Gridspace

GRIDSPACE IAP 2024 LECTURE 4
Tools for LLM Memory

January 18, 2024

Questions from last time

- Why do you think ChatGPT did poorly on the 3-back task, even though it supposedly has superior memory to humans?

  It could be...

  - misinterpreting the task (even though it could recite a definition)

  - not effectively retrieving the information from the context

  - bad at counting the number of steps

Questions from last time

- As LLMs get larger and larger, and trained with more and more data. How would you expect the models to perform on each of the memory categories?

  - Working Memory: No change as long as the context length stays the same. However, it might get better at extracting relevant data from the context

  - Episodic Memory: Same as working memory

  - Semantic Memory: Depends on the data, it could see more of the same facts be repeated and remember things better. Or it could get conflicting data and become less effective.

Questions from last time

- LLM memorizing certain data can be a concern (e.g. private data, copyrighted data). What are some ways to alleviate this.

  There are many ways including:

  - Reducing repetition in data to reduce the chance of it remembering word for word

  - Using anonymized data (For cases of private data)

  - Create synthetic data (if applicable to use case)

LO

nch.

"Wh

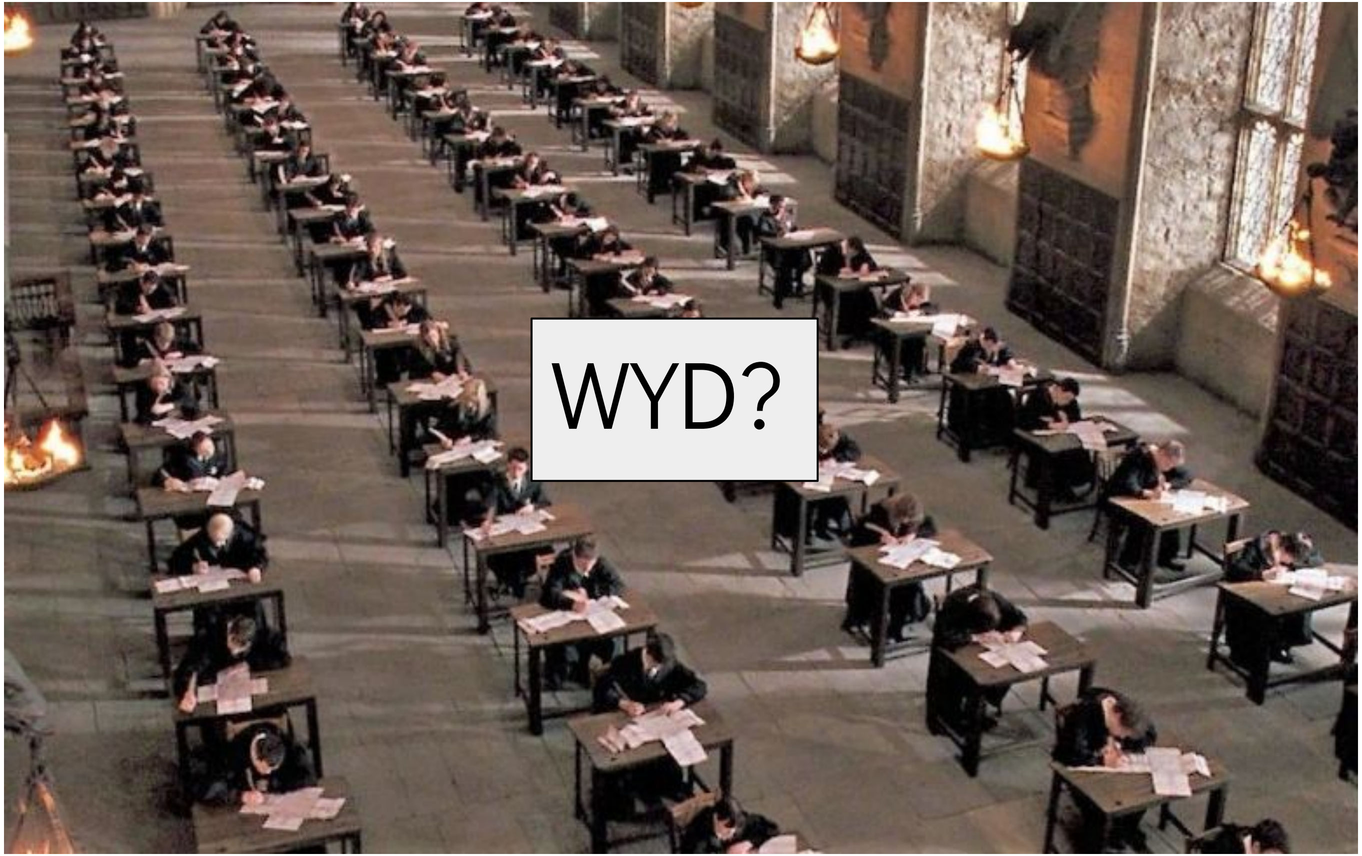
"Lec

)

Outline

- Problem: Information retrieval
- Solution 0: …
- Solution 1: …
- Solution 2: …
- Solution 3: …
- Solution 4: …

Scenario: closed book exam in 2 weeks

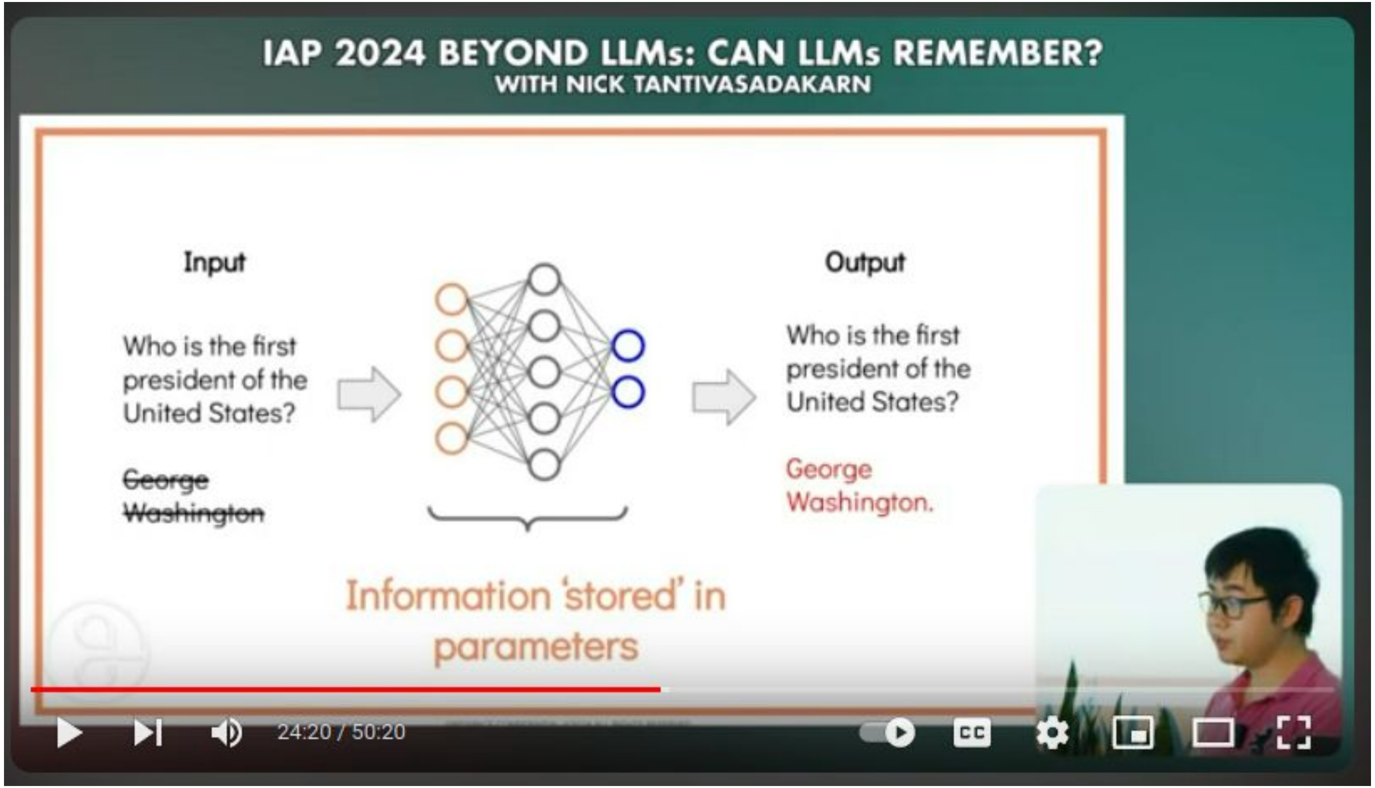


WYD?

Outline

- Problem: Information retrieval
- Solution 0: Memory in training data
- Solution 1: …
- Solution 2: …
- Solution 3: …
- Solution 4: …

Recall from Nick



IAP 2024 BEYOND LLMs: CAN LLMs REMEMBER?
WITH NICK TANTIVASADAKARN

Input

Who is the first president of the United States?

~~George Washington~~

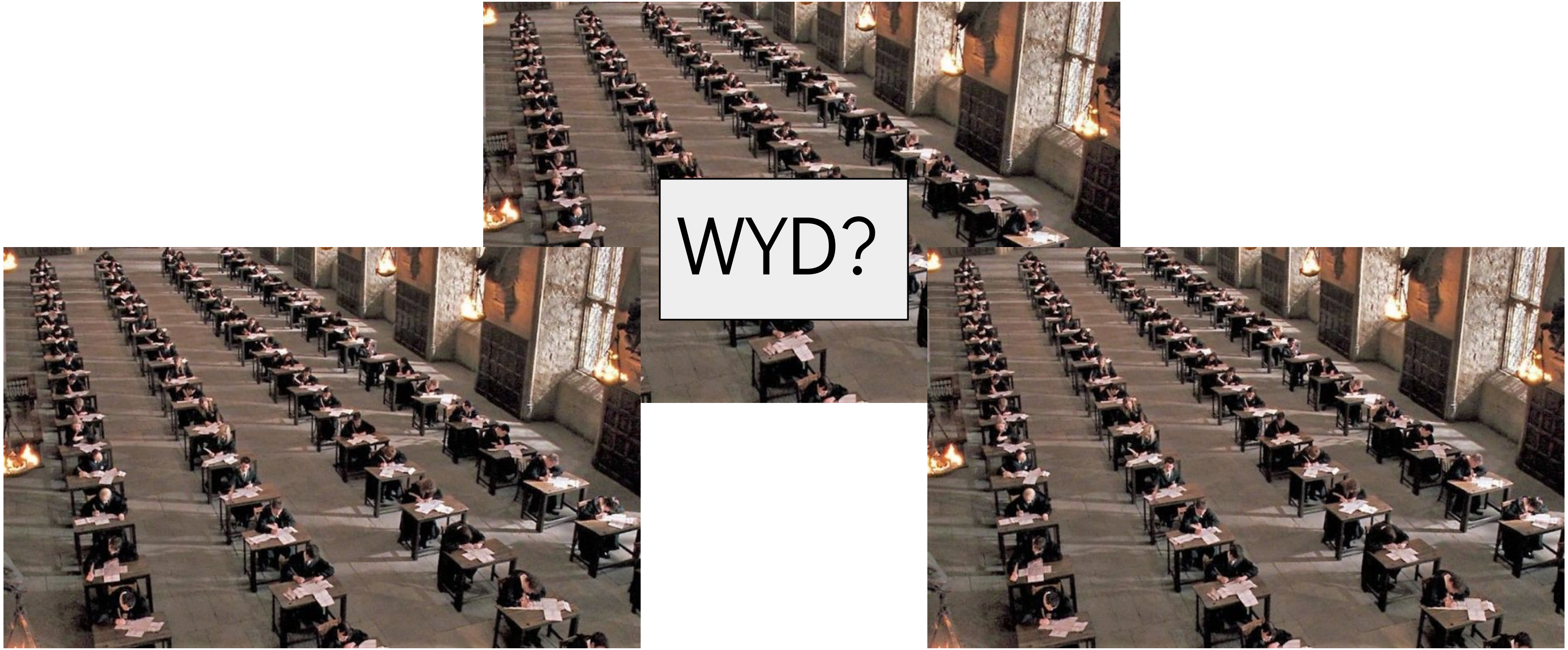
Output

Who is the first president of the United States?

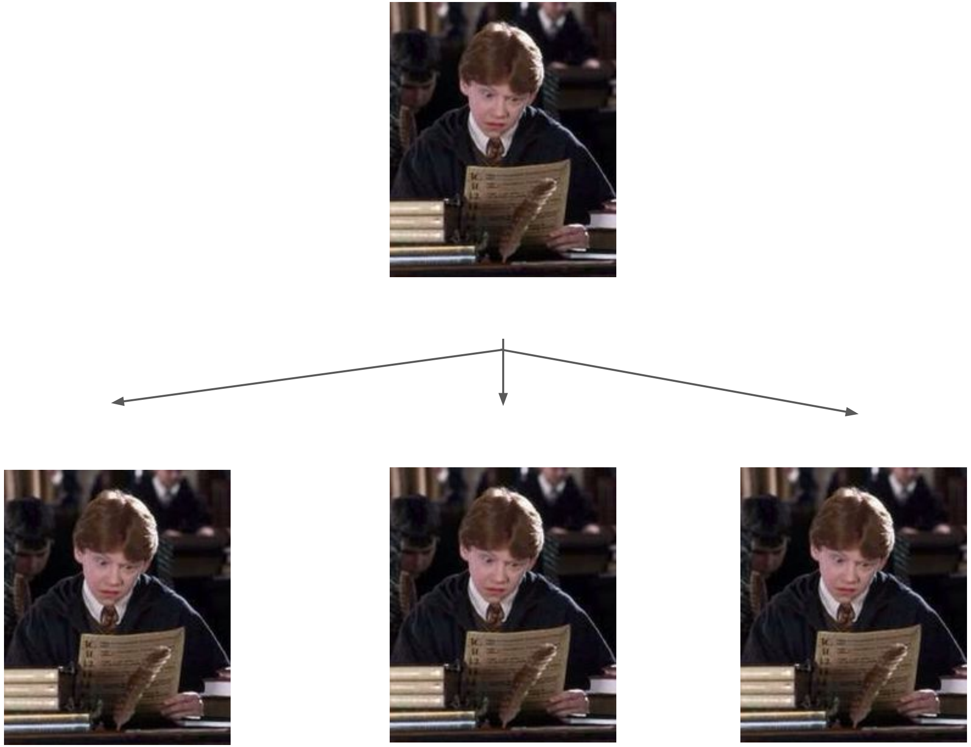
George Washington.

Information 'stored' in parameters

Scenario: You have three exams at the same time!
(But you can clone yourself)

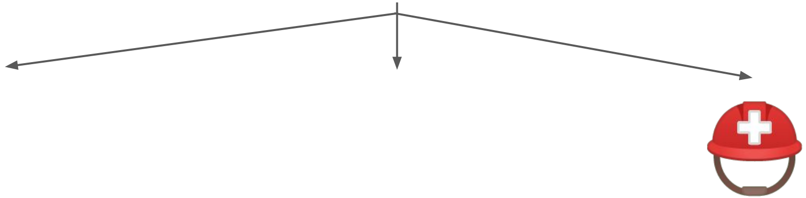


WYD?



Memory in training data

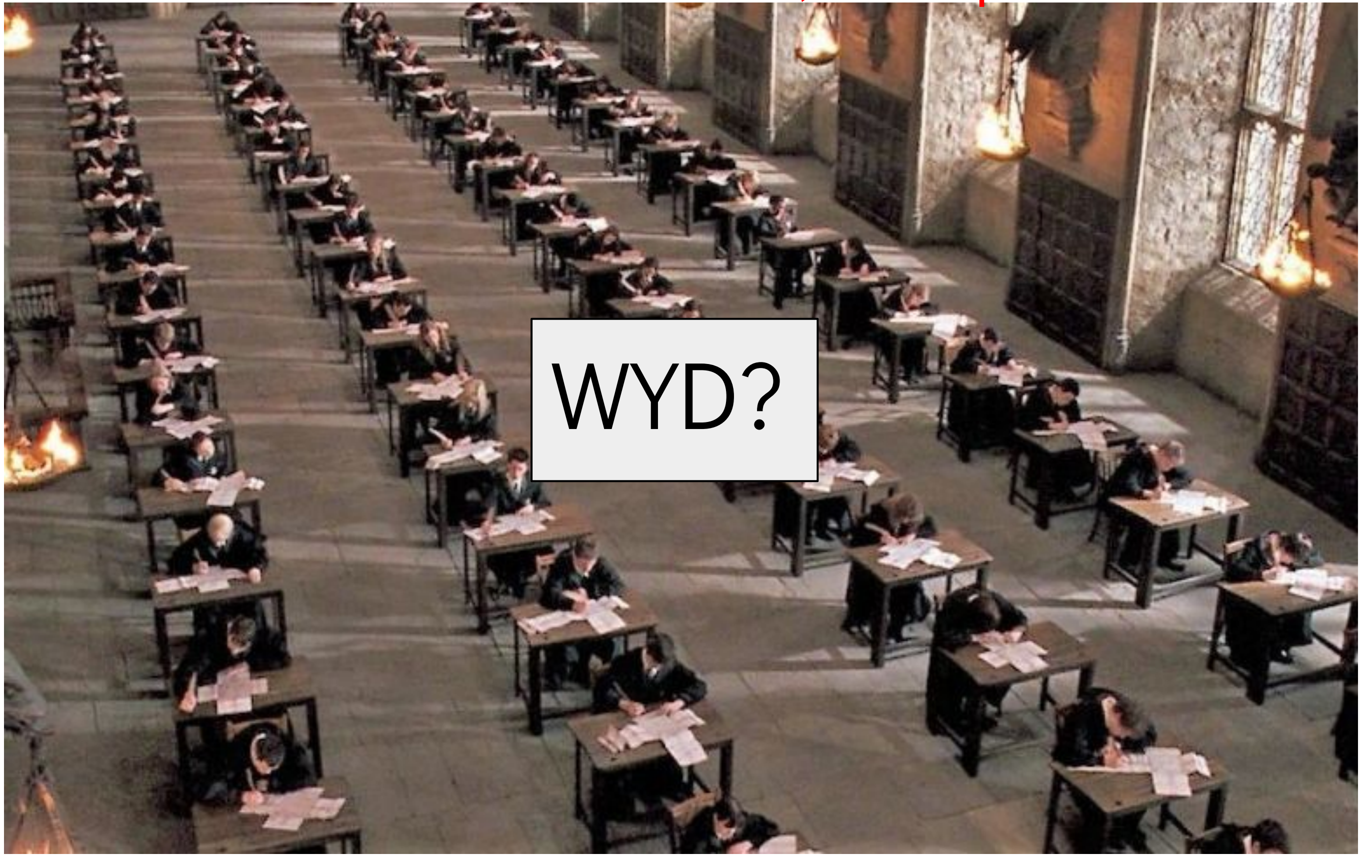
- One base model
- (Possibly multiple) delta on base model for domain specialization
- Train delta on new dataset



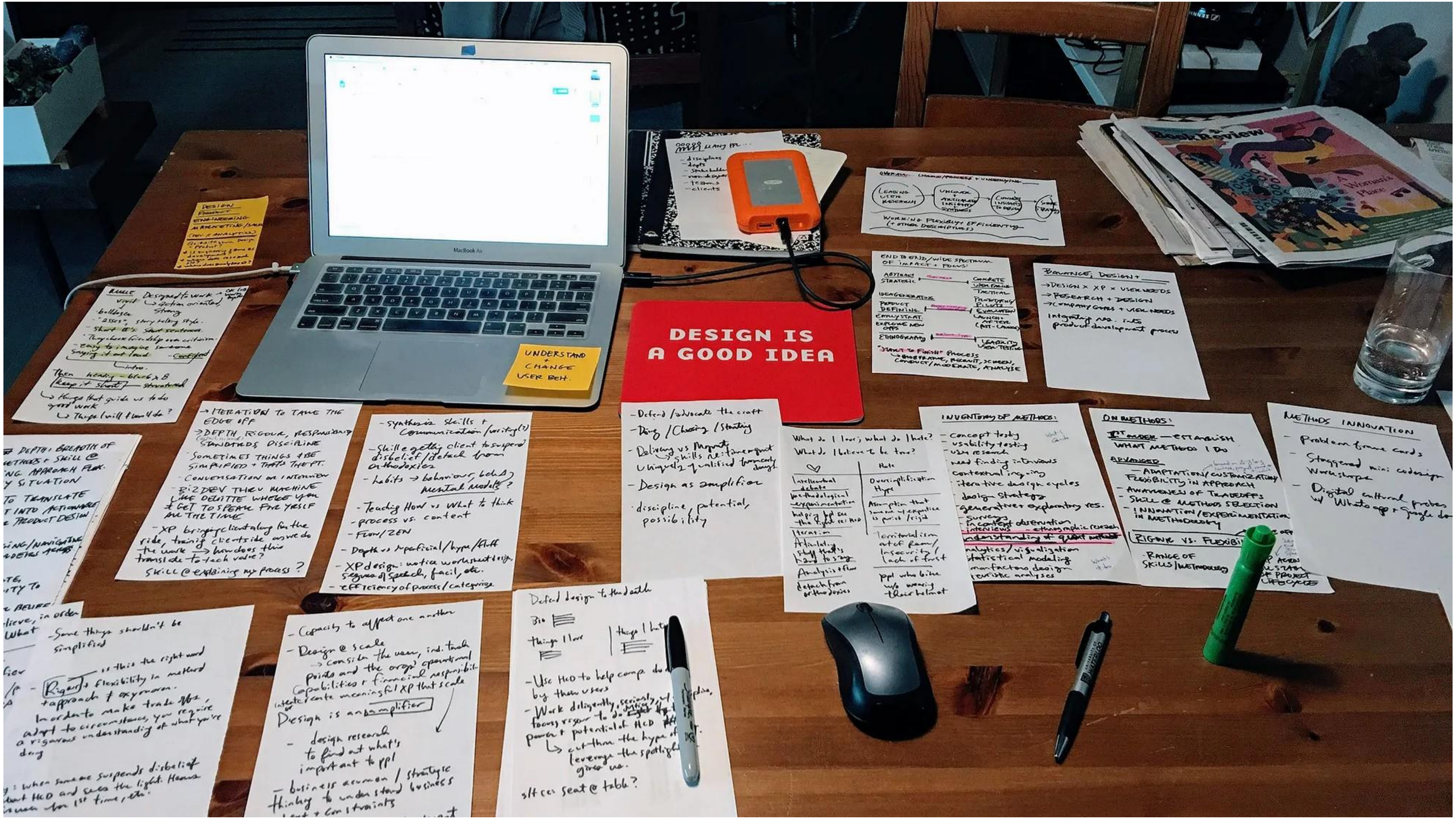
Memory in training data

- Relatively “static”, best for foundation knowledge, eg. language grammar, common sense (eg. the sky is blue)
- Expensive to retrain model when new knowledge is acquired

Scenario: exam in 1 hour, but open notes



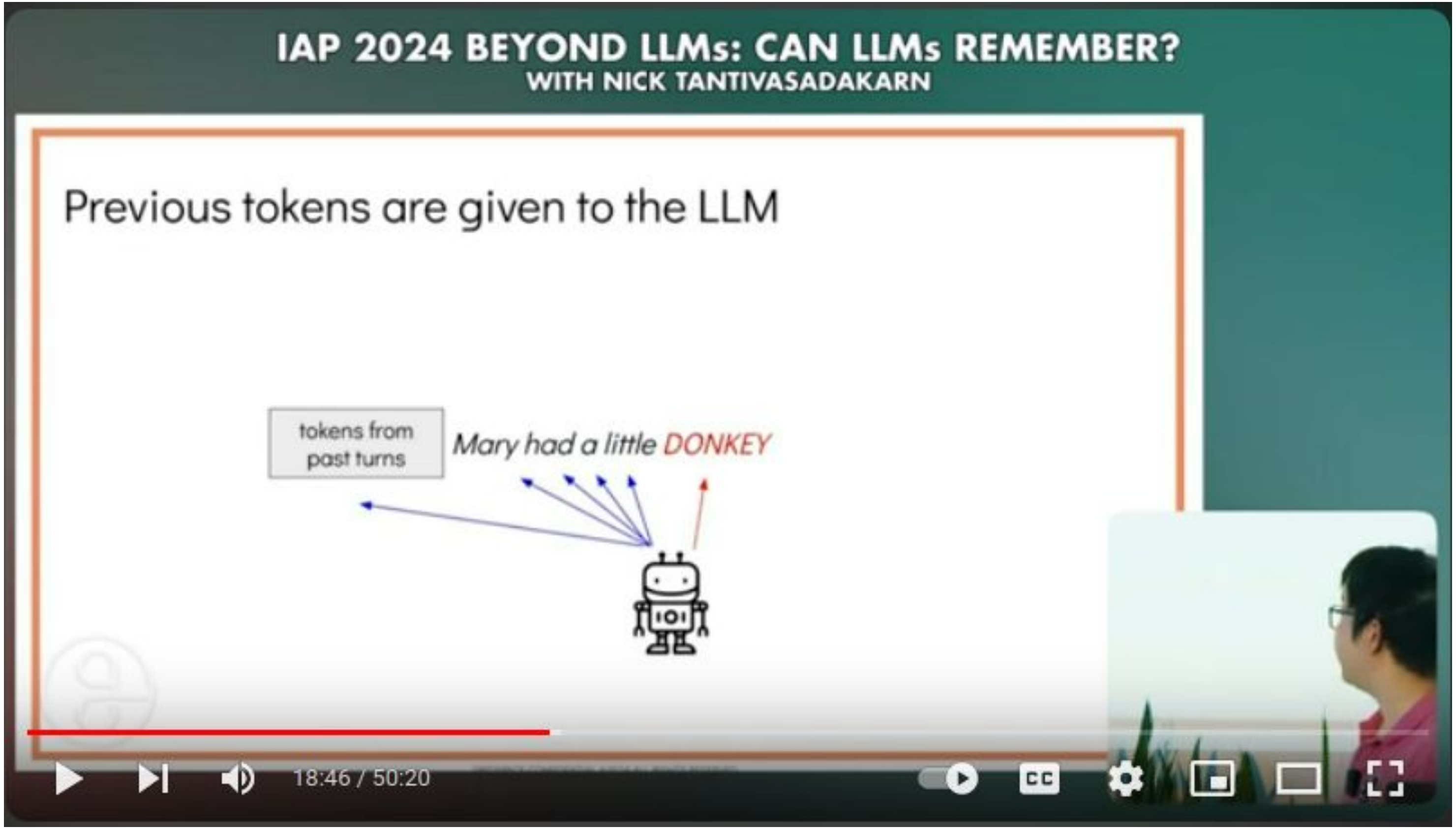
WYD?


DESIGN IS A GOOD IDEA
UNDERSTAND + CHANGE USER BEH.
MacBook Air
INVENTORY OF METHODS:
ON METHODS:
METHODS INNOVATION
BALANCE, DESIGN +
Defend design to the death
What do I love; what do I hate?
Design as amplifier
Design is an amplifier
Some things shouldn't be simplified
Capacity to affect one another
Design @ scale
Concept testing
Usability testing
User research
Problem frame cards
Staggered mini codesign workshops
Digital cultural probes

Outline

- Problem: Information retrieval
- Solution 0: Memory in training data
- Solution 1: Memory in LLM Context
- Solution 2: …
- Solution 3: …
- Solution 4: …

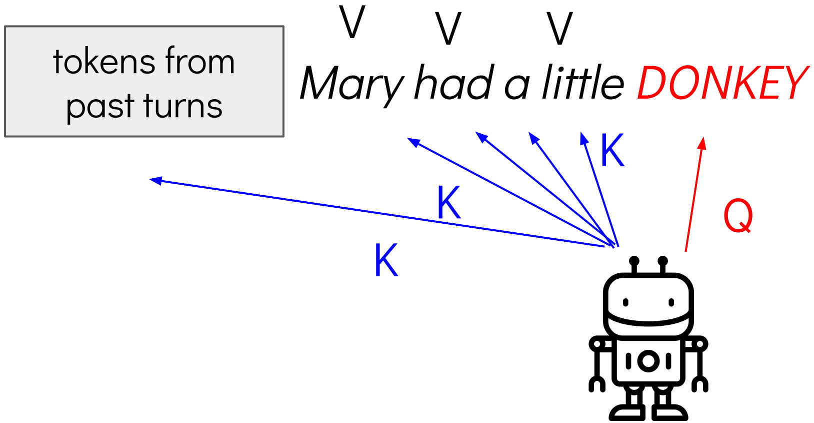
Recall from Nick



IAP 2024 BEYOND LLMs: CAN LLMs REMEMBER?
WITH NICK TANTIVASADAKARN

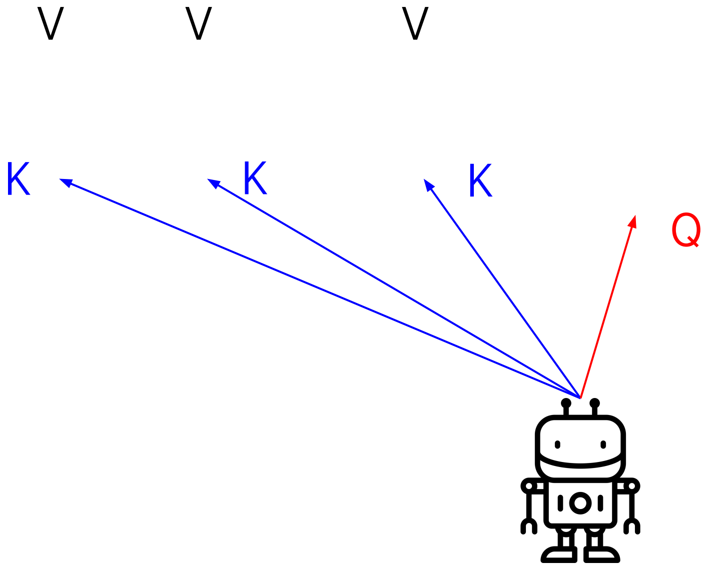
Previous tokens are given to the LLM

Memory in LLM Context

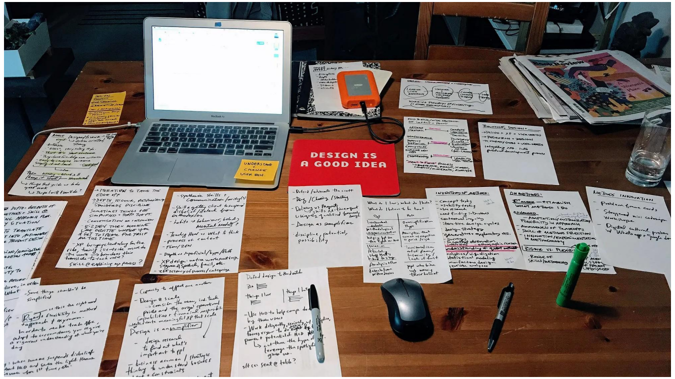


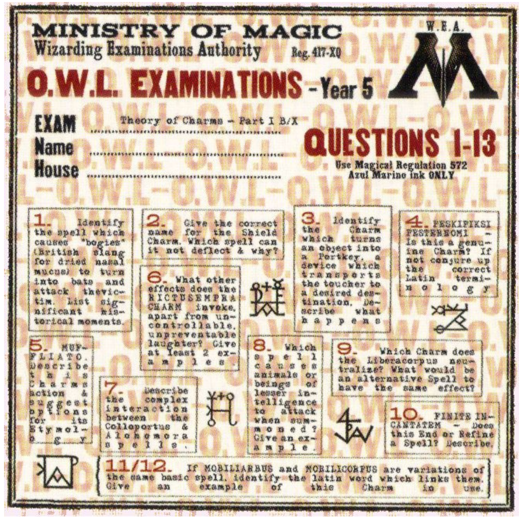
tokens from past turns

Attention: Q, K, V



Memory in LLM Context





MINISTRY OF MAGIC
Wizarding Examinations Authority Reg. 417-XQ
W.E.A.

O.W.L. EXAMINATIONS - Year 5

EXAM Theory of Charms - Part I B/X

Name ....

House ....

QUESTIONS 1-13

Use Magical Regulation 572
Axul Marino ink ONLY

1. Identify the spell which causes "bogies" (British slang for dried nasal mucus) to turn into bats and attack thevictim. List significant historical moments.

2. Give the correct name for the Shield Charm. Which spell can it not deflect & why?

3. Identify the Charm which turns an object into a Portkey, a device which transports the toucher to a desired destination. Describe what happens

4. PESKIPIKSI PESTERNOMI - Is this a genuine Charm? If not conjure up the correct Latin terminology

5. MUFFLIATO. Describe this Charms action & suggest options for its Etymology

6. What other effects does the RICTUSEMPRA CHARM invoke, apart from uncontrollable, unpreventable laughter? Give at least 2 examples.

7. Describe the complex interaction between the Colloportus & Alohomora spells.

8. Which spell causes animals or beings of lesser intelligence to attack when summoned? Give an example.

9. Which Charm does the Liberacorpus neutralize? What would be an alternative Spell to have the same effect?

10. FINITE INCANTATEM - Does this End or Refine a Spell? Describe.

11/12. If MOBILIARBUS and MORILICORPUS are variations of the same basic spell, identify the latin word which links them. Give an example of this Charm in use.

K

K

K

Q

Mucus ad Nauseam

Memory in LLM Context

- Storing K,V takes a lot of GPU memory
- Storing K, V can be messy (fragmented)

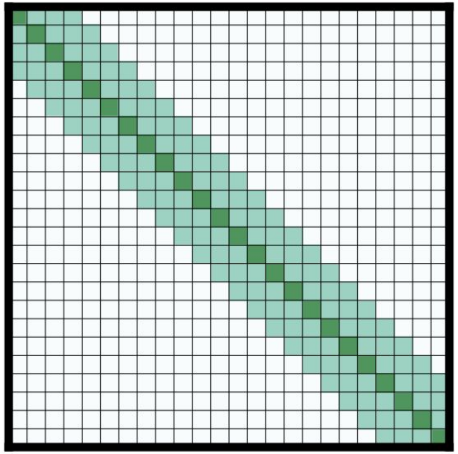
Memory in LLM Context

- Economize, reduce size of K,V
  - Sliding window attention
  - Quantization
  - Group Query Attention

- Reduce memory waste due to messy storage (fragmentation)
  - PagedAttention

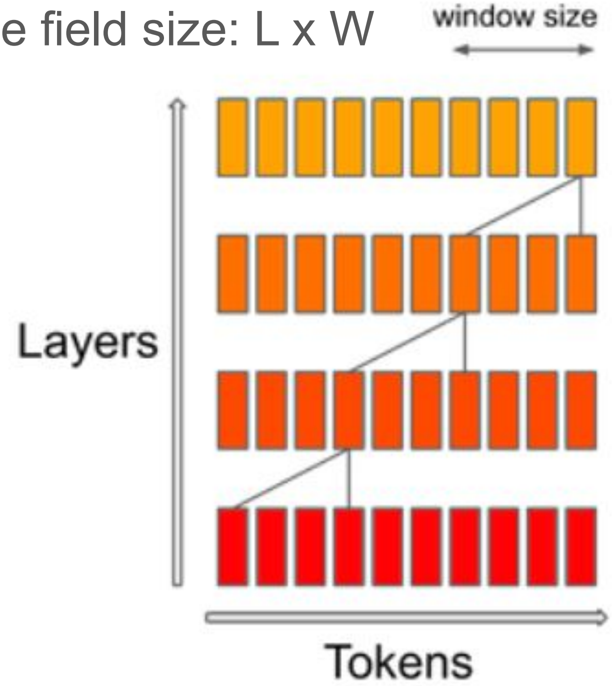
Memory in LLM Context

Window size W,
a transformer of L layers,
Receptive field size: L x W



(b) Sliding window attention

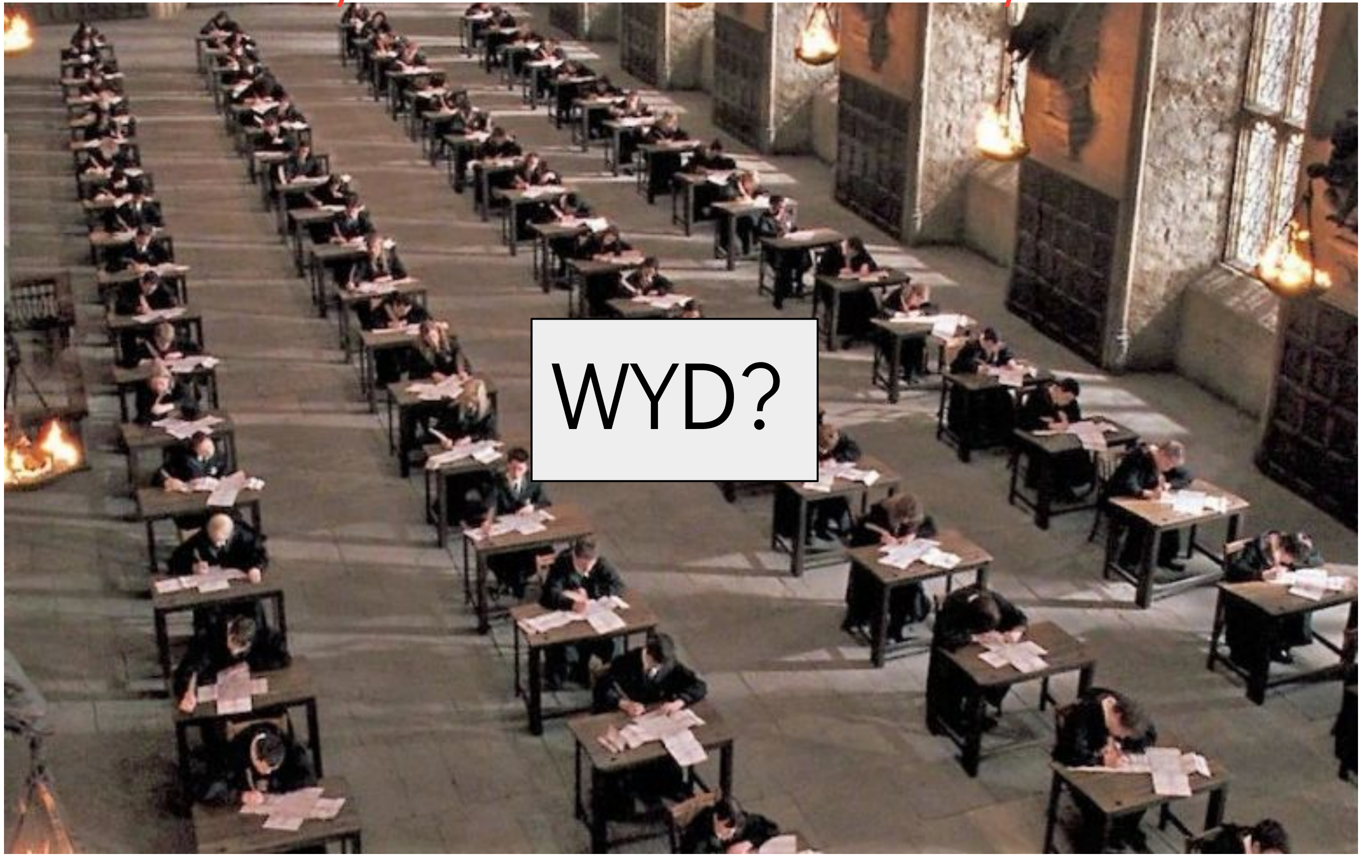
https://paperswithcode.com/method/sliding-window-attention



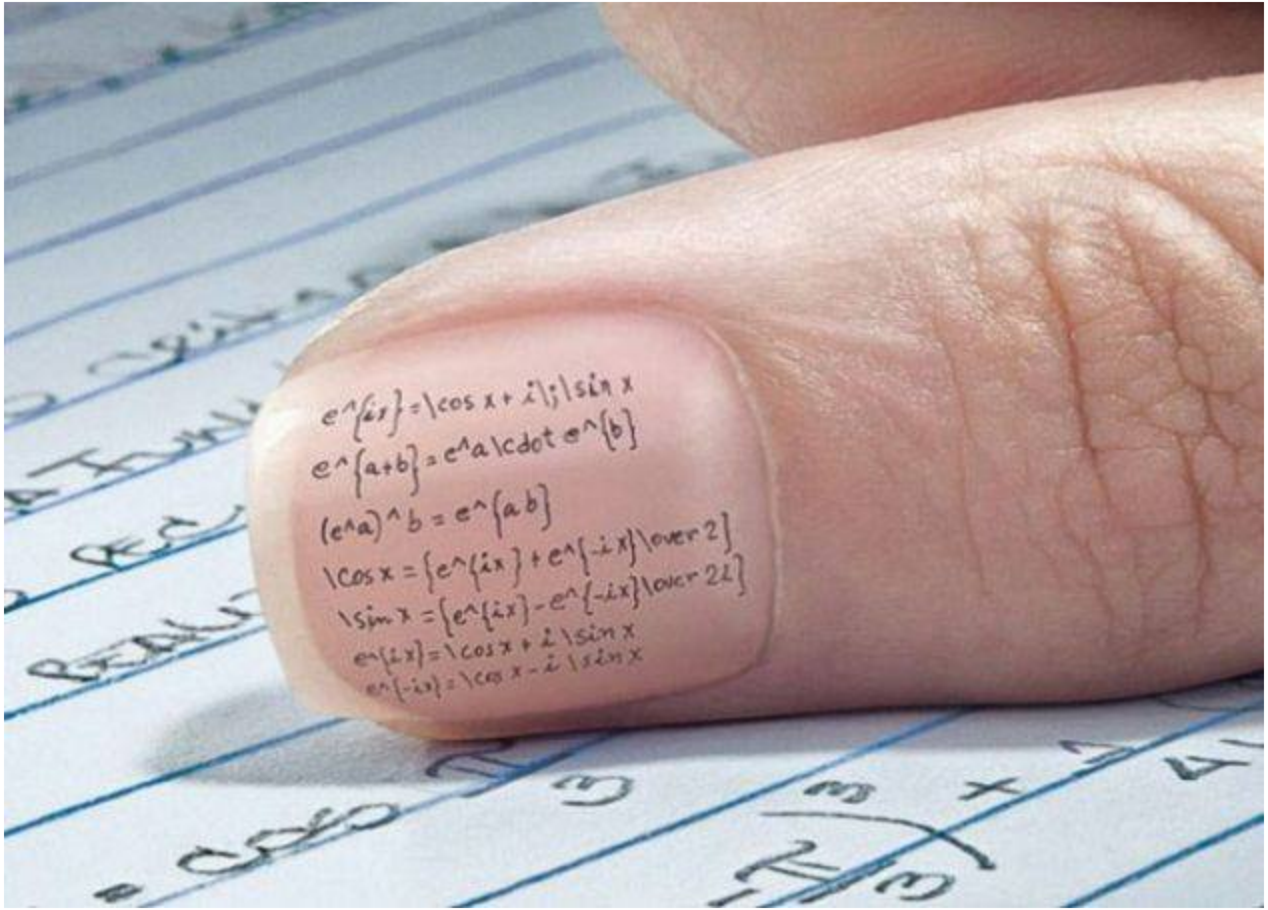
Memory in LLM Context

- Still limited, 32k tokens
- Still computationally expensive

Scenario: your desk cannot fit all your notes!



WYD?

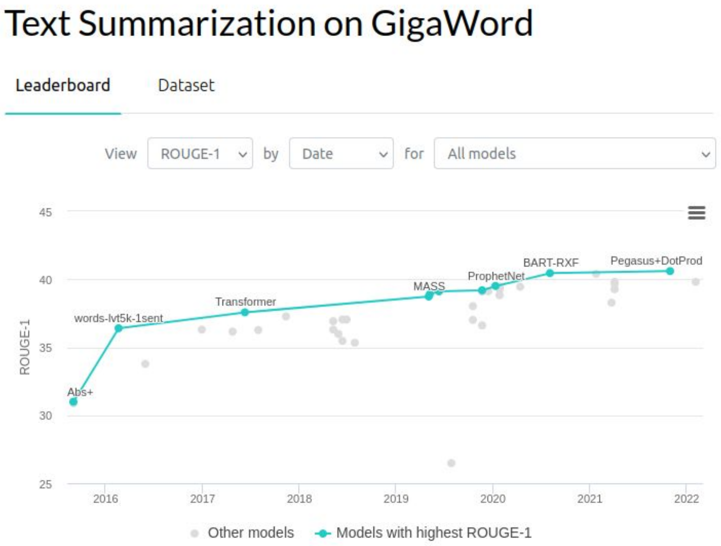

e^{ix} = \cos x + i\;\sin x
e^{a+b} = e^a\cdot e^{b}
(e^a)^b = e^{ab}
\cos x = {e^{ix} + e^{-ix}\over 2]
\sin x = {e^{ix} - e^{-ix}\over 2i]
e^{ix} = \cos x + i \sin x
e^{-ix} = \cos x - i \sin x

Outline

- Problem: Information retrieval
- Solution 0: Memory in training data
- Solution 1: Memory in LLM Context
- Solution 2: Compress chat history/context
- Solution 3: …
- Solution 4: …

Compress chat history/context

- LLMs are really good at summarization



Text Summarization on GigaWord

Leaderboard Dataset

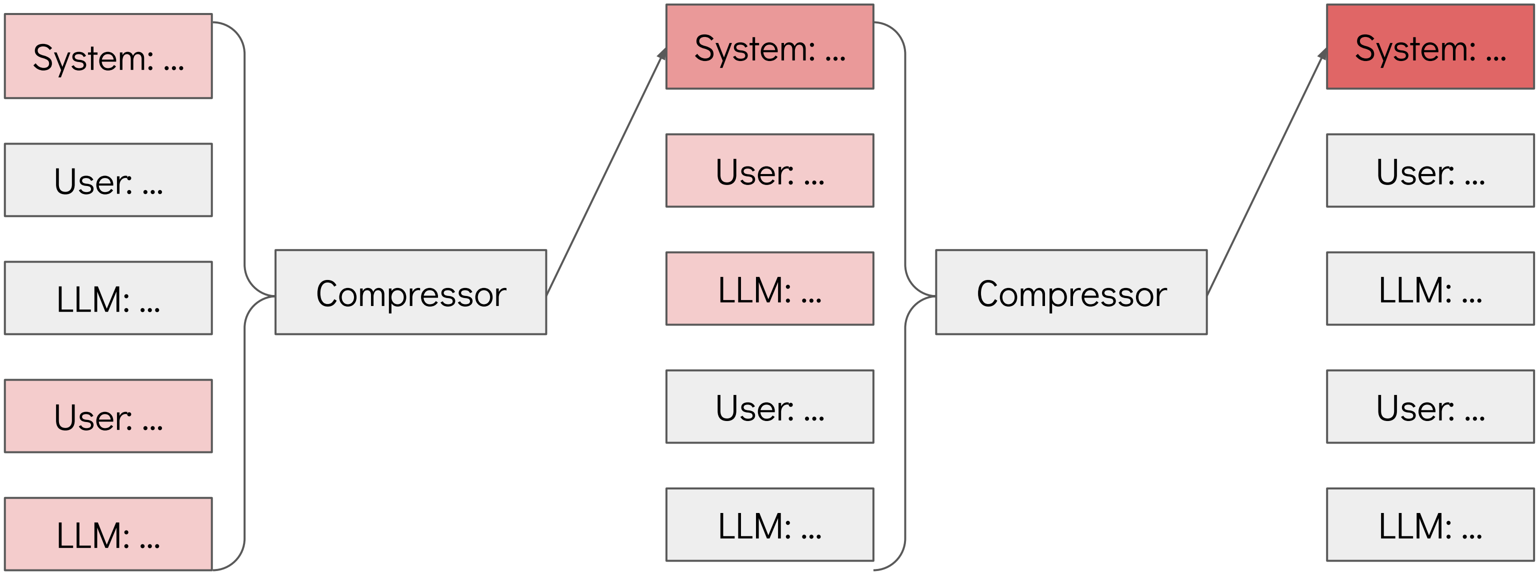
View ROUGE-1 by Date for All models

- Let's summarize chat history/context so far!

Compress chat history/context

- “In at most 1000 words, generate a summary of the chat history so far with sufficient detail to act as a replacement for the chat history in case we lose it. Pay special attention to instructions given by the user and system.”

Compress chat history/context



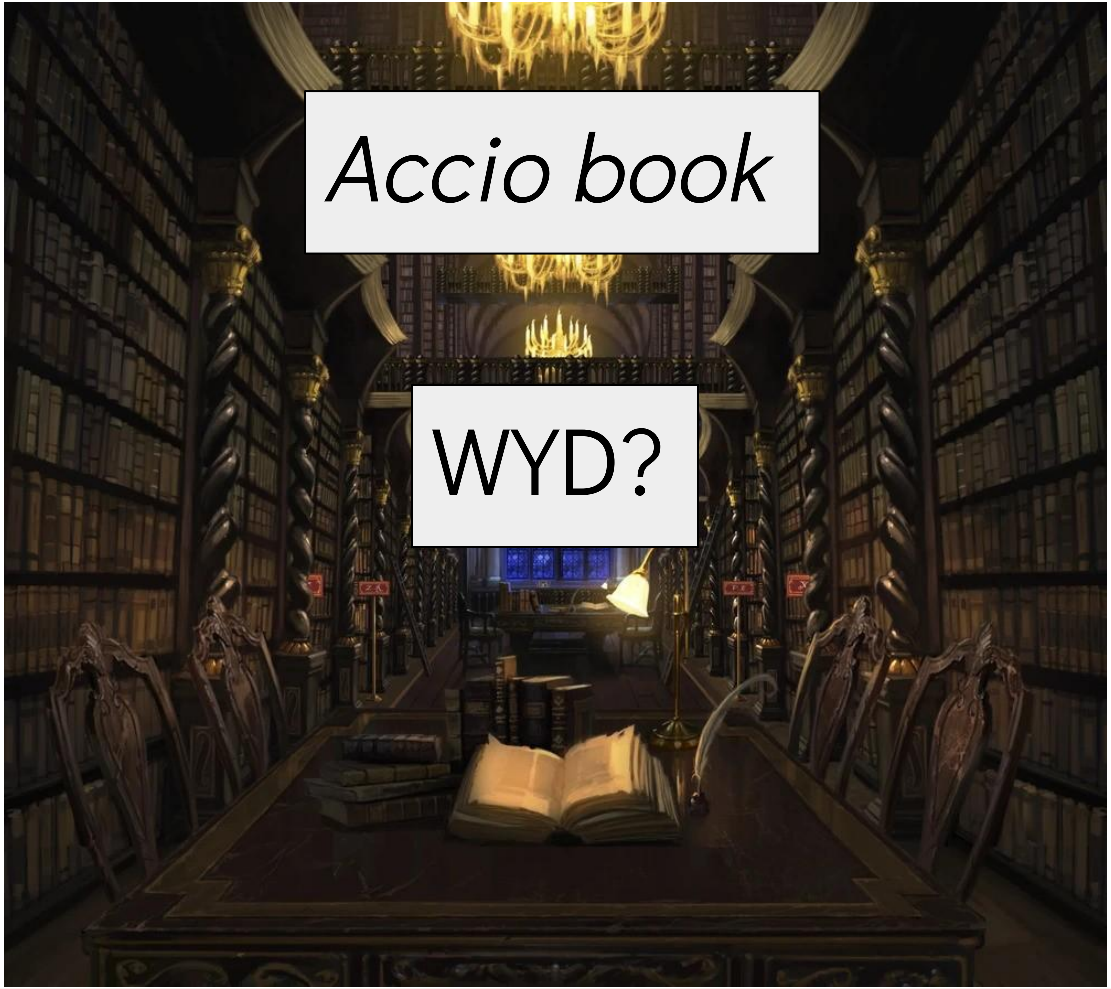
Compress chat history/context

- Needs to run LLM compressor regularly
- Some information is lost

Scenario: you have incantation to summon an object

Accio

Summoning Charm. This charm summons an object to the caster, potentially over a significant distance.



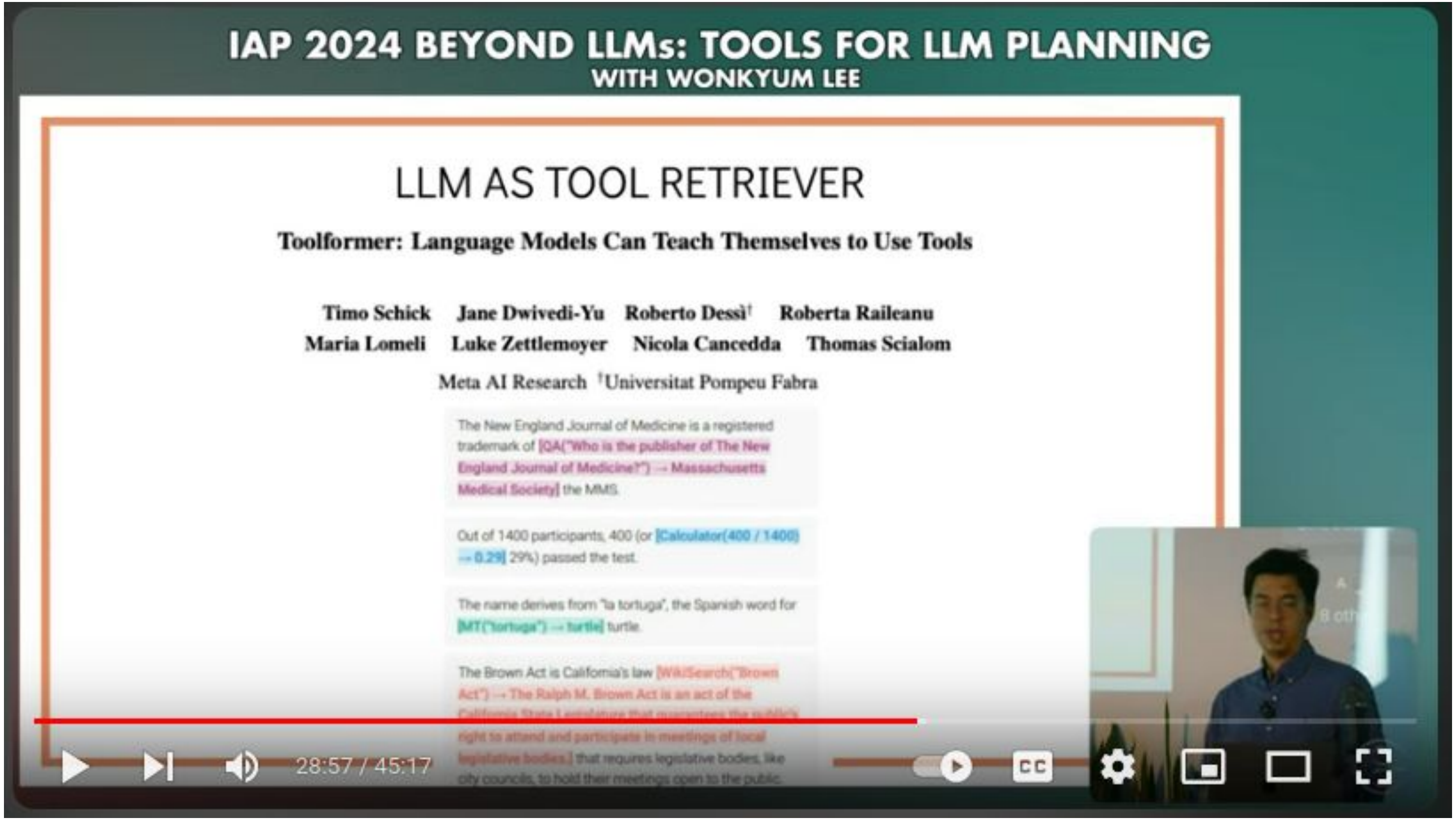
Accio book

WYD?

Outline

- Problem: Information retrieval
- Solution 0: Memory in training data
- Solution 1: Memory in LLM Context
- Solution 2: Compress chat history
- Solution 3: Memory in external database
- Solution 4: …

Recall from week 1: LLM can use tools



IAP 2024 BEYOND LLMs: TOOLS FOR LLM PLANNING
WITH WONKYUM LEE

LLM AS TOOL RETRIEVER

Toolformer: Language Models Can Teach Themselves to Use Tools

Timo Schick Jane Dwivedi-Yu Roberto Dessì[†] Roberta Raileanu
Maria Lomeli Luke Zettlemoyer Nicola Cancedda Thomas Scialom

Meta AI Research [†]Universitat Pompeu Fabra

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

Recall from week 1: LLM can use tools

System: You may use these tools: “check_temperature”, args: … usage: …

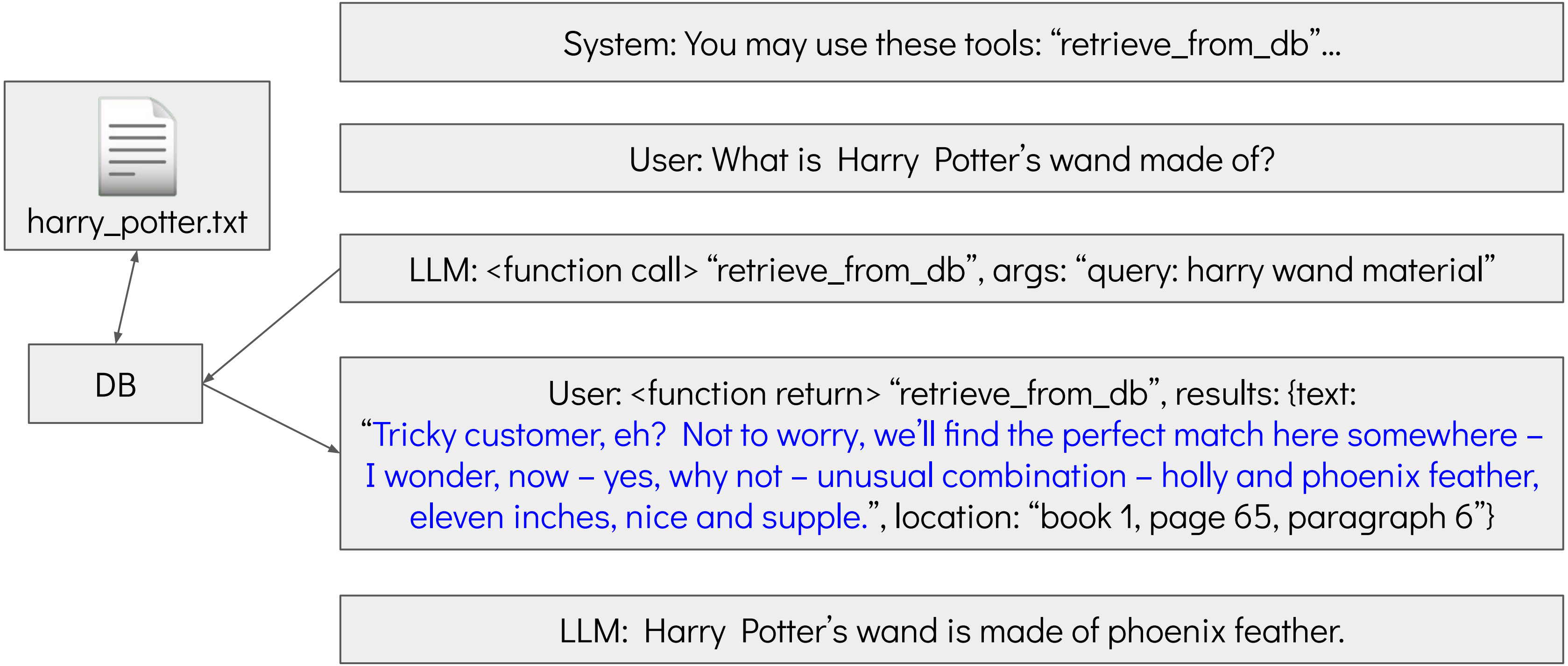
User: What is the temperature in LA now?

LLM: <function call> “check_temperature”, args: “location: LA, time: now”

User: <function return> “check_temperature”, results: “60F”

LLM: The temperature in LA is currently 60F! Nice day for hiking!

Memory in external database



System: You may use these tools: “retrieve_from_db”...

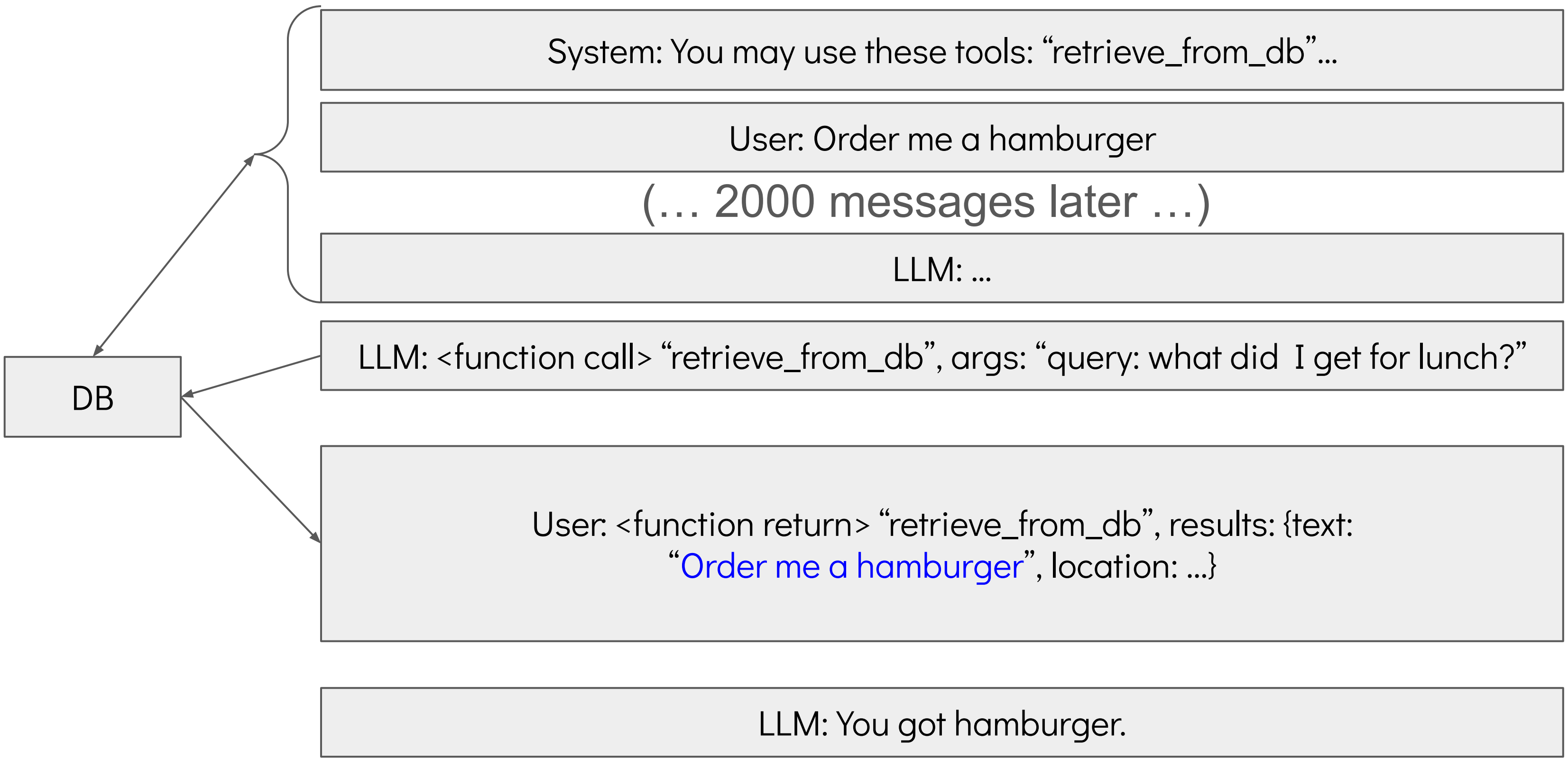
User: What is Harry Potter’s wand made of?

LLM: <function call> “retrieve_from_db”, args: “query: harry wand material”

User: <function return> “retrieve_from_db”, results: {text: “Tricky customer, eh? Not to worry, we’ll find the perfect match here somewhere – I wonder, now – yes, why not – unusual combination – holly and phoenix feather, eleven inches, nice and supple.”, location: “book 1, page 65, paragraph 6”}

LLM: Harry Potter’s wand is made of phoenix feather.

Memory in external database



System: You may use these tools: “retrieve_from_db”...

User: Order me a hamburger

(… 2000 messages later …)

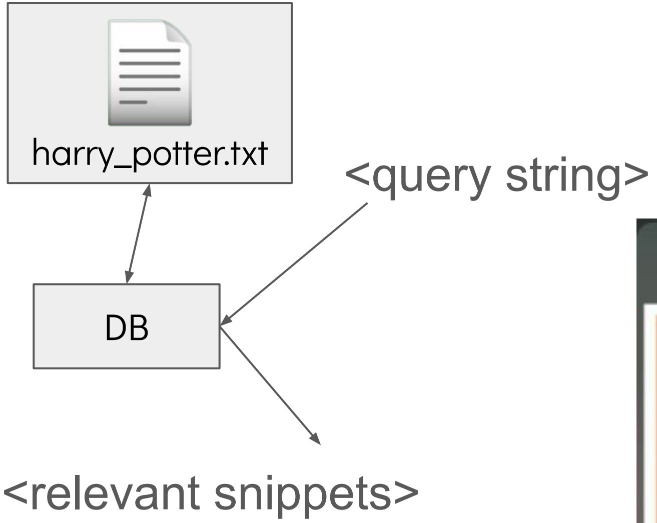
LLM: ...

LLM: <function call> “retrieve_from_db”, args: “query: what did I get for lunch?”

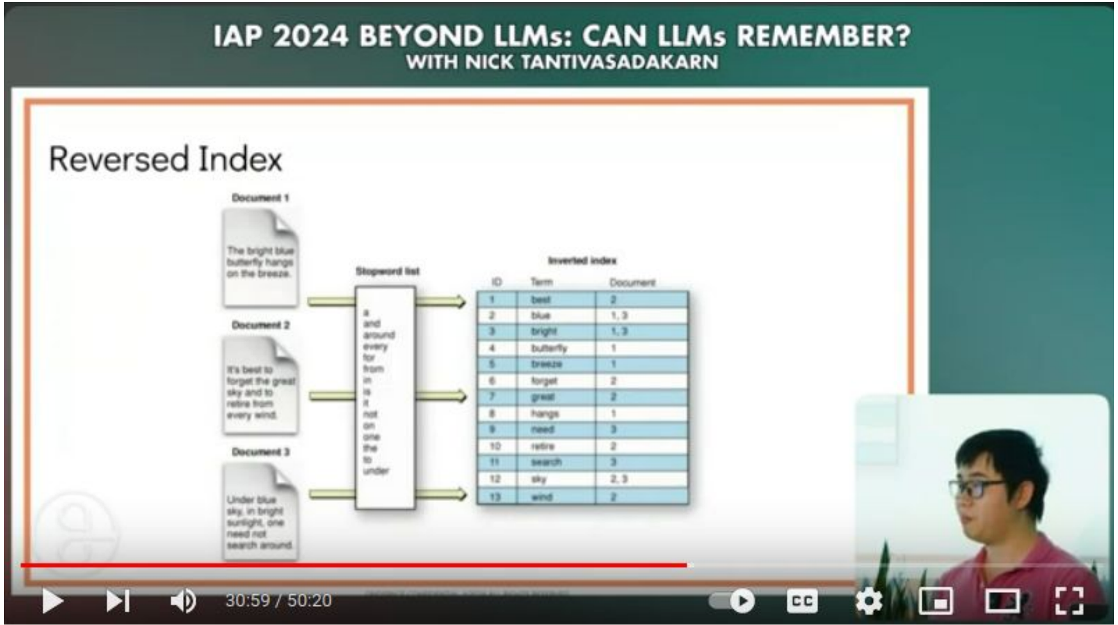
User: <function return> “retrieve_from_db”, results: {text: “Order me a hamburger”, location: ...}

LLM: You got hamburger.

Memory in external database (Reversed Index)



- Can be BM25



IAP 2024 BEYOND LLMs: CAN LLMs REMEMBER?

WITH NICK TANTIVASADAKARN

Reversed Index

Document 1

The bright blue butterfly hangs on the breeze.

Document 2

It's best to forget the great sky and to retire from every wind.

Document 3

Under blue sky, in bright sunlight, one need not search around.

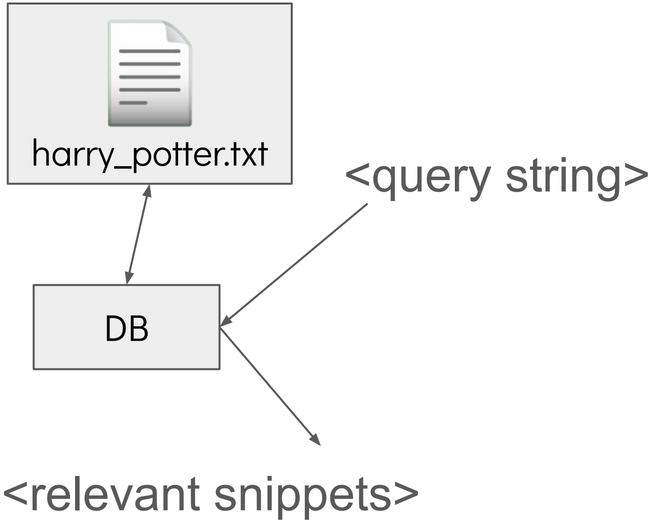
Stopword list

a
and
around
every
for
from
in
is
it
not
on
one
the
to
under

Inverted index

ID	Term	Document
1	best	2
2	blue	1, 3
3	bright	1, 3
4	butterfly	1
5	breeze	1
6	forget	2
7	great	2
8	hangs	1
9	need	3
10	retire	2
11	search	3
12	sky	2, 3
13	wind	2

Memory in external database (Transformer-based retrieval)



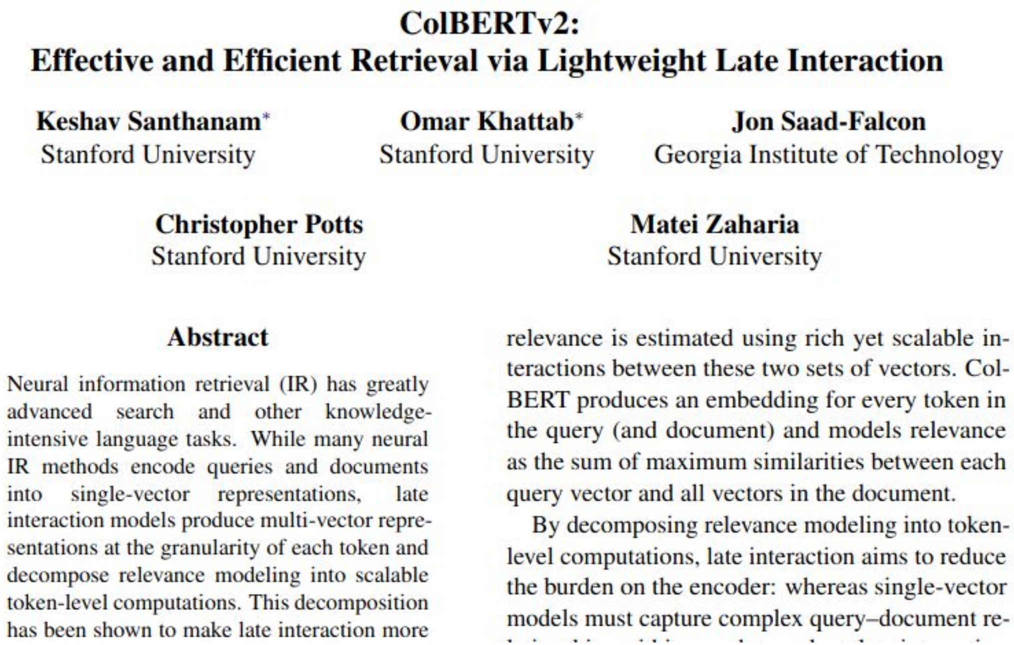
harry_potter.txt

DB

<query string>

<relevant snippets>

- Can be transformer-based
- eg. Colbertv2



ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction

Keshav Santhanam* Stanford University
Omar Khattab* Stanford University
Jon Saad-Falcon Georgia Institute of Technology
Christopher Potts Stanford University
Matei Zaharia Stanford University

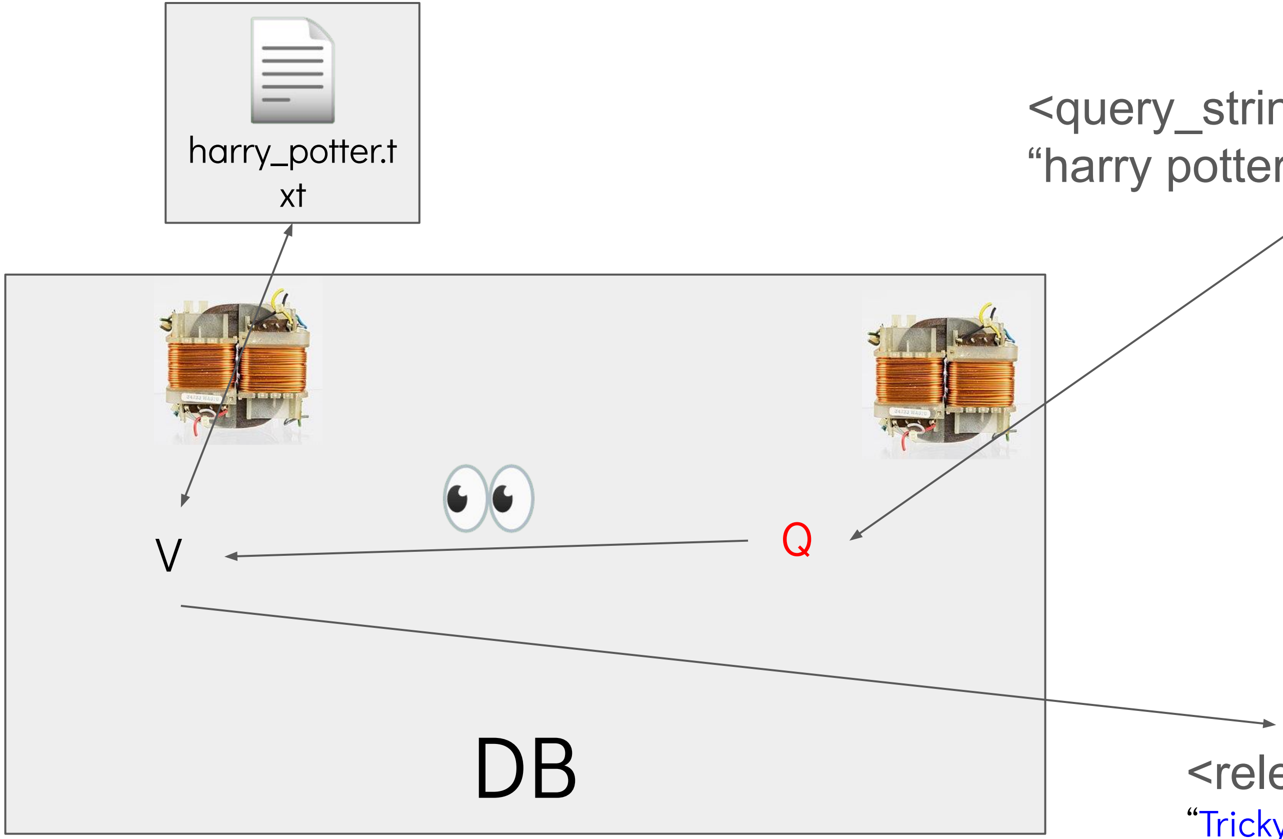
Abstract

Neural information retrieval (IR) has greatly advanced search and other knowledge-intensive language tasks. While many neural IR methods encode queries and documents into single-vector representations, late interaction models produce multi-vector representations at the granularity of each token and decompose relevance modeling into scalable token-level computations. This decomposition has been shown to make late interaction more

relevance is estimated using rich yet scalable interactions between these two sets of vectors. ColBERT produces an embedding for every token in the query (and document) and models relevance as the sum of maximum similarities between each query vector and all vectors in the document.

By decomposing relevance modeling into token-level computations, late interaction aims to reduce the burden on the encoder: whereas single-vector models must capture complex query–document re-

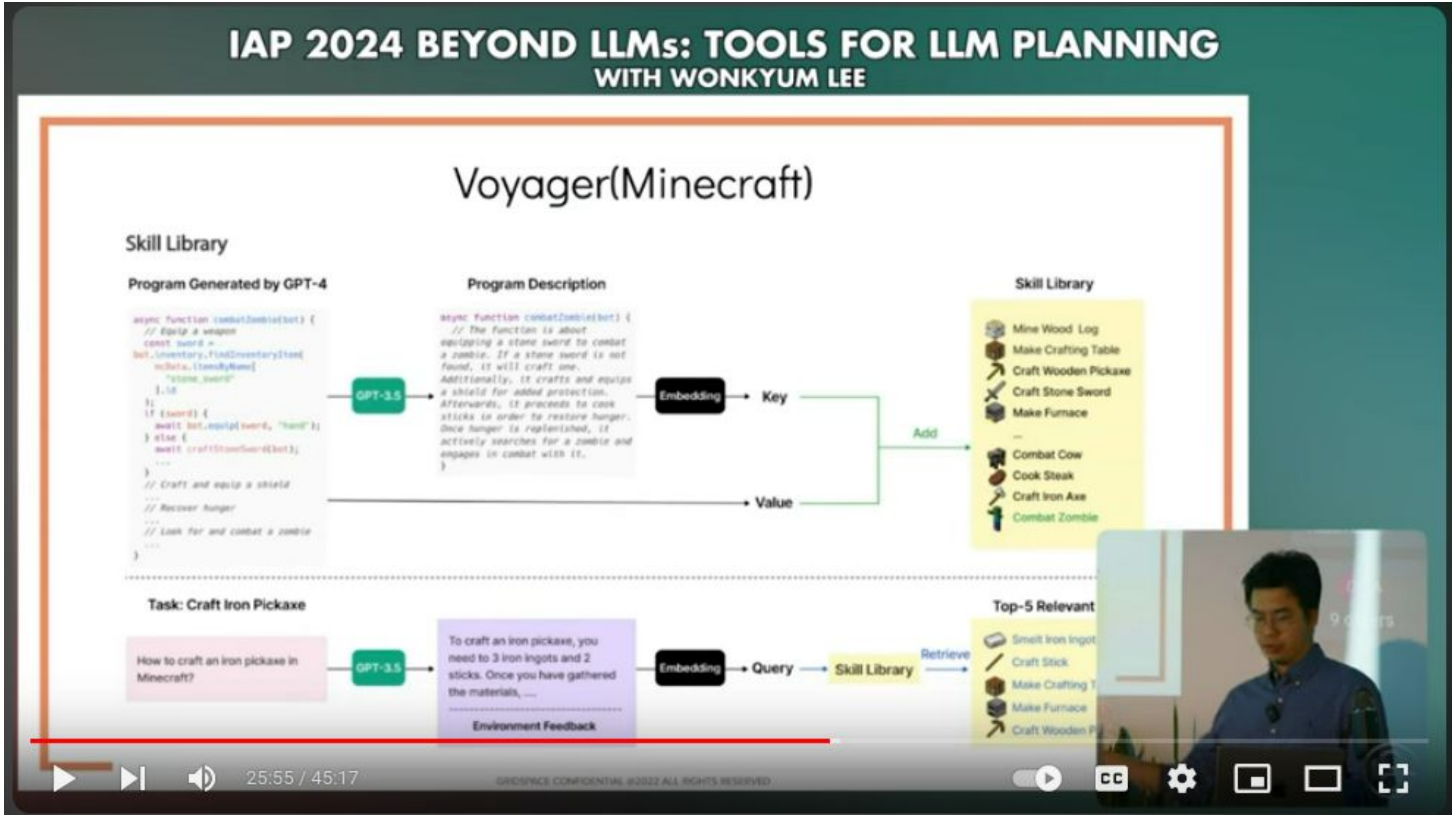
Memory in external database (Transformer-based retrieval)



<query_string> eg.
“harry potter wand material”

<relevant snippets> eg.
“Tricky customer, eh?...”

Recall from week 1: Voyager skill retrieval



IAP 2024 BEYOND LLMs: TOOLS FOR LLM PLANNING
WITH WONKYUM LEE

Voyager(Minecraft)

Skill Library

Program Generated by GPT-4

```
async function combatZombie(bot) {
  // Equip a weapon
  const sword =
bot.inventory.findInventoryItem(
    mcData.itemsByName[
      "stone_sword"
    ].id
  );
  if (sword) {
    await bot.equip(sword, "hand");
  } else {
    await craftStoneSword(bot);
    ...
  }
  // Craft and equip a shield
  ...
  // Recover hunger
  ...
  // Look for and combat a zombie
  ...
}
```

Program Description

```
async function combatZombie(bot) {
  // The function is about
equipping a stone sword to combat
a zombie. If a stone sword is not
found, it will craft one.
Additionally, it crafts and equips
a shield for added protection.
Afterwards, it proceeds to cook
sticks in order to restore hunger.
Once hunger is replenished, it
actively searches for a zombie and
engages in combat with it.
}
```

Skill Library

- Mine Wood Log
- Make Crafting Table
- Craft Wooden Pickaxe
- Craft Stone Sword
- Make Furnace
- ...
- Combat Cow
- Cook Steak
- Craft Iron Axe
- Combat Zombie

Task: Craft Iron Pickaxe

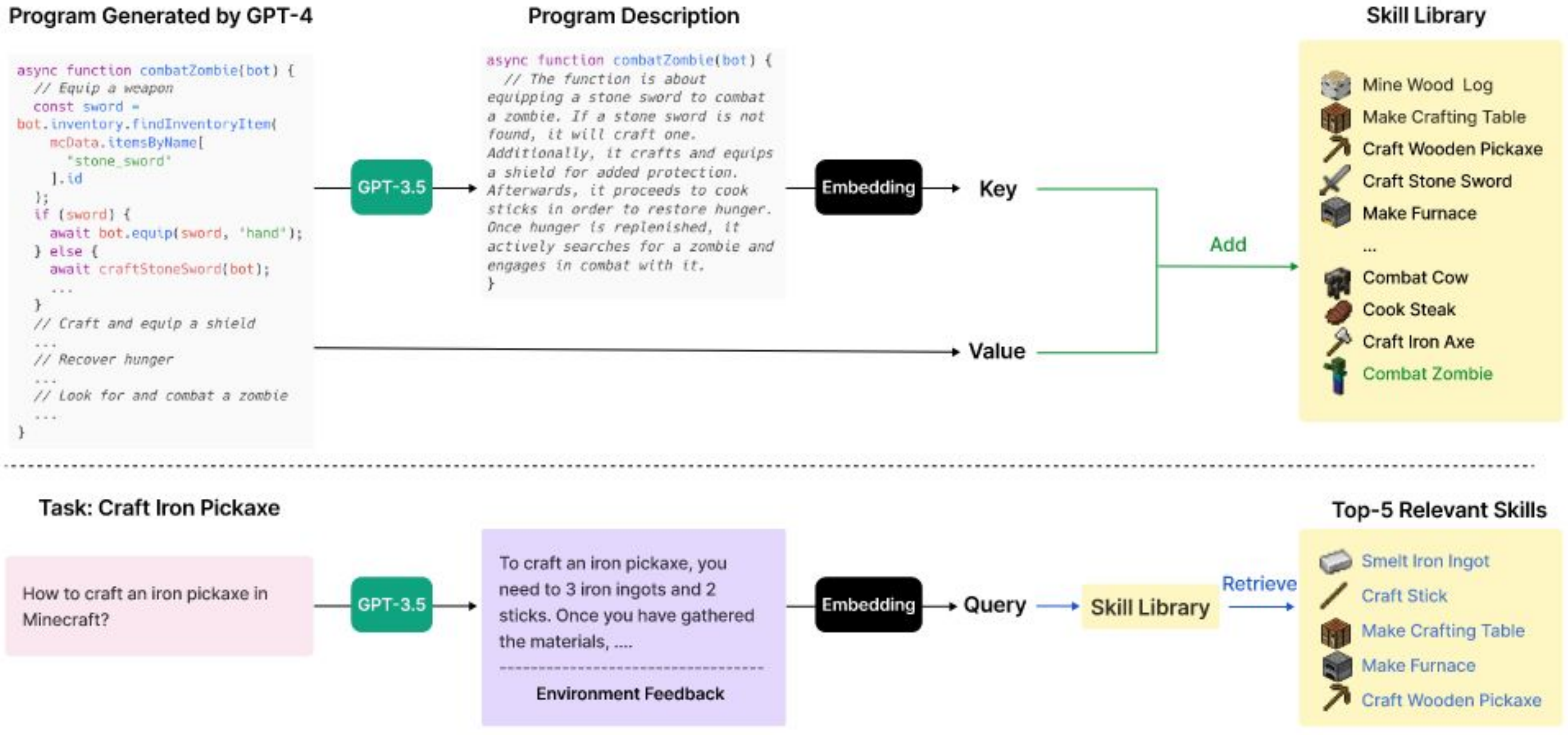
How to craft an iron pickaxe in Minecraft?

To craft an iron pickaxe, you need to 3 iron ingots and 2 sticks. Once you have gathered the materials, ...

Environment Feedback

Top-5 Relevant

- Smelt Iron Ingot
- Craft Stick
- Make Crafting T
- Make Furnace
- Craft Wooden P



Program Generated by GPT-4

```
async function combatZombie(bot) {
  // Equip a weapon
  const sword =
bot.inventory.findInventoryItem(
    mcData.itemsByName[
      "stone_sword"
    ].id
  );
  if (sword) {
    await bot.equip(sword, 'hand');
  } else {
    await craftStoneSword(bot);
    ...
  }
  // Craft and equip a shield
  ...
  // Recover hunger
  ...
  // Look for and combat a zombie
  ...
}
```

Program Description

```
async function combatZombie(bot) {
  // The function is about
equipping a stone sword to combat
a zombie. If a stone sword is not
found, it will craft one.
Additionally, it crafts and equips
a shield for added protection.
Afterwards, it proceeds to cook
sticks in order to restore hunger.
Once hunger is replenished, it
actively searches for a zombie and
engages in combat with it.
}
```

Skill Library

- Mine Wood Log
- Make Crafting Table
- Craft Wooden Pickaxe
- Craft Stone Sword
- Make Furnace
- ...
- Combat Cow
- Cook Steak
- Craft Iron Axe
- Combat Zombie

Task: Craft Iron Pickaxe

How to craft an iron pickaxe in Minecraft?

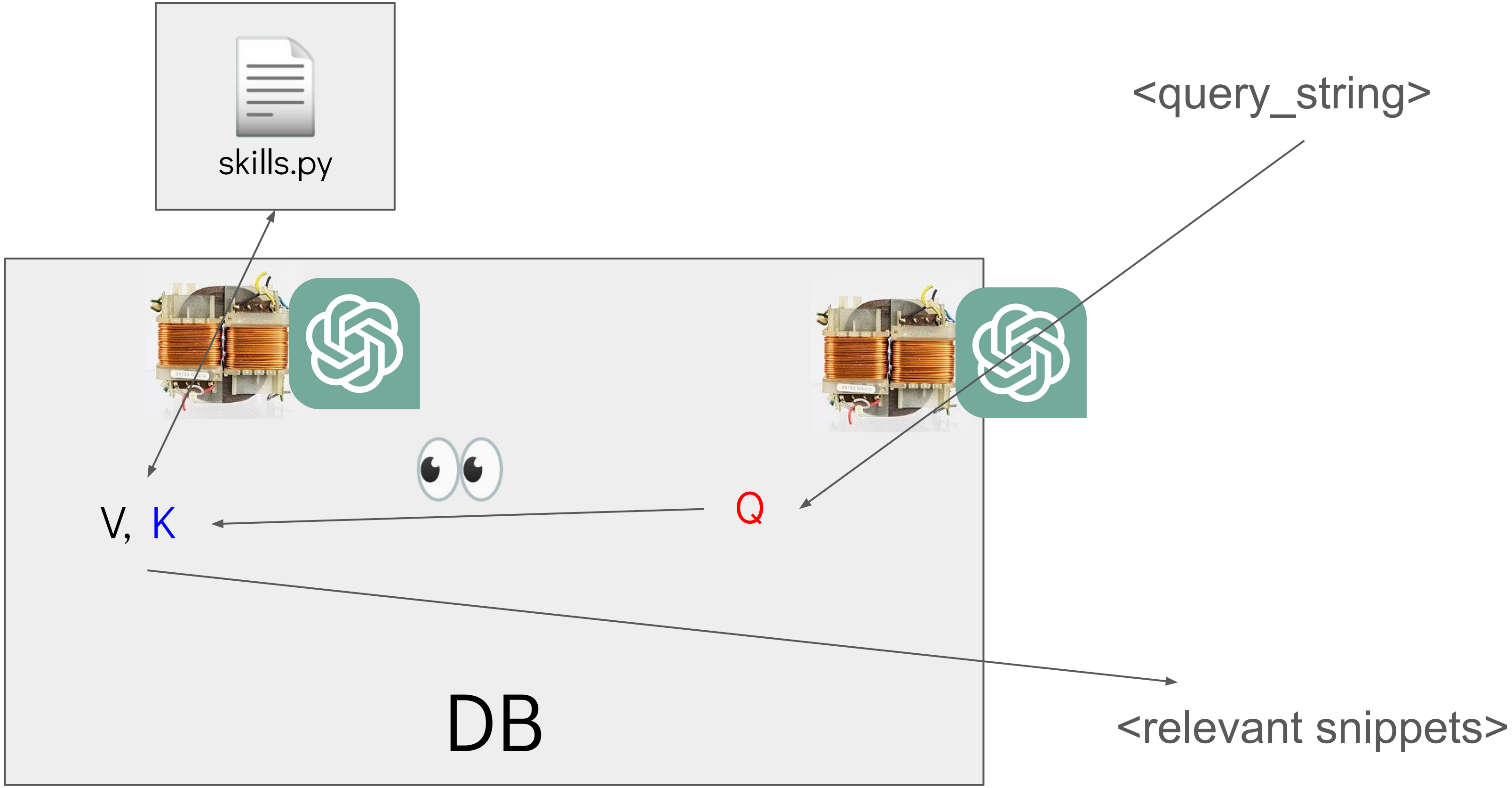
To craft an iron pickaxe, you need to 3 iron ingots and 2 sticks. Once you have gathered the materials, ...

Environment Feedback

Top-5 Relevant Skills

- Smelt Iron Ingot
- Craft Stick
- Make Crafting Table
- Make Furnace
- Craft Wooden Pickaxe

Memory in external database (Transformer-based retrieval)



Memory in external database

the Good:

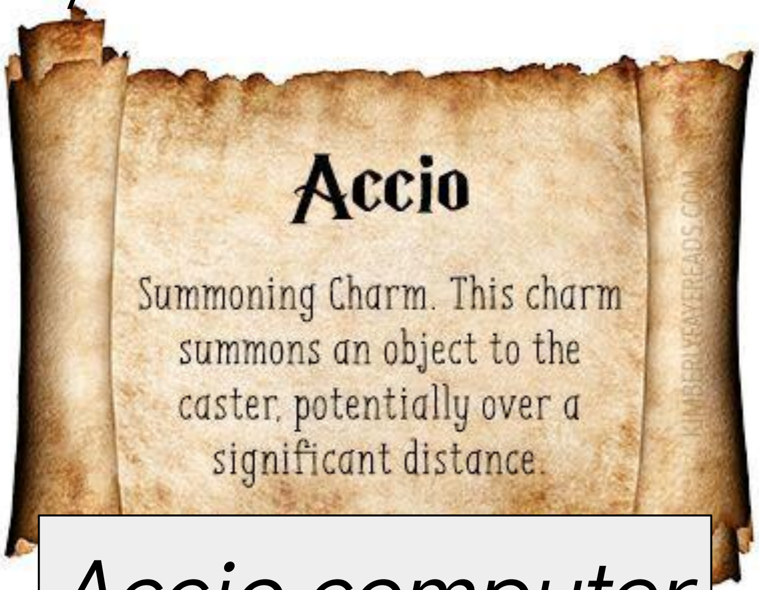
- Infinite storage

the Bad:

- Needs to build DB
- DB can be tricky to maintain

- Is there already an existing large DB on the shelf?

Scenario: Hogwarts have wifi and nothing in the rules says you cannot summon a computer



Accio

Summoning Charm. This charm summons an object to the caster, potentially over a significant distance.

Accio computer

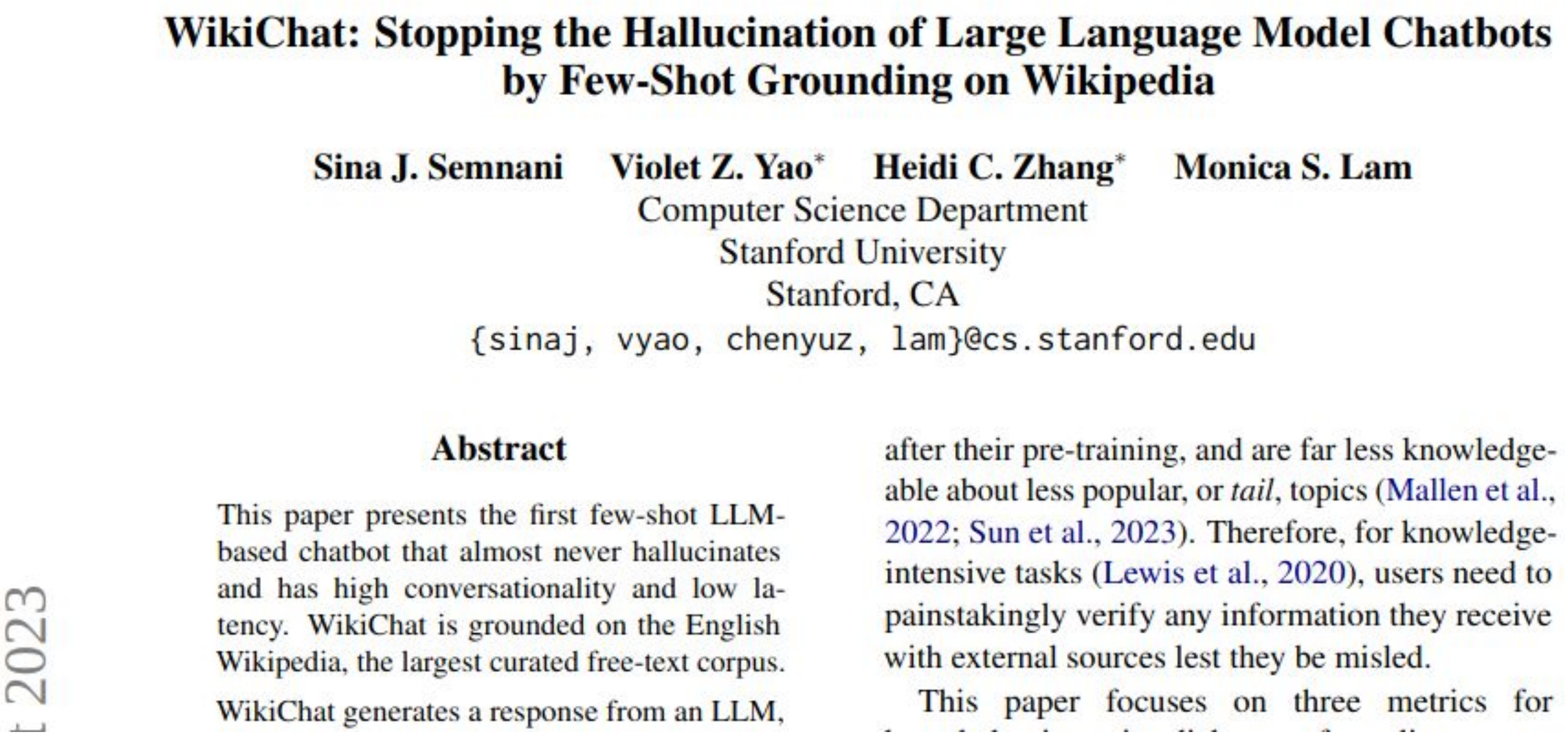
11:45 AM
://www.google.co...
Images Local News more
Google
Google Search
Location unavailable update
Sign in Help
iGoogle Settings Classic
Google in: Mobile | Classic

Outline

- Problem: Information retrieval
- Solution 0: Memory in training data
- Solution 1: Memory in LLM Context
- Solution 2: Compress chat history
- Solution 3: Memory in external database
- Solution 4: Let LLM browse the internet

Let LLM browse the internet

- Lookup Google
- Lookup Wikipedia
  - Can do “fact checks” too!



WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia

Sina J. Semnani   Violet Z. Yao*   Heidi C. Zhang*   Monica S. Lam
Computer Science Department
Stanford University
Stanford, CA
{sinaj, vyao, chenyuz, lam}@cs.stanford.edu

Abstract

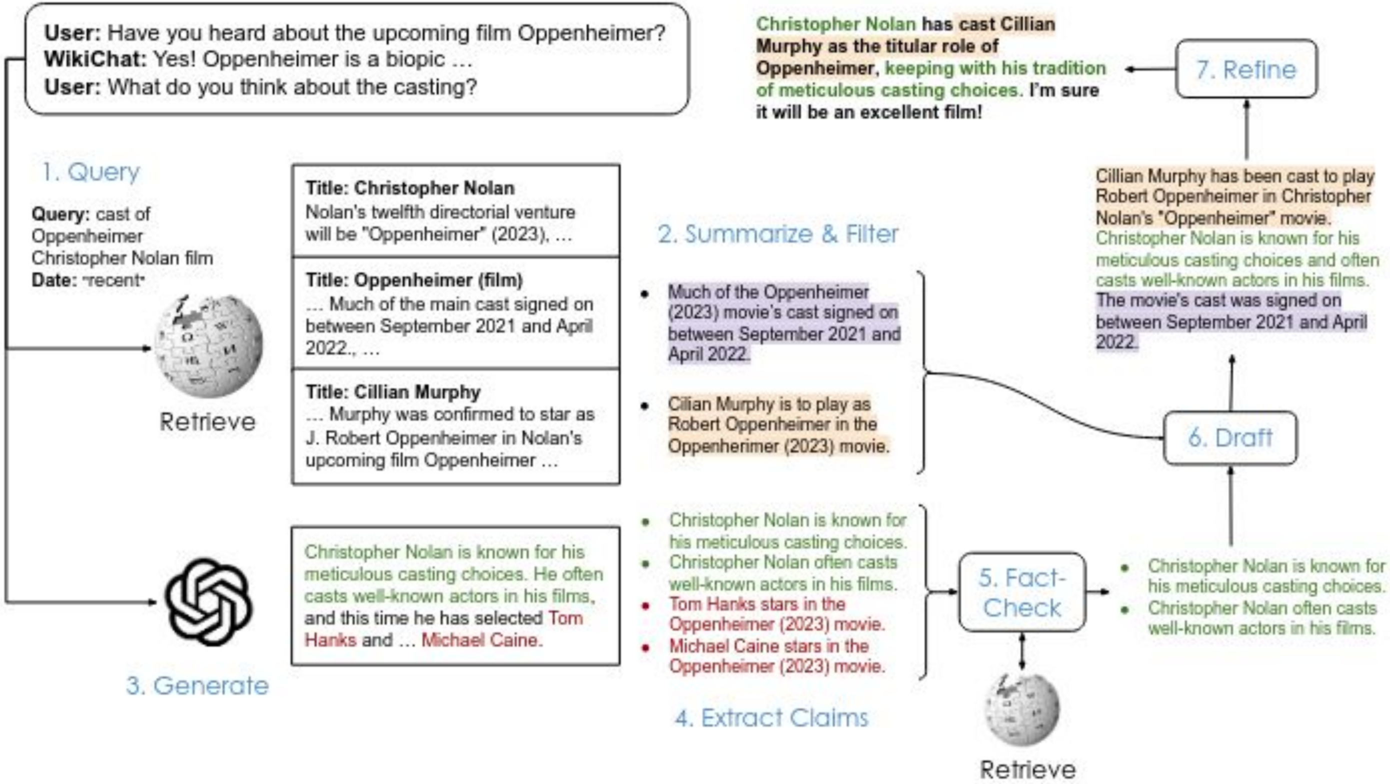
This paper presents the first few-shot LLM-based chatbot that almost never hallucinates and has high conversationality and low latency. WikiChat is grounded on the English Wikipedia, the largest curated free-text corpus.

WikiChat generates a response from an LLM,

after their pre-training, and are far less knowledgeable about less popular, or *tail*, topics (Mallen et al., 2022; Sun et al., 2023). Therefore, for knowledge-intensive tasks (Lewis et al., 2020), users need to painstakingly verify any information they receive with external sources lest they be misled.

This paper focuses on three metrics for

https://arxiv.org/pdf/2305.14292v2.pdf



User: Have you heard about the upcoming film Oppenheimer?
WikiChat: Yes! Oppenheimer is a biopic …
User: What do you think about the casting?

1. Query

Query: cast of Oppenheimer Christopher Nolan film
Date: "recent"

Retrieve

Title: Christopher Nolan
Nolan's twelfth directorial venture will be "Oppenheimer" (2023), …

Title: Oppenheimer (film)
… Much of the main cast signed on between September 2021 and April 2022., …

Title: Cillian Murphy
… Murphy was confirmed to star as J. Robert Oppenheimer in Nolan's upcoming film Oppenheimer …

2. Summarize & Filter

- Much of the Oppenheimer (2023) movie's cast signed on between September 2021 and April 2022.
- Cilian Murphy is to play as Robert Oppenheimer in the Oppenherimer (2023) movie.

3. Generate

Christopher Nolan is known for his meticulous casting choices. He often casts well-known actors in his films, and this time he has selected Tom Hanks and … Michael Caine.

4. Extract Claims

- Christopher Nolan is known for his meticulous casting choices.
- Christopher Nolan often casts well-known actors in his films.
- Tom Hanks stars in the Oppenheimer (2023) movie.
- Michael Caine stars in the Oppenheimer (2023) movie.

5. Fact-Check

Retrieve

- Christopher Nolan is known for his meticulous casting choices.
- Christopher Nolan often casts well-known actors in his films.

6. Draft

Cillian Murphy has been cast to play Robert Oppenheimer in Christopher Nolan's "Oppenheimer" movie. Christopher Nolan is known for his meticulous casting choices and often casts well-known actors in his films. The movie's cast was signed on between September 2021 and April 2022.

7. Refine

Christopher Nolan has cast Cillian Murphy as the titular role of Oppenheimer, keeping with his tradition of meticulous casting choices. I'm sure it will be an excellent film!

Let LLM browse the internet

the Good:

- Simple to implement
- Cheap computation (on our side)

the Bad:

- Beware of internet trolls
- Dependent on third party servers

Outline

- Problem: Information retrieval
- Solution 0: Memory in training data
- Solution 1: Memory in LLM Context
- Solution 2: Compress chat history
- Solution 3: Memory in external database
- Solution 4: Let LLM browse the internet
- Q&A

Exercises:

- Mistral-7B-v0.1 uses sliding window attention. What's its window length & number of layers? What's its theoretical attention span?
- Can you suggest some reasons that the attention span in practice is lower than theoretical span? How to remedy?
- What are some of the pain points of depending on third-party search engines?

References

- https://mistral.ai/news/mixtral-of-experts/
- https://huggingface.co/blog/moe
- S Semnani, et al., WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia. https://arxiv.org/abs/2305.14292
- https://paperswithcode.com/method/sliding-window-attention
- https://huggingface.co/blog/4bit-transformers-bitsandbytes
- https://paperswithcode.com/method/grouped-query-attention
- PagedAttention: https://arxiv.org/abs/2309.06180
- https://voyager.minedojo.org/
- https://mistral.ai/news/announcing-mistral-7b/

Image Credits

- https://dwellingondreamspodcast.com/2019/08/28/classes-at-hogwarts-we-wish-wed-heard-more-about/
- https://imgflip.com/memetemplate/124827840/You-know-nothing
- https://twitter.com/TheHPfacts/status/848409690266533888
- https://harrypotter.fandom.com/wiki/Ordinary_Wizarding_Level
- https://harrypotter.fandom.com/wiki/Hogwarts_Library
- https://twitter.com/HPotterUniverse/status/628967209252032512
- https://uxdesign.cc/what-kind-of-designer-researcher-are-you-2da7598b17d4
- https://www.independent.co.uk/life-style/health-and-families/features/why-being-messy-can-be-a-positive-trait-according-to-researchers-a6774206.html
- https://gradecalculator.mes.fm/memes/fingernail-cheat-sheet
- https://www.gsm-earpiece.com/howto/tips-on-cheating-exam/
- https://en.wikipedia.org/wiki/Transformer