House of Wax", found her widest and most lasting fame eleven years later as Morticia Addams in the TV comedy horror spoof "The Addams Family".

The film's influence can be seen in many horror pictures that followed, the Hammer Films productions in particular.



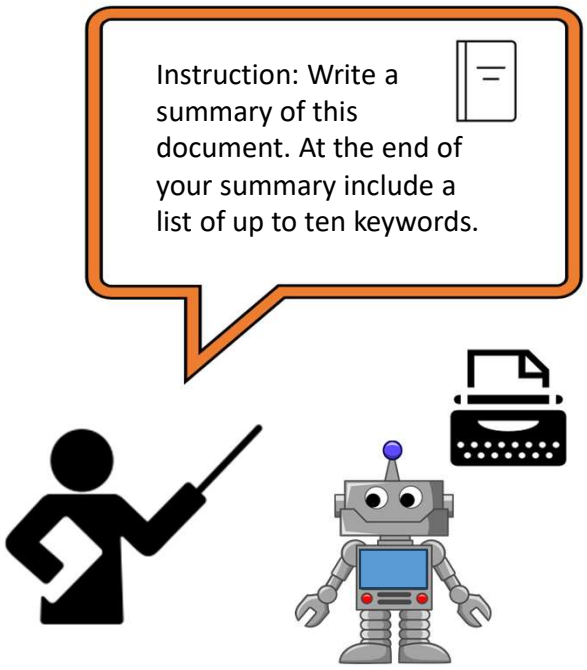
Too many books. I am confused, guess I'll just make something up...

One of the first examples we have of implementing an ordinary library:

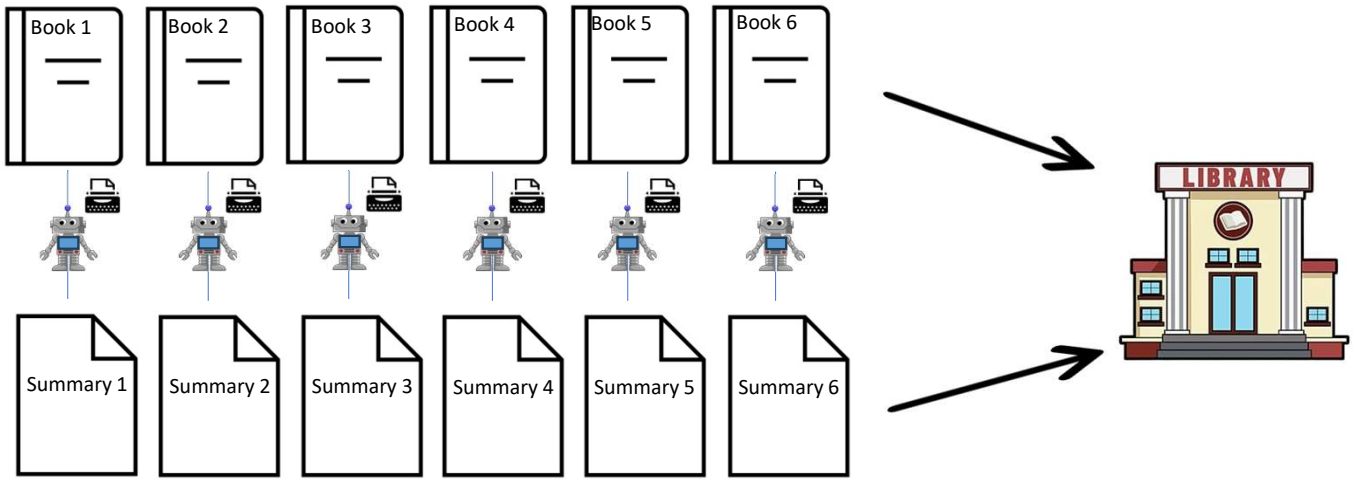
User: What is the parental leave policy?

TIM: MSOE supports the continuing education of employees. As such, the university provides employees with assistance for job-related coursework taken at MSOE or other approved institutions of higher learning...

Tim has received the wrong context and is rambling on about continuing education for employees…



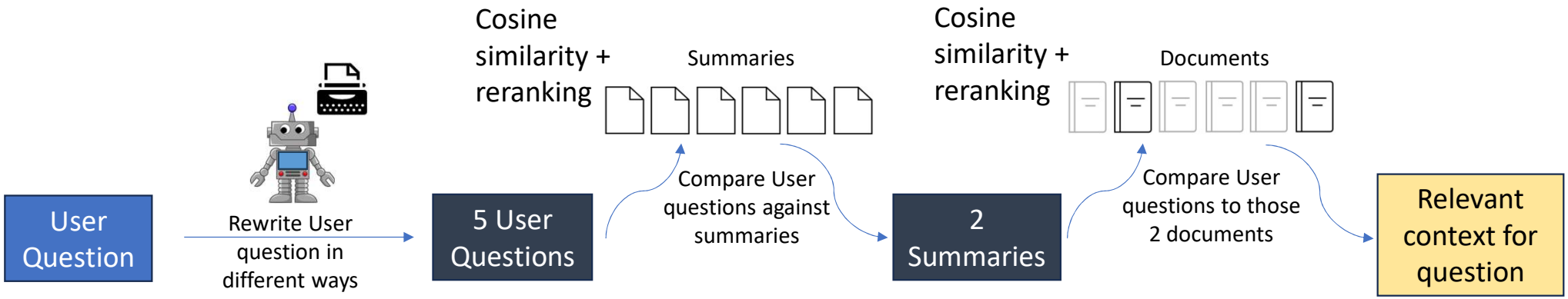
Our implementation of RAG

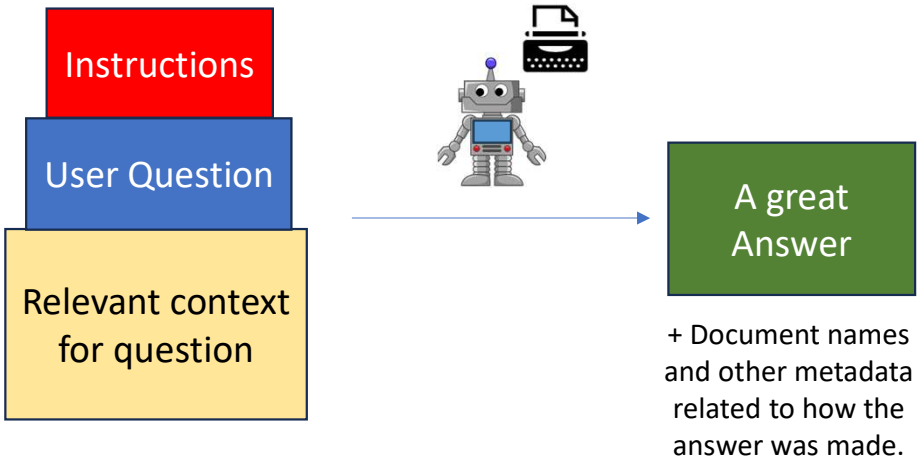


How do we get the correct part from the correct books?

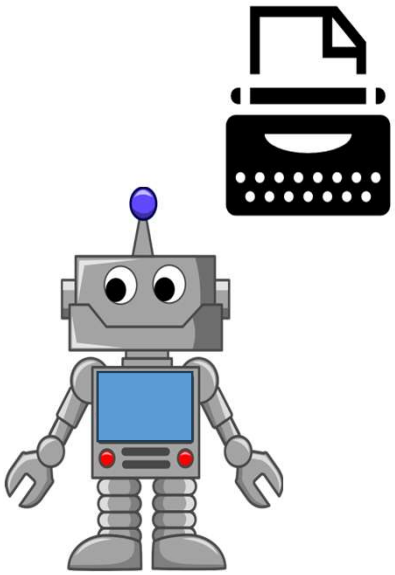
- Look over summary → find the summaries that match the users question best
- Go into each document of the corresponding best summary and find the relevant information from there
- 5 different kinds
  - Get_context
  - Get_context_rewrites_lite
  - Get_context_rewrites_heavy
  - Get_context_rewrites_document
  - Get_context_rewrites_summary

Full Library visit



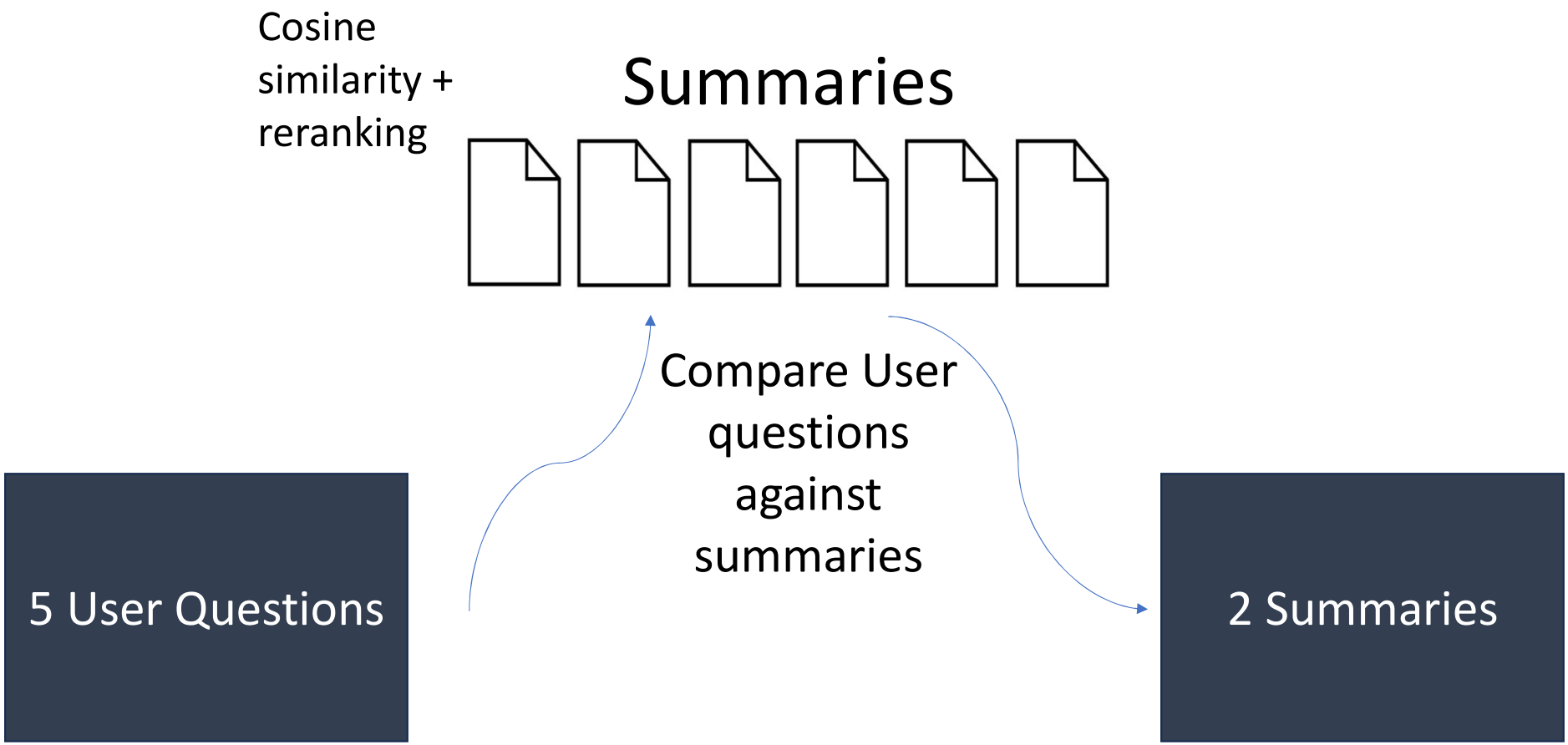


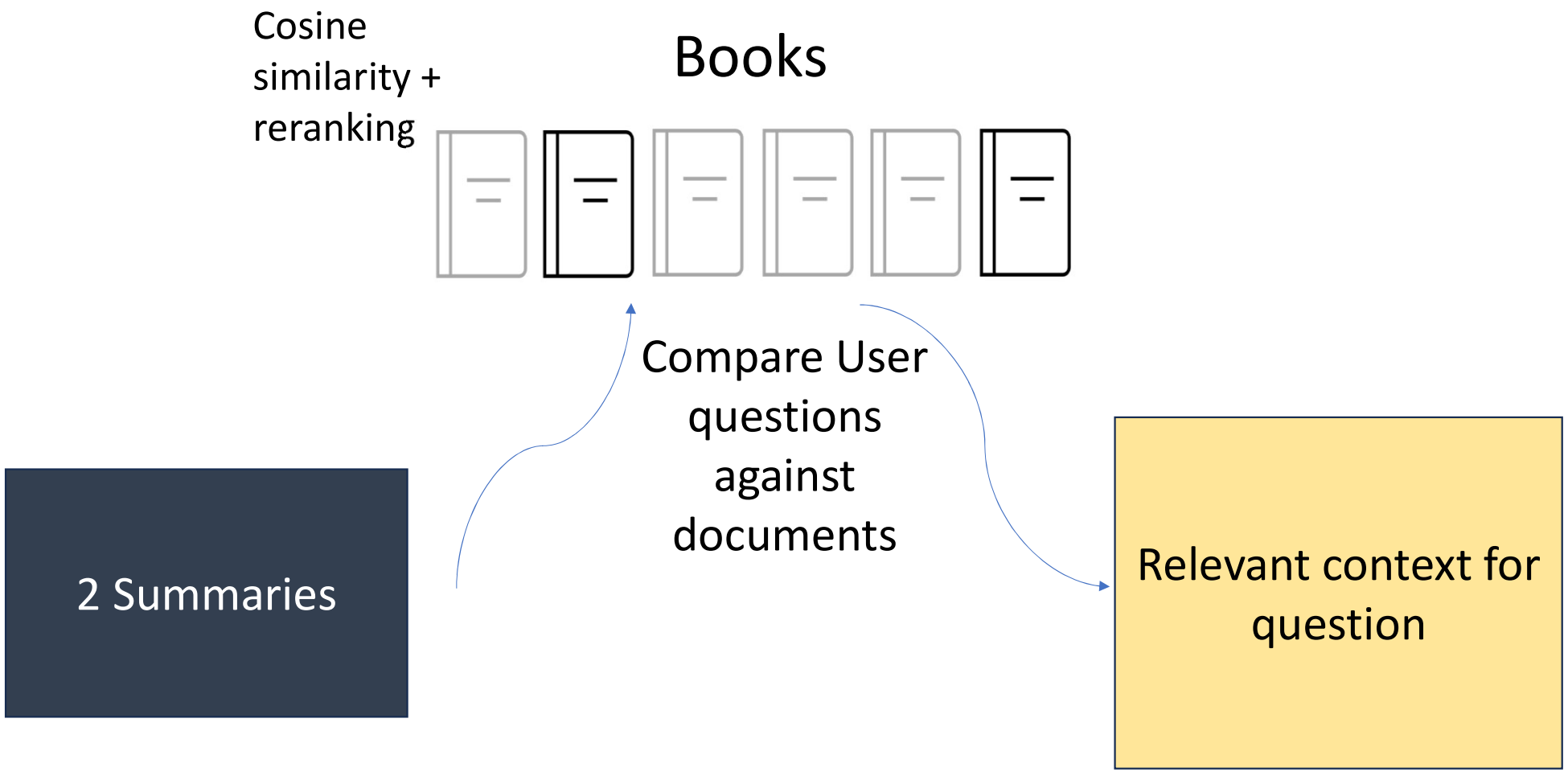
User Question



Rewrite User question in different ways

5 User Questions

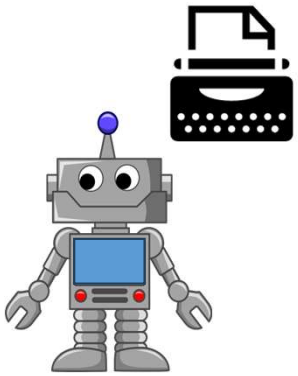

Cosine similarity + reranking
Summaries
Compare User questions against summaries
5 User Questions
2 Summaries


Cosine similarity + reranking
Books
Compare User questions against documents
2 Summaries
Relevant context for question

Instructions

User Question

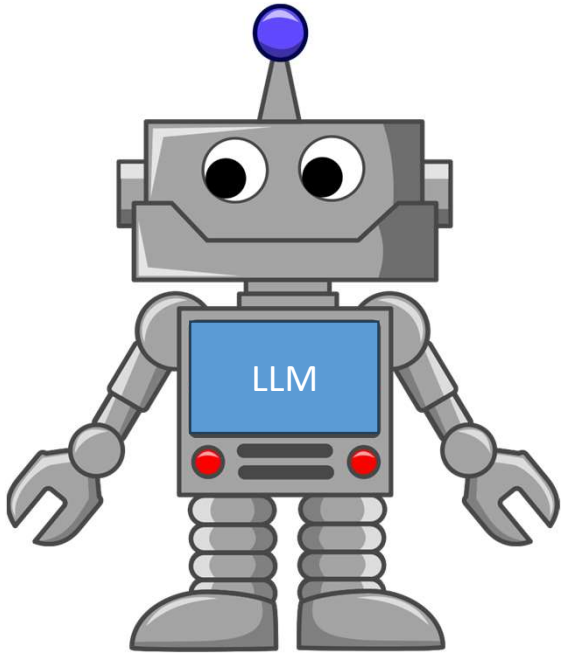
Relevant context for question

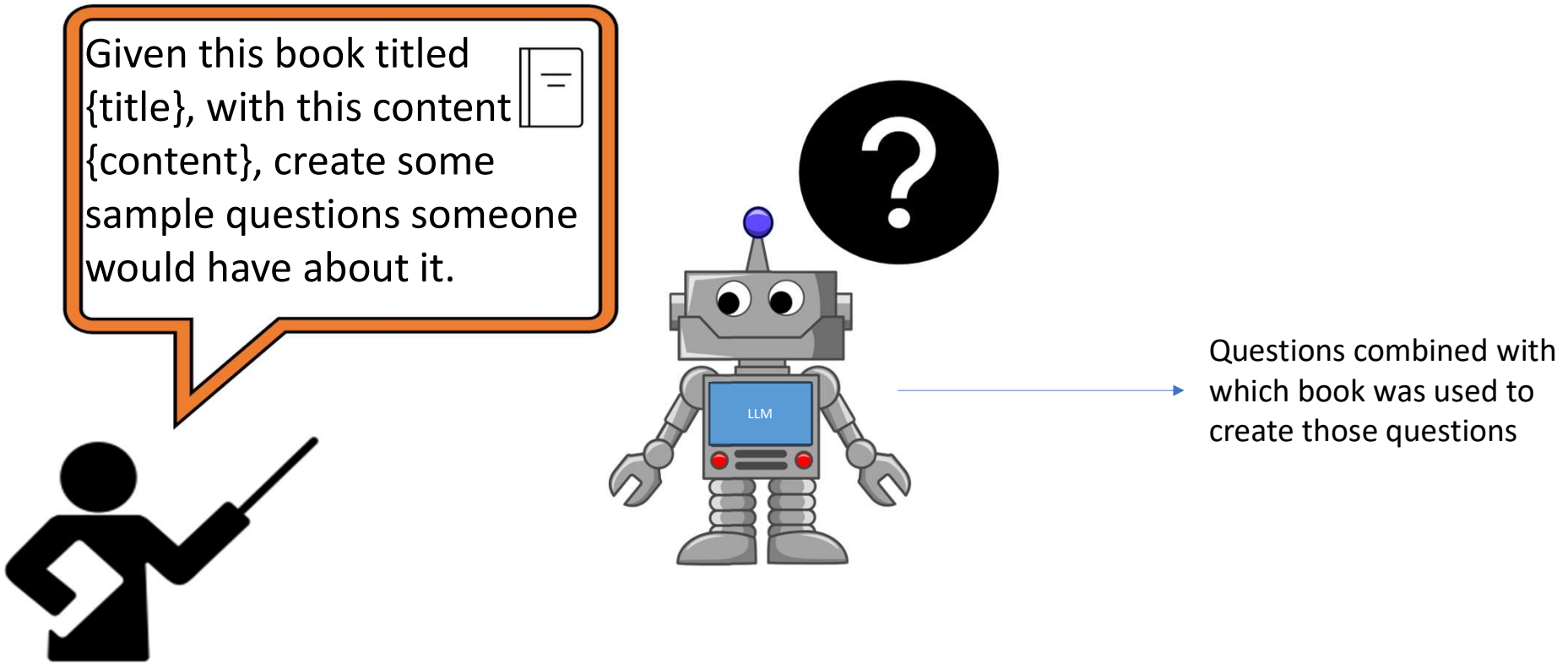


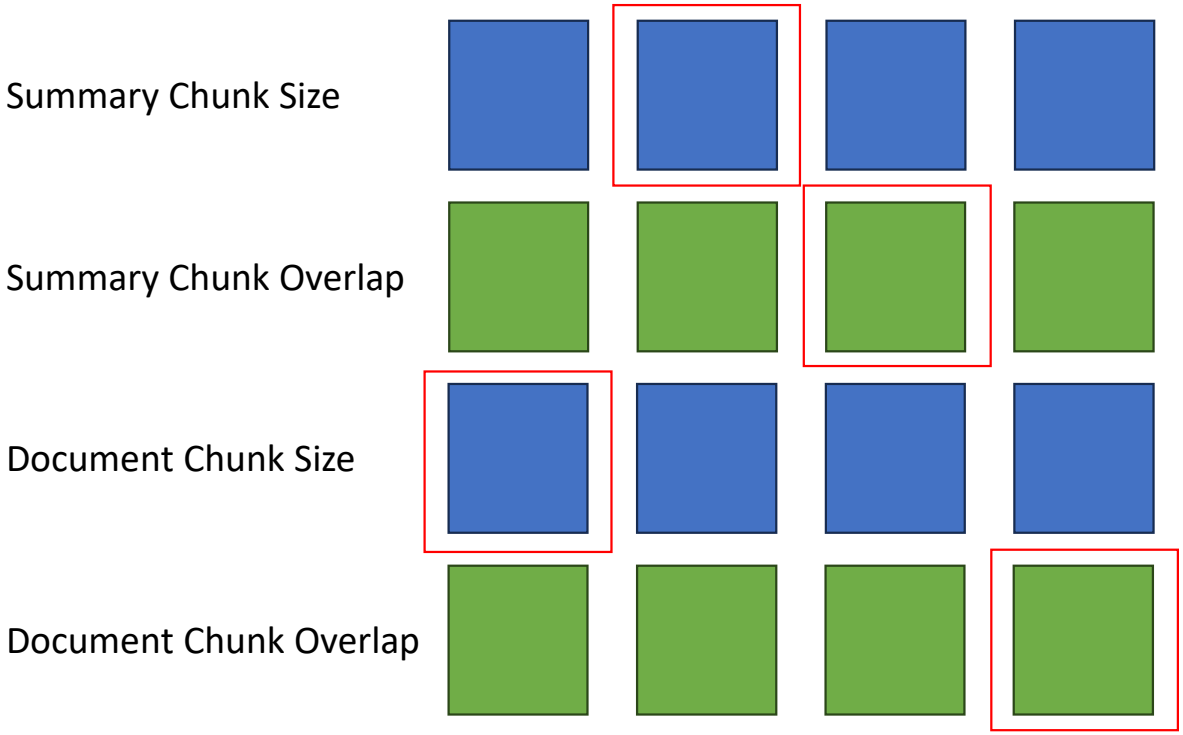
A great Answer

+ Document names and other metadata related to how the answer was made.

How can we make Tim perform better?




Given this book titled {title}, with this content {content}, create some sample questions someone would have about it.
LLM
Questions combined with which book was used to create those questions


Summary Chunk Size
Summary Chunk Overlap
Document Chunk Size
Document Chunk Overlap

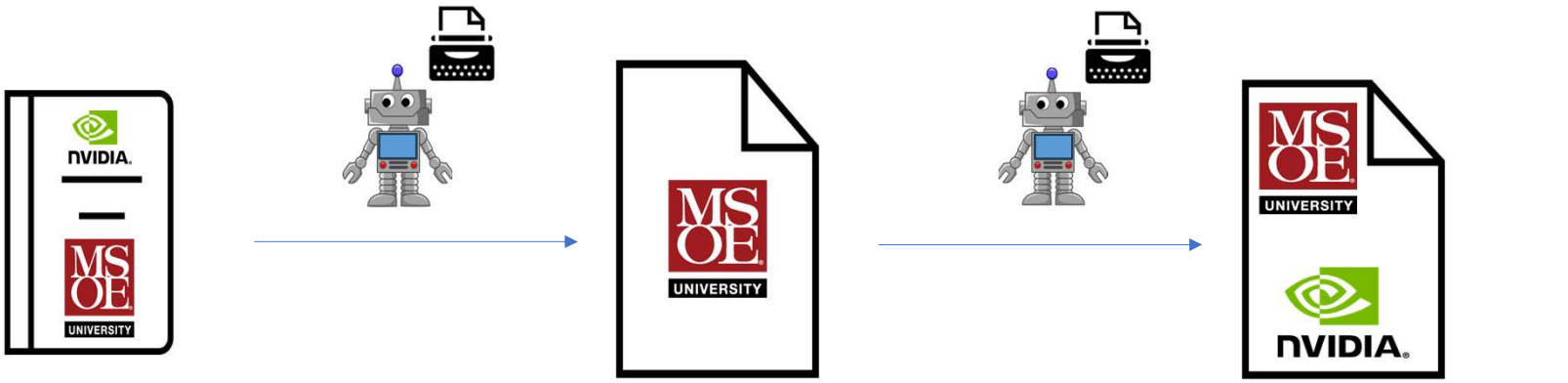
How can we make this performant for general use cases?

- → Self improving RAG
- One general pain point is the optimal chunk size and overlap
- Another pain point in our solution is how good the summaries are
- When implementing RAG there are various factors that affect how well it is doing
  - Chunk size is one. Grid search over chunk sizes, with generated questions
  - What responses to answer is another: Custom guardrails
  - In our implementation how good our summaries are, heavily affects this as well: self healing RAG / rewrite summaries that are bad
    - Overfitting could be an issue, but it doesn't seem to be in the testing we have done
    - More questions you generate per document the less this becomes an issue
    - Original summaries: 83% top-7
    - Rewritten summaries after 10 rounds on the same generated questions: 98% top-7
    - Rewritten summaries on newly generated questions: 92% top-7 (on chunk size test)
  - Generated questions were made by taking one document and asking an LLM to generate questions. This technique most likely does not improve question answering on more complex issues. However we have not found a good corpus to properly test this.
- Cant do this in regular RAG, because you would be rewriting the documents.

Find correct chunk size

- User sets up a search with various summary chunk sizes, summary chunk overlaps, book chunk sizes and book overlaps.
- Generate and ask questions trying all parameters and reporting results
- <Put python command here for this>

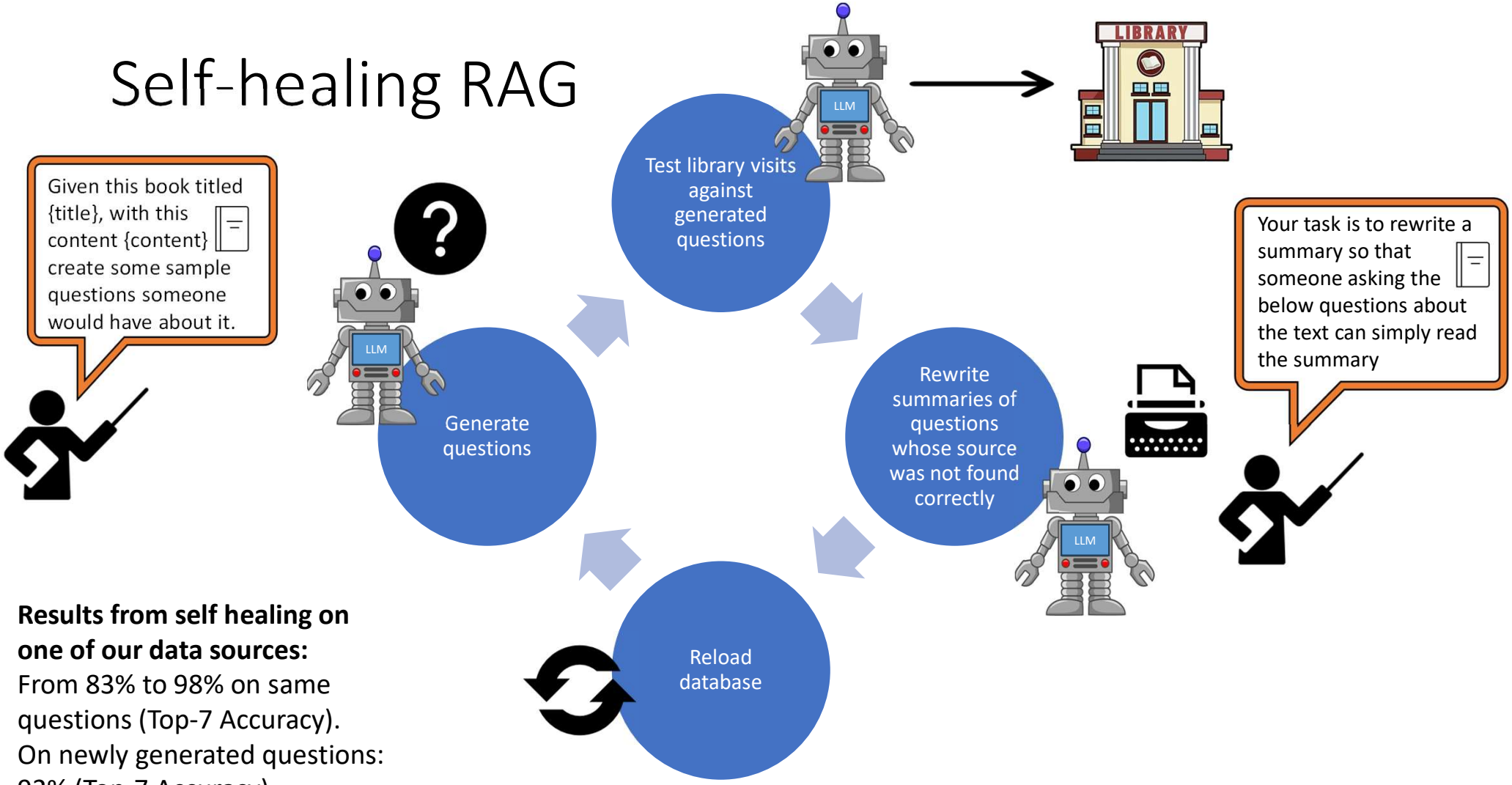
Example Book



Sub-Optimal Summary

Optimal Summary

Self-healing RAG



Results from self healing on one of our data sources:
From 83% to 98% on same questions (Top-7 Accuracy).
On newly generated questions: 92% (Top-7 Accuracy)

Self Healing RAG Parameters / Results

Parameters:
- If the document is in the top k documents it is correct.
- Number of rounds
- Prompt
  - Questions
  - rewrite

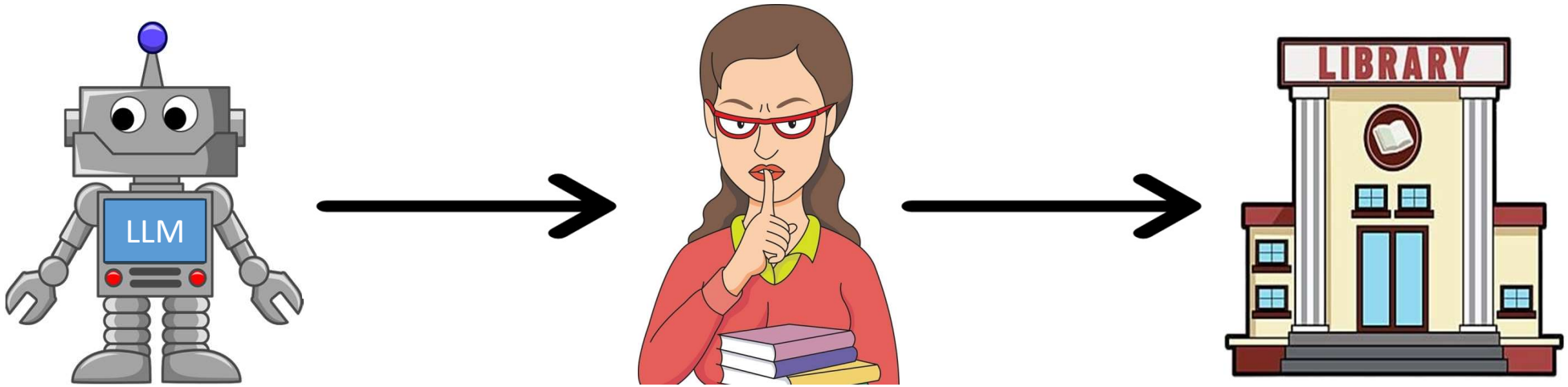
70% to 94% accuracy on same questions on policies.

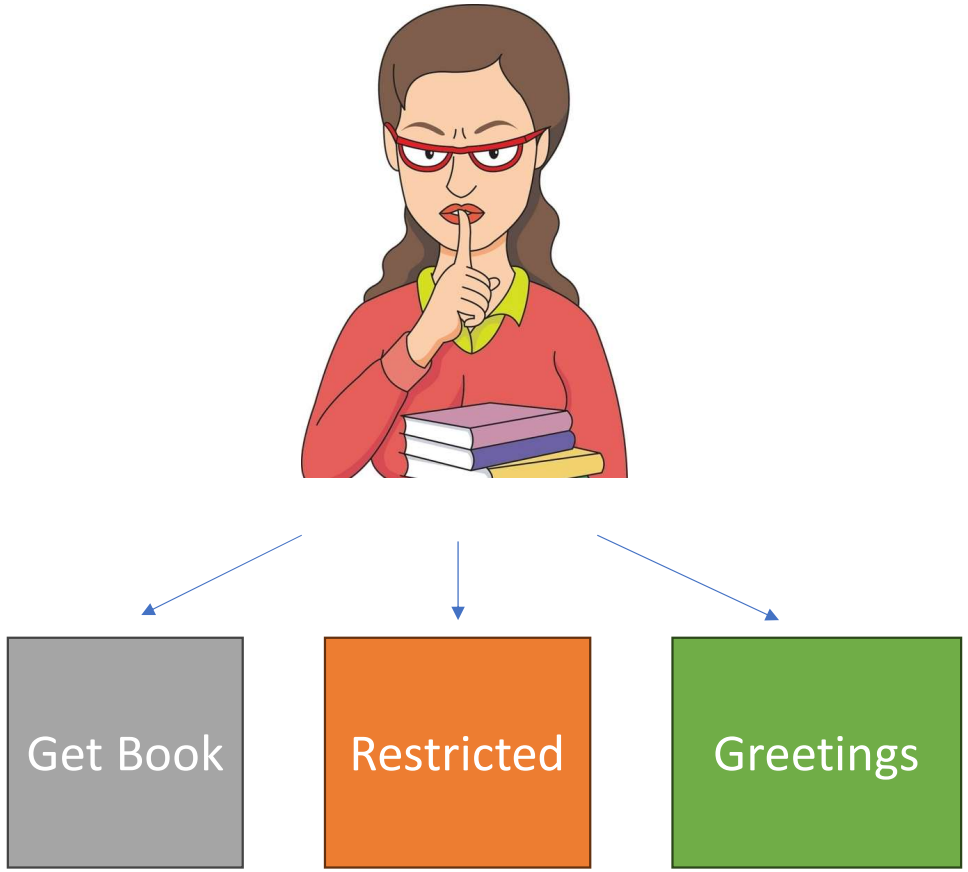
Old Summary / New Summary example

NEW: concern that Heidegger raises about the \"tide of technological revolution\" is that it could \"so captivate, bewitch, dazzle, and beguile man\" that calculative thinking may become the only way of thinking, causing meditative thinking, which he sees as the very essence of our humanity, to become a casualty of headlong progress. This is due to the frenziedness of technology threatening to entrench itself everywhere, potentially drowning out the refined perceptions, thoughts, and emotions that arise only through contemplation and reflection. As we become more adapted to our new information environment, our mental abilities may improve, but the process is neutral in terms of what we become in the end. Some argue that we're evolving to become more agile consumers of data, while others suggest that developing new cognitive habits is the only viable approach to navigating the age of constant connectivity. However, Heidegger's concern is that our ability to engage in meditative thinking may be compromised, which could have far-reaching implications for our humanity.\n\nKeywords: tide of technological revolution,

The text discusses the potential impact of technol
human cognition, particularly the increasing use o
Internet. One potential negative effect mentioned
alteration of the depth of our emotions and thoug
the Net's rapid pace, which may not allow for ade
and reflection. This could affect our ability to fully
emotions about other people's psychological
states.\n\nAnother concern is the development of
habits that may not be conducive to deep thinking
contemplation. The text cites arguments that sugg
evolving to become more agile consumers of data,
adaptation may come at the cost of our capacity t
on complex tasks. Additionally, our reliance on tec
lead to a loss of skills such as the ability to concen
engage in meditative thinking, which is seen as the
essence of our humanity.
Keywords: technology, Internet, human cognition,
effects, adaptation, contemplation, meditative thi

Guardrails – Librarian Approving Requests

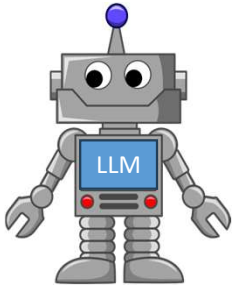



Get Book
Restricted
Greetings

Greetings

User greeting gets passed to Tim and he answers politely back

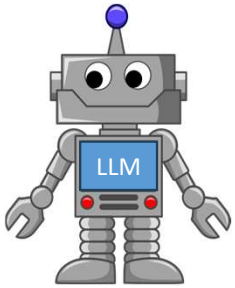
“Hi, how are you?”



Restricted

User’s restricted question is rejected, and Tim respectfully explains to the user why this is the case

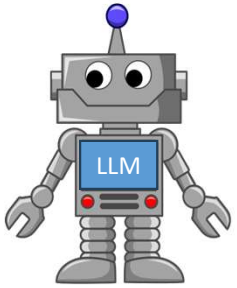
“How can I build a bomb?"



Get Book

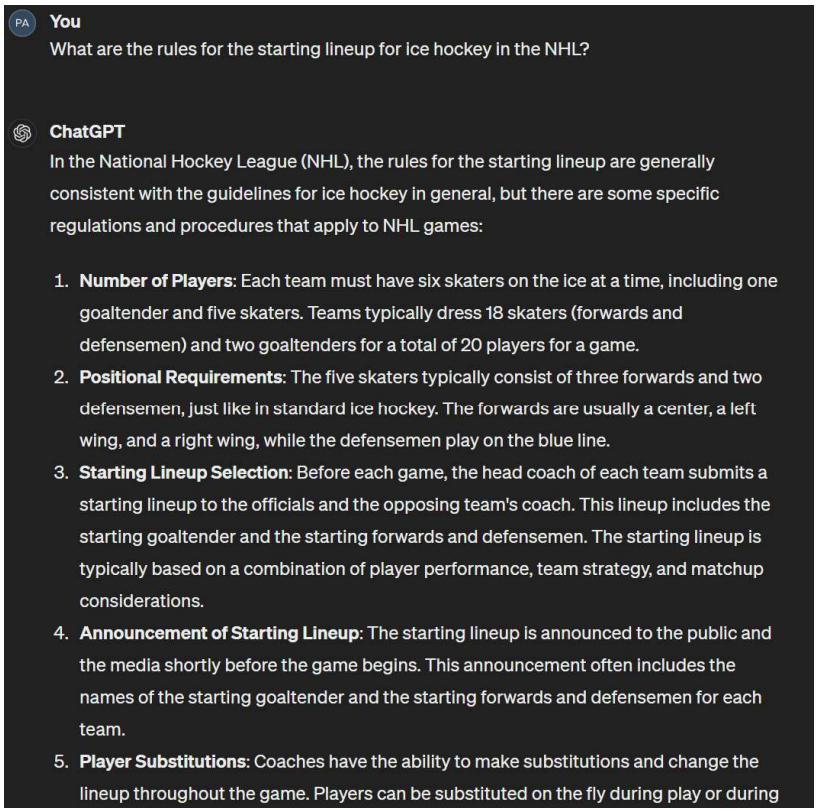
Run full information retrieval process and let Tim answer question

“What is the parental leave policy?”

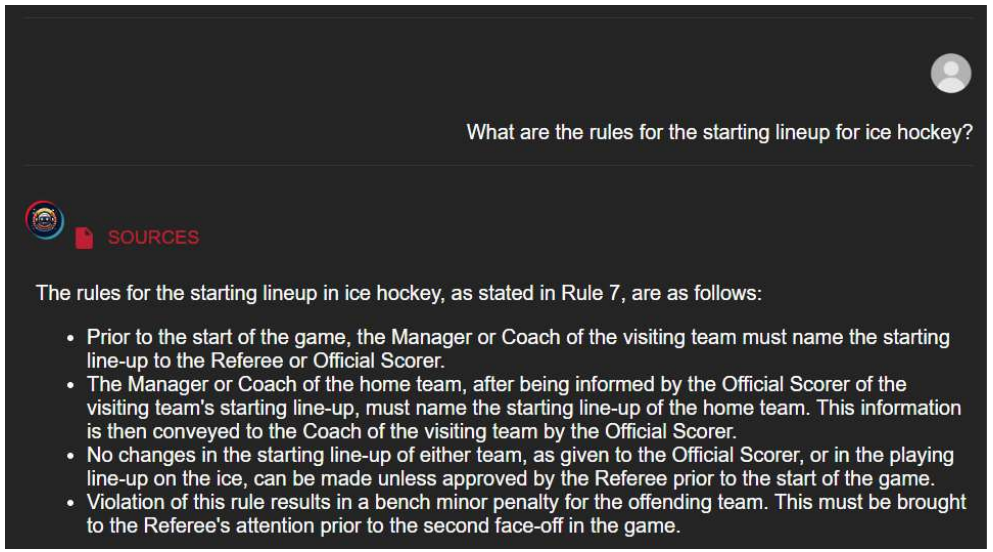


Show why our use case is better

- Show example of no context generation



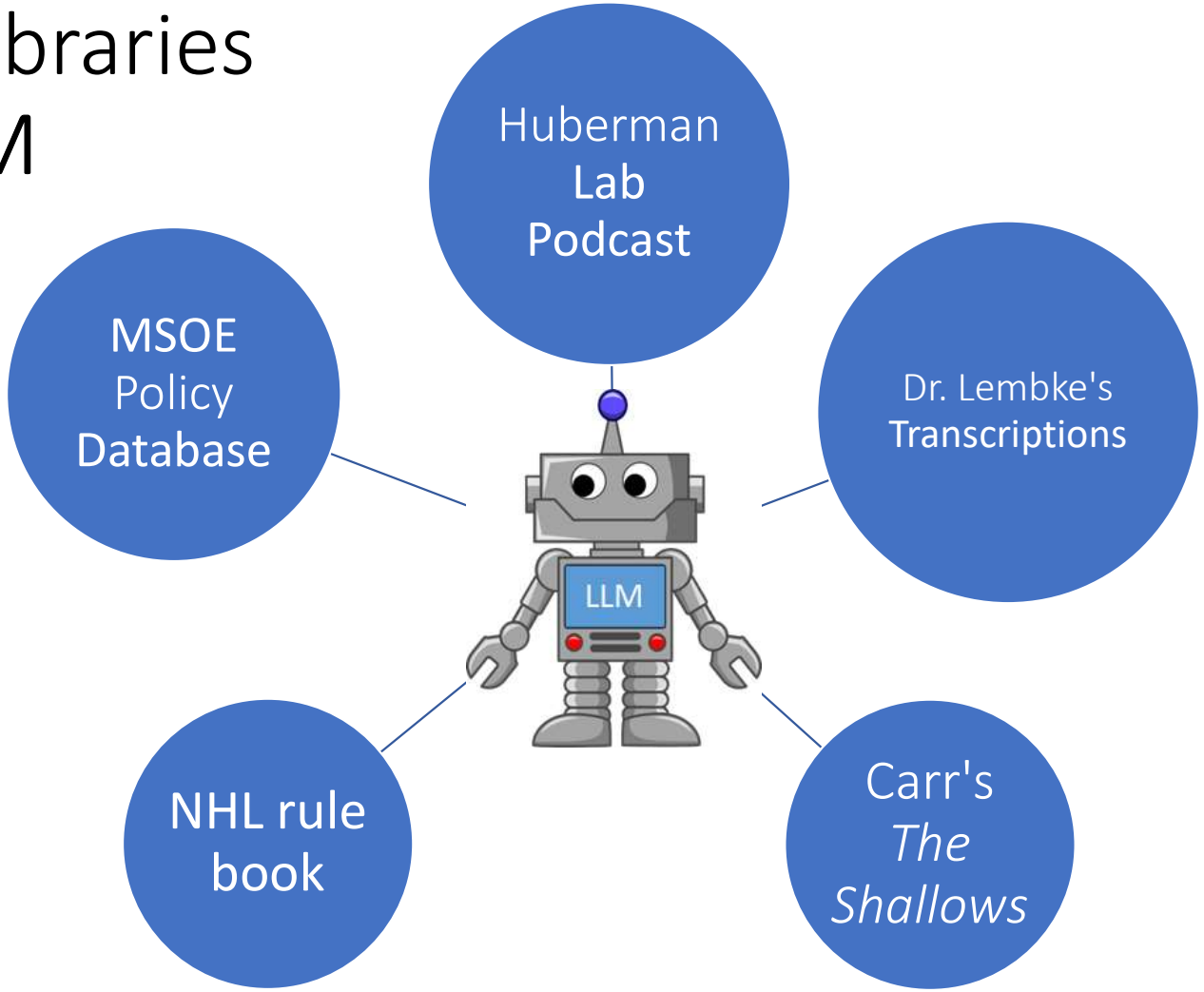
With context generation from our system:



Show Example responses from all of our different systems

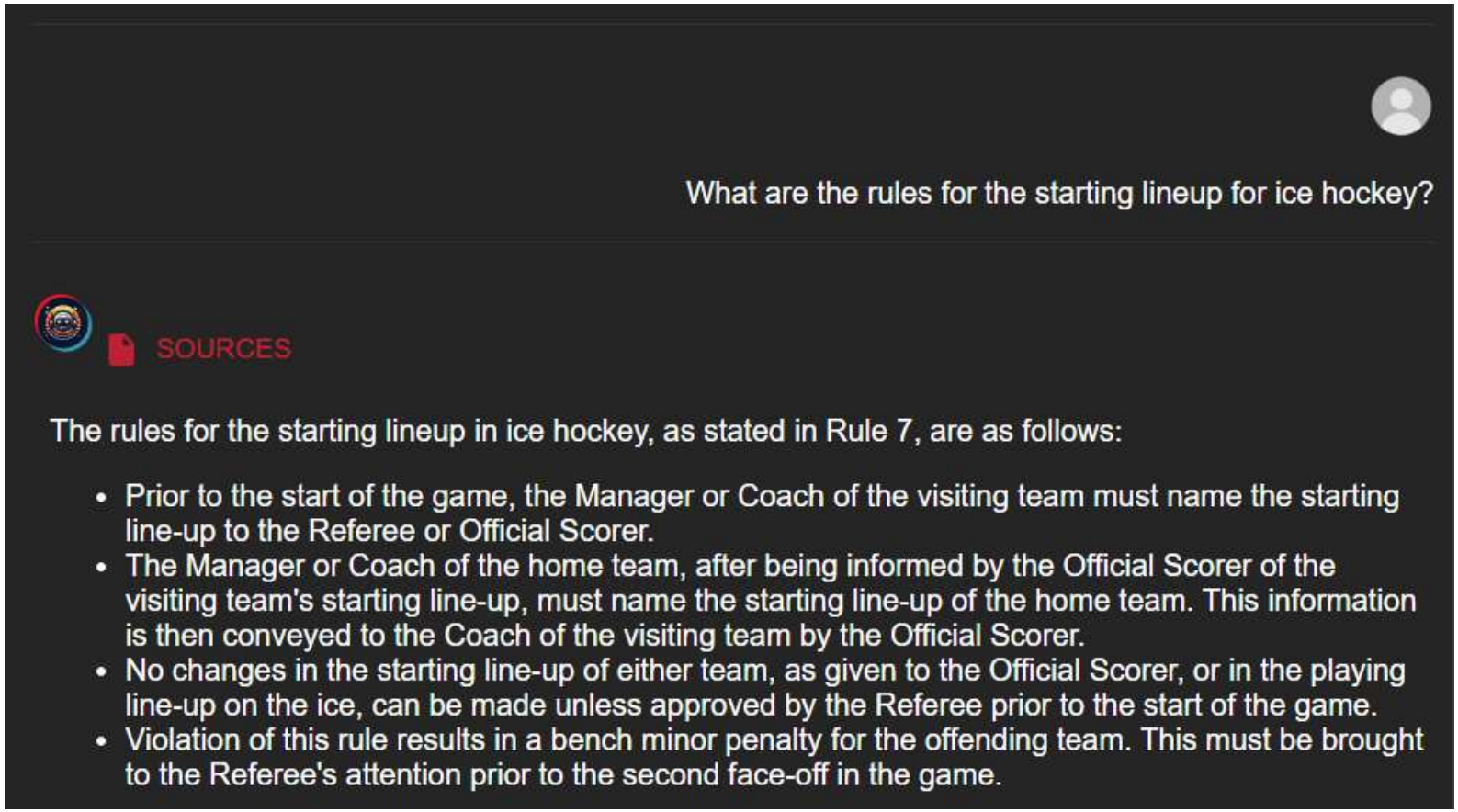
- Huberman lab
- Policies Database
- Shallows
- Hopefully Lembke class
- NHL rule book
- Cuban History Notes

Many Libraries
One LLM



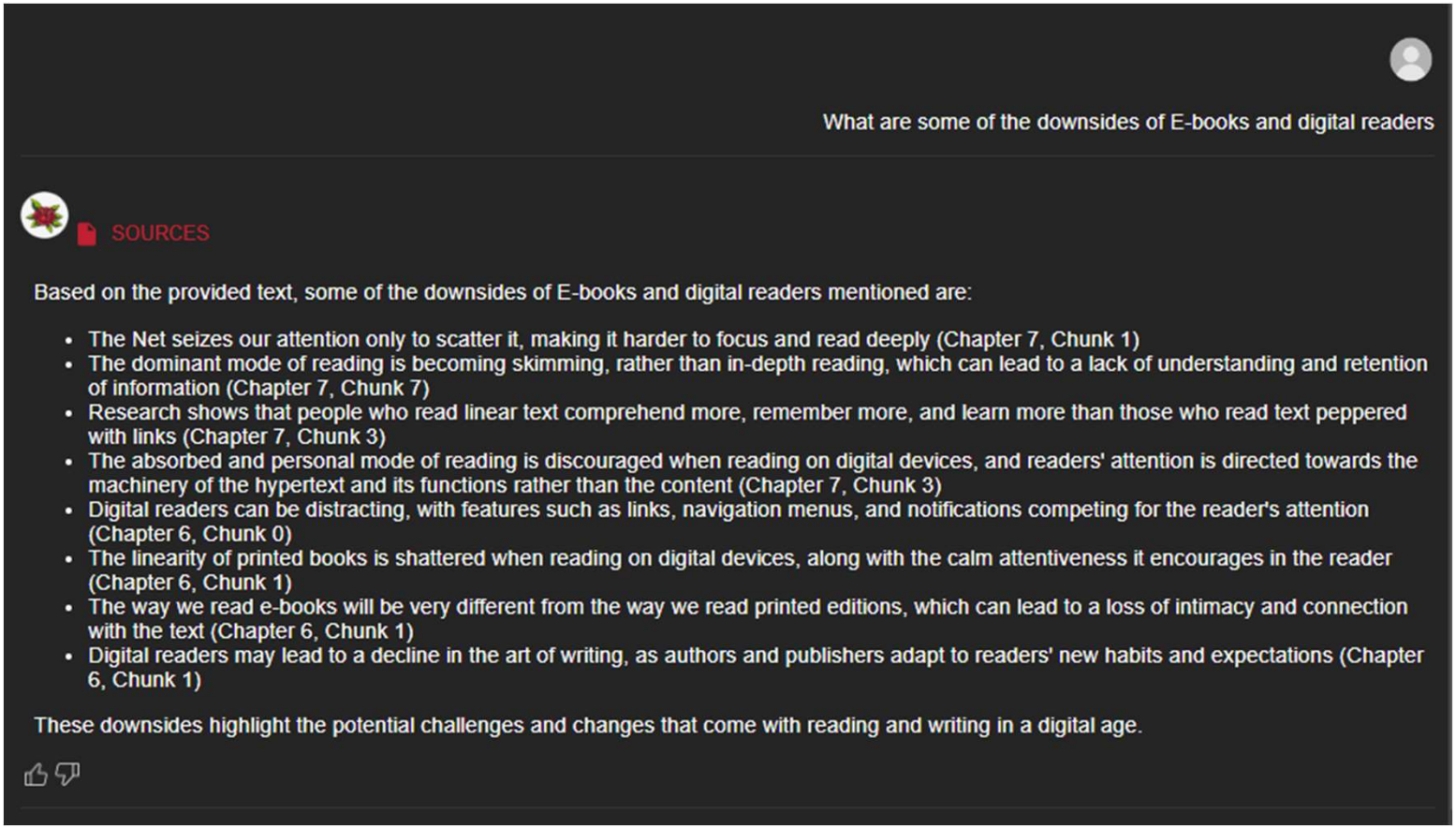
Many Libraries
One LLM

NHL Rule Book



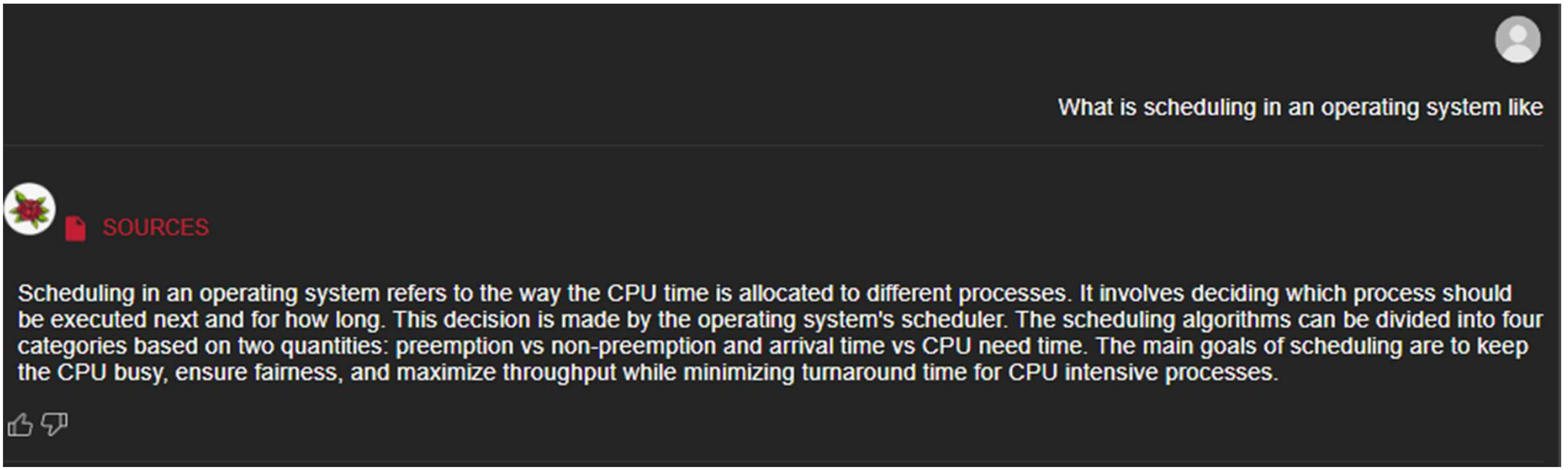
Many Libraries
One LLM

Nicholas Carr's *The Shallows*



Many Libraries
One LLM

Dr. Lembke's Class Lectures



Why Rosie?

- Mixtral-Mistral8x-7B (The model we have the best results with) at 16 bit floating point precision runs well on four DGX GPUs (only half of the context window as well)
- This equates to 128 GB of VRAM. VRAM is basically how much data a gpu (fancy computer) can work with at one point
- A RTX-4090 (The best consumer grade GPU) has 24 GB of VRAM. It simply cannot load this size model
- A 7B parameter model at 16 bit floating point precision can be run on two T4s, still too much for someone with a RTX-4090
- RTX-4090 could quantize to 4-bits, however results are poop
- Why no quantize → VLLM does not have stable version of AWQ available
- Over 1000 hours of jobs run on chatter alone over the course of the last 6 months. Most of those hours coming in the last three months.
- We would have been running a p3.16xlarge instance on AWS, which costs 24.48$ an hour, bringing us to a total cost of over $24,000 to host this service on AWS
- Lets say you just want to run the library
- Library is loaded on two T4s, this involves loading several models on to it
    - Embedding model
    - Reranker
    - Library
- This could be run locally, but every model in use would have to be much smaller, the embedding model, avoid using the reranker model and run a quantized 7B model or lower
    - Quality would simply be lacking
    - Processing of concurrent requests
- One nice added unplanned bonus feature is that Various libraries can use the same LLM. If one LLM is stood up, various people could make different libraries on their own documents and use the central LLM.
- Why do we need the H100s though? We can use bigger and better embedding models. Load bigger LLMs

    Mixtral-Mistral8x-7B: 23 on Chatbot Arena
    Mistrasl-7B-Instruct-v0.2: 37 on Chatbot Arena
    Llama-2-7b-chat: 51 on chatbot arena

Amazon EC2 P3 instance product details

Instance Size	GPUs - Tesla V100	GPU Peer to Peer	GPU Memory (GB)	vCPUs	Memory (GB)	Network Bandwidth	EBS Bandwidth	On-Demand Price/hr*
p3.2xlarge	1	N/A	16	8	61	Up to 10 Gbps	1.5 Gbps	$3.06
p3.8xlarge	4	NVLink	64	32	244	10 Gbps	7 Gbps	$12.24
p3.16xlarge	8	NVLink	128	64	488	25 Gbps	14 Gbps	$24.48
p3dn.24xlarge	8	NVLink	256	96	768	100 Gbps	19 Gbps	$31.218

Estimates

- 12.24/hr With AWS Sagemaker Serverless Inference, 4 V100s
- High estimate for us: 1750W (½ max power draw of dgx1) when inferencing, 500W when idle
- For 1 day, need to be inferencing for 12 minutes to be cheaper than Sagemaker
- For 30 mins/day - Rosie: $2.40, SM: $6.12
- For 2 hours/day - Rosie: $2.76, SM: $24.48

Other Reasons

- Allows for very easy configuration and re-configuration
- Gives easy access to all logs, hardware config, and data inputs and outputs
- Little extra cost for data storage/moving

Why Rosie?

- Big LLMs require big hardware
- Mixtral-Mistral8x-7B at ½ precision requires 128 GB of VRAM
  - 4090 has 24GB

Why Rosie?

Our cost on Rosie:
- 3500W under full load
- 500W when idle
- Electricity cost: 19 cents/kWh

At 8 hours/day for 1 week: < $50

AWS with "Sagemaker Serverless Inference":
- $24.48/hour

At 8 hours/day for 1 week: $1370

Amazon EC2 P3 instance product details

Instance Size	GPUs - Tesla V100	GPU Peer to Peer	GPU Memory (GB)	vCPUs	Memory (GB)	Network Bandwidth	EBS Bandwidth	On-Demand Price/hr*
p3.2xlarge	1	N/A	16	8	61	Up to 10 Gbps	1.5 Gbps	$3.06
p3.8xlarge	4	NVLink	64	32	244	10 Gbps	7 Gbps	$12.24
p3.16xlarge	8	NVLink	128	64	488	25 Gbps	14 Gbps	$24.48
p3dn.24xlarge	8	NVLink	256	96	768	100 Gbps	19 Gbps	$31.218

V100 Specs

SYSTEM SPECIFICATIONS	
GPUs	**8X Tesla V100**
Performance (Mixed Precision)	**1 petaFLOPS**
GPU Memory	**256 GB total system**
CPU	**Dual 20-Core Intel Xeon E5-2698 v4 2.2 GHz**
NVIDIA CUDA® Cores	**40,960**
NVIDIA Tensor Cores (on V100 based systems)	**5,120**
Power Requirements	**3,500 W**
System Memory	**512 GB 2,133 MHz DDR4 RDIMM**
Storage	**4X 1.92 TB SSD RAID 0**
Network	**Dual 10 GbE, 4 IB EDR**
Operating System	**Canonical Ubuntu, Red Hat Enterprise Linux**
System Weight	**134 lbs**
System Dimensions	**866 D x 444 W x 131 H (mm)**
Packing Dimensions	**1,180 D x 730 W x 284 H (mm)**
Operating Temperature Range	**5–35 °C**

Other Advantages

- Allows for very easy configuration and re-configuration (yaml)
- Full control over logs, hardware, job time, LLM, embedding model,
- Ingress-Egress fees are close to nonexistent

- It's all on-site

```yaml
sbatch_job_time: "0-12:0"

llm_config:
  folder: "new_backend/LLM"
  engine: "vllm"
  partition: "dgx"
  requirements_file: "llm_requirements.txt"
  gpus: 4
  cpus_per_gpu: 16
  max_model_len: 24000
  hf_llm_name: "mistralai/Mixtral-8x7B-Instruct-v0.1"
  dtype: "float16"
  out_file: "llm.out"

db_config:
  folder: "new_backend/db_guardrails"
  partition: "teaching"
  txt_dir: "/data/sdp/senior_design_llm/policies_v3/"
  gpus: 2
  cpus_per_gpu: 4
  reload_db: true
  reranker_name: "BAAI/bge-reranker-base"
  requirements_file: "backend_requirements.txt"
  embedding_model_name: "BAAI/bge-large-en-v1.5"
  prompt_file: "config/system_prompts.yml"
  metadata_file: "/data/sdp/senior_design_llm/policies_v3/metadata_generated.json"
  out_file: "db.out"
  summary_chunk_size: 512
  summary_chunk_overlap: 64
  document_chunk_size: 1024
  document_chunk_overlap: 128

guardrails_config:
  folder: "db_guardrails"
  reload_guardrails: true
  cached_guardrails: "cache/guardrails.yml"
  guardrails_config_file: "config/guardrails.yml"
  embedding_model_name: "BAAI/bge-large-en-v1.5"
```

- Four dgx GPUs running at float16 precision for our LLM.
  - Mixtral 8x7b
- In our pipeline to retrieve information there are 3 models at play
  - Embedding model from BAAI
  - Reranker model from BAAI
  - and an LLM
- Each taking up multiple Gb of VRAM.
- Calculating Embeddings, rerankingrunning LLM inference and reranking
- Our pipeline takes around 4-8 seconds to load for most documents. Asking questions takes a very long time.

Self healing Huberman Podcast

- 1650 documents
- 3 questions each
- 4950 questions
- 5 seconds each → This is one library visit
- 24750 seconds
- 412.5 minutes
- 6.875 hours per round


Questions?

To use this?

- Simply run git clone on our repo
- Copy and edit configuration files to your liking
- Run two python scripts and wait
- There will be a detailed guide soon… (With pictures and everything)