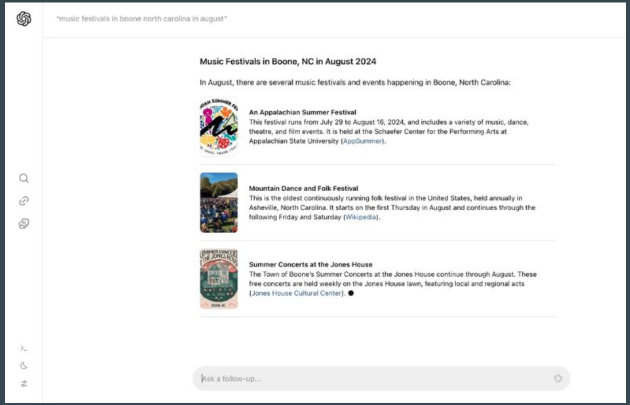
Temporality of LLMs

Dr. Wei Zhao
15/10/2024

Why this topic?

- Timing
  - Perplexity AI and SearchGPT
  - No temporal component
- Impacts
  - Hundreds of millions of users
  - ~30% user queries are time-sensitive (Archive Query Log)
  - Public disappointment
  - AI winter
  - ..



Why this topic?

- This topic is hard
  - Challenges
    - temporal hallucination and causes
    - temporal trustworthiness
    - temporal complexity
    - temporal evaluation
    - temporal dynamics (forecasting)
    - temporal agents (autonomous update)
    - grounding LLMs in time

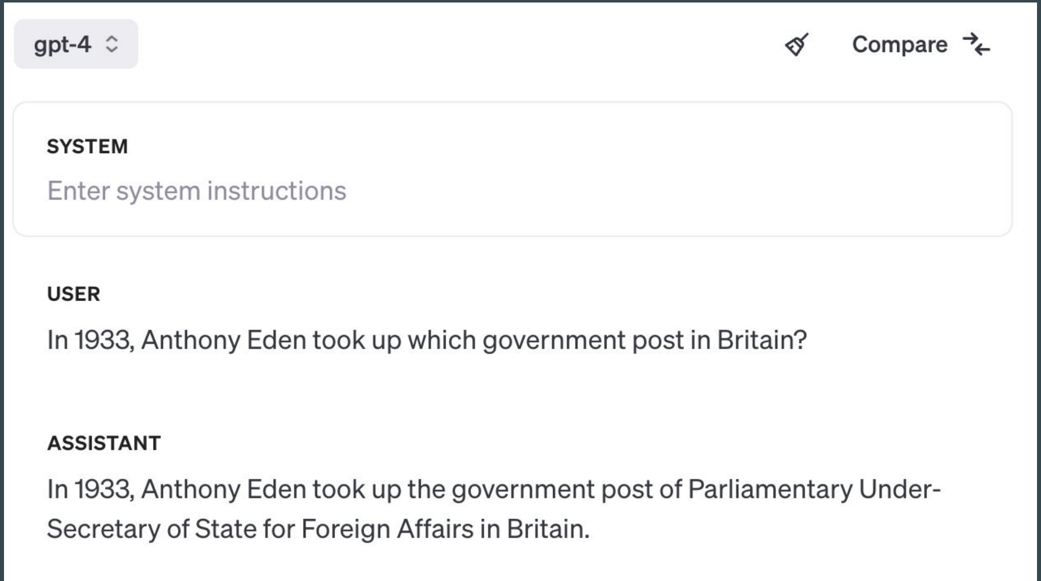
Temporal challenges

•••

Temporal Hallucination and Complexity

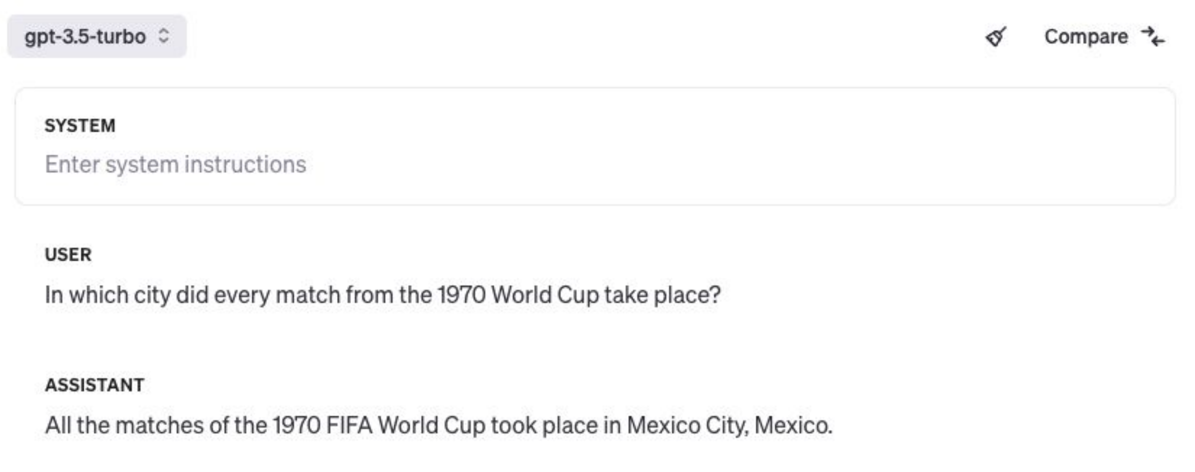
Temporal Hallucination

- Fabrication
  - For a query with no answer, LLMs invent a false answer



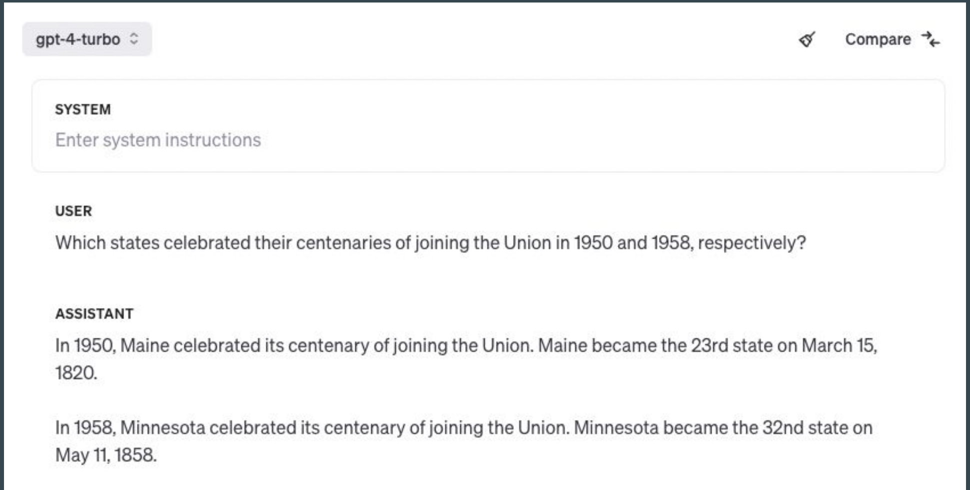
Temporal Hallucination

- Omission
  - For a query with multiple answers, LLMs provide an incomplete answer
    - GT answer: Mexico City, Guadalajara, León, Puebla, Toluca



Temporal Hallucination

- Omission
  - For a query with multiple answers, LLMs provide a partly correct answer
    - GT answer: California, Minnesota

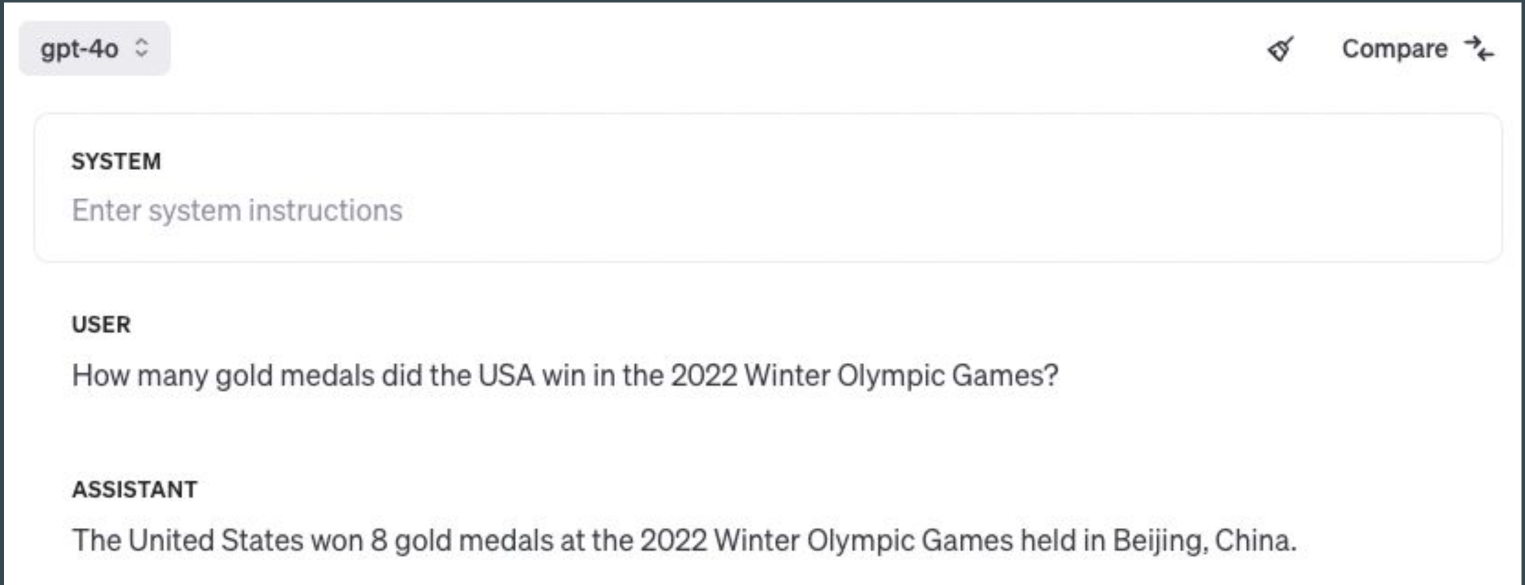


Temporal Hallucination

- Misattribution
  - For a query with the answer to be a proper noun (person name/profession, etc), LLMs attribute to a wrong entity
    - GT answer: Secretary of State for Foreign Affairs

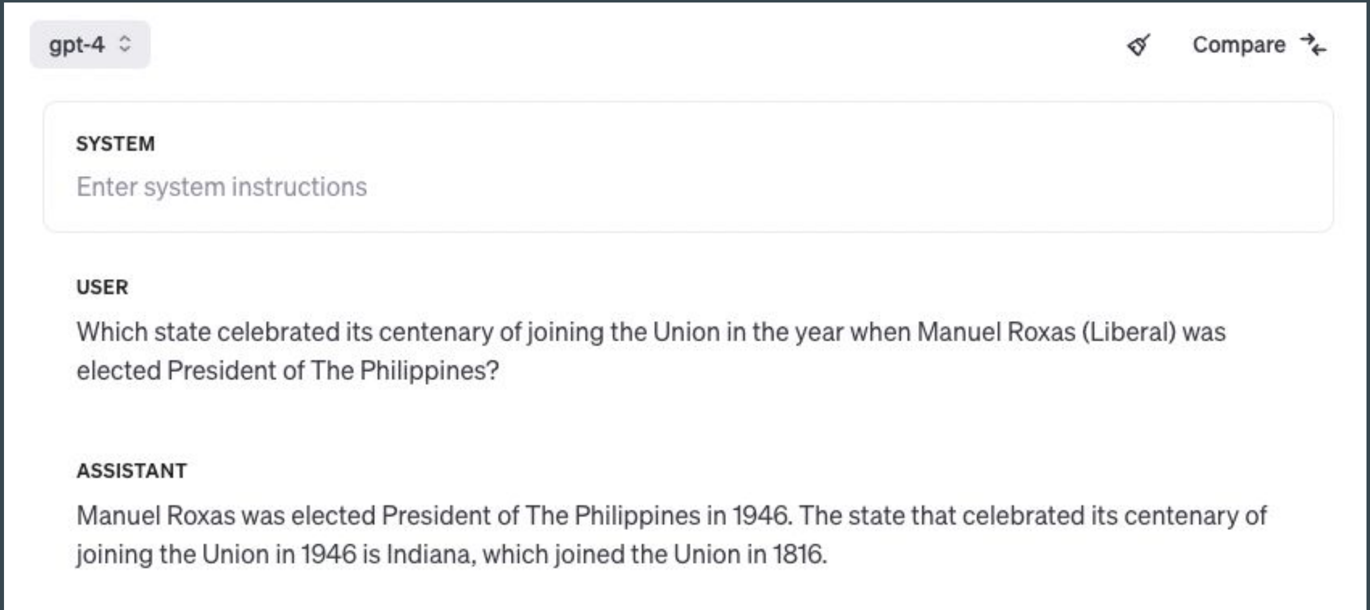
Temporal Complexity

- 1-hop: timestamp is stated explicitly
  - GT answer: 9 gold medals



Temporal Complexity

- 2-hop: timestamp is stated implicitly
  - GT answer: Iowa, American state



Project results

...

Research questions

- RQ1 [Data]: How we benchmark temporality of LLMs?
- RQ2 [LLMs]: What are their temporalities?
- RQ3 [LLMs]: How do their temporalities change over time?
  - How do different aspects of hallucination change over time?
  - Can LLMs answer a question with knowledge change over time?

Evaluation setup

- Models
  - GPT-4, GPT-4o, GPT-3.5
    - closed-source
    - versatile across domains and modalities
  - Claude-3.5, Claude-3
    - closed-source
    - safety and ethical use
  - Llama-3-70b, Llama-3-8b
    - open-source

Evaluation setup

- Evaluation Metrics
  - Exact Match (True or False)
    - true if GT contained in model answer
  - F1 (Precision and Recall)
    - word overlap between GT and model answer
  - PEDANTS (Li et al 2024)
    - g(F1, prec, recall, query, GT, model answer)
    - good correlation with human on QA datasets

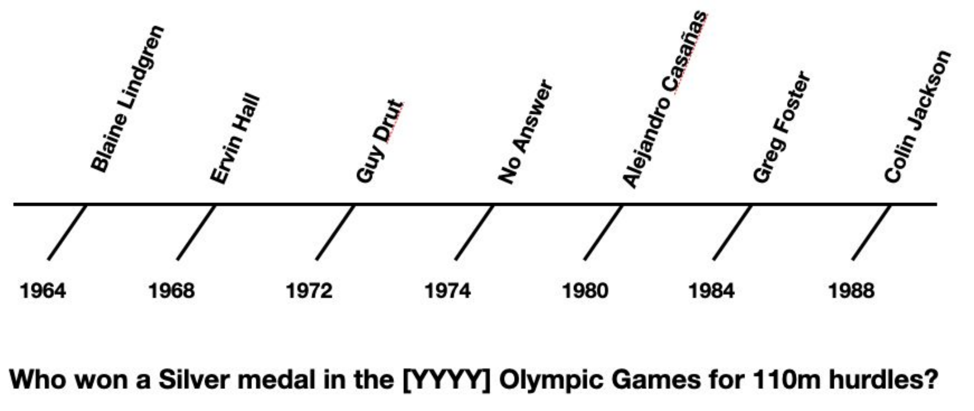
Dataset

Our dataset in a nutshell (RQ1)

- A temporal benchmark dataset for time-sensitive queries
  - temporal extension of the TriviaQA dataset
  - semi-automatic human annotation
  - 22 decade groups from 1330s to now (likely part of training data)
  - #answers: 0, 1, 1+
    - fabrication and omission
  - answer type: person, location, organization, etc
    - misattribution
  - temporal complexity: 1-hop and 2-hop
    - implicit vs. explicit timestamp
  - domains: history, geography, science, sports, entertainment, etc.

RQ1: Data construction

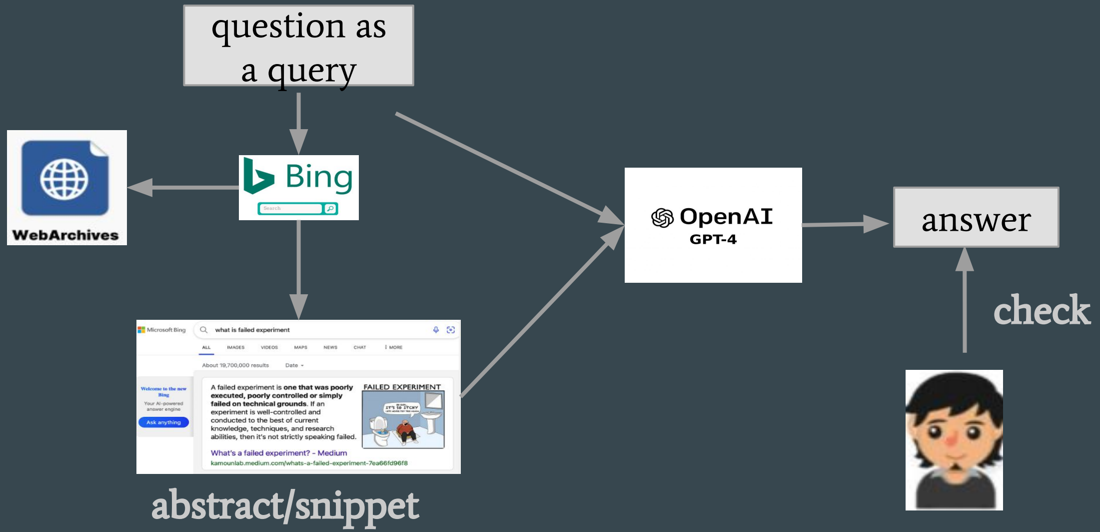
- Procedure for dataset construction
  1. select temporal questions from TriviaQA
  2. filter out questions if their answers do not change over time
  3. expand questions at different points in time



person, 1-hop, points in time, 1960s-1980s, 0 or 1 answer, sports

Data construction in a nutshell

4. annotate answers for the questions (semi-automatic process)
    - twice cheaper than human annotation, but Bing API access is expensive: free to use for only 1000 web queries per month



Data construction in a nutshell

- Procedure for dataset construction
  5. generate questions
     - with no answers to test fabrication
       - lifecycle of knowledge
     - with multiple answers to test omission
       - merge different time points in time into a time period

Data construction in a nutshell

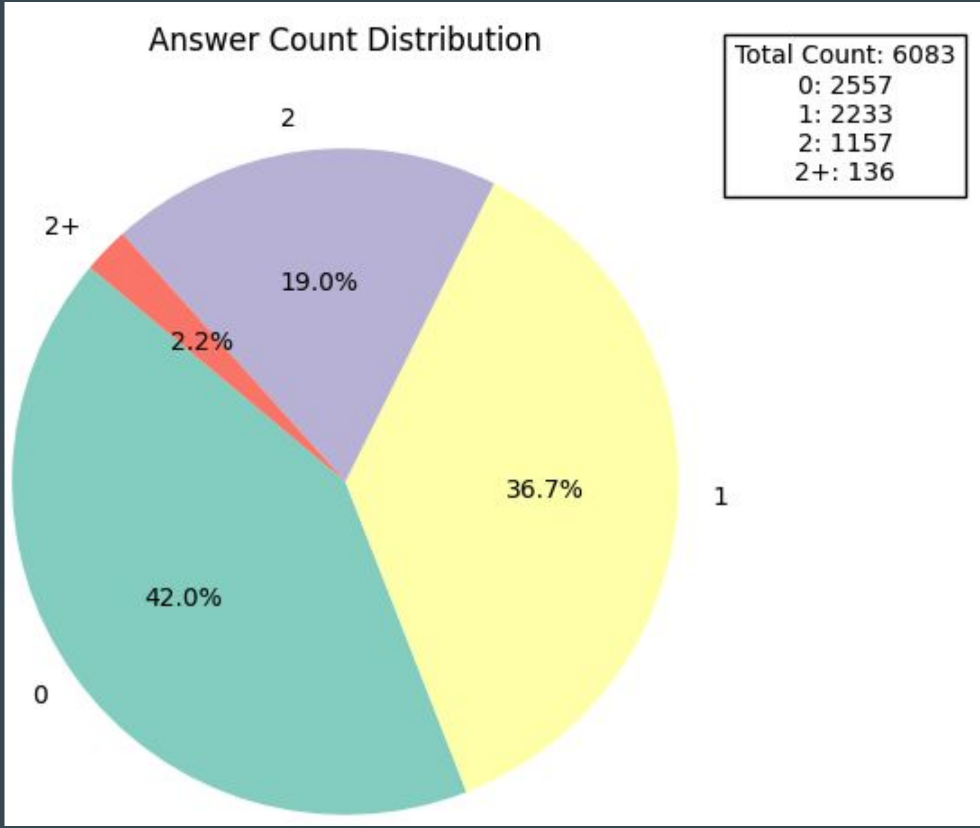
6. generate two-hop questions (time is stated implicitly)
   - converted from 1-hop questions
     - pair 1-hop questions dated in the same year
     - merge into a two-hop question (LLM + human)
7. Assign an answer type to each question (LLM + human)
   - person name, location, number, time, etc

Data statistics

- Aim for ~20,000 QA pairs, based on 500 QA pairs from TriviaQA
- current results are based on ~6,000 QA pairs

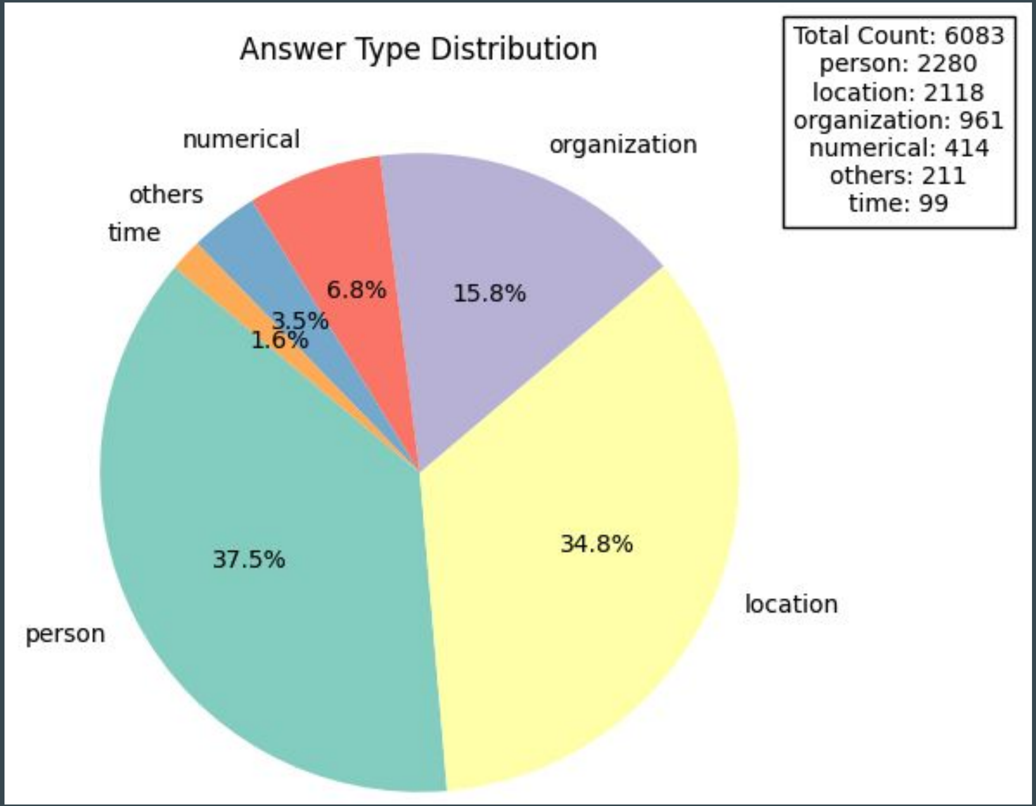
Data statistics

- Fabrication:
  - #answer = 0
- Omission:
  - #answer > 1



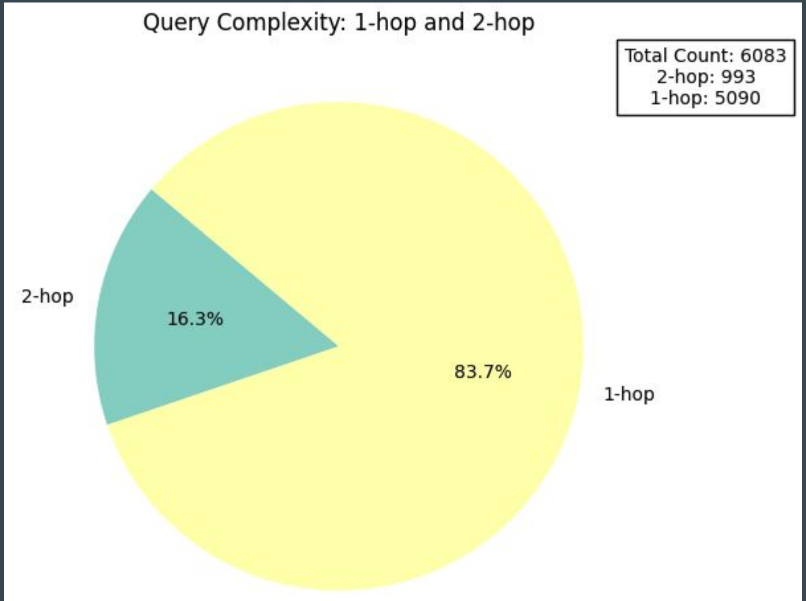
Data statistics

- Misattribution
  - Organization
  - Location
  - Person name



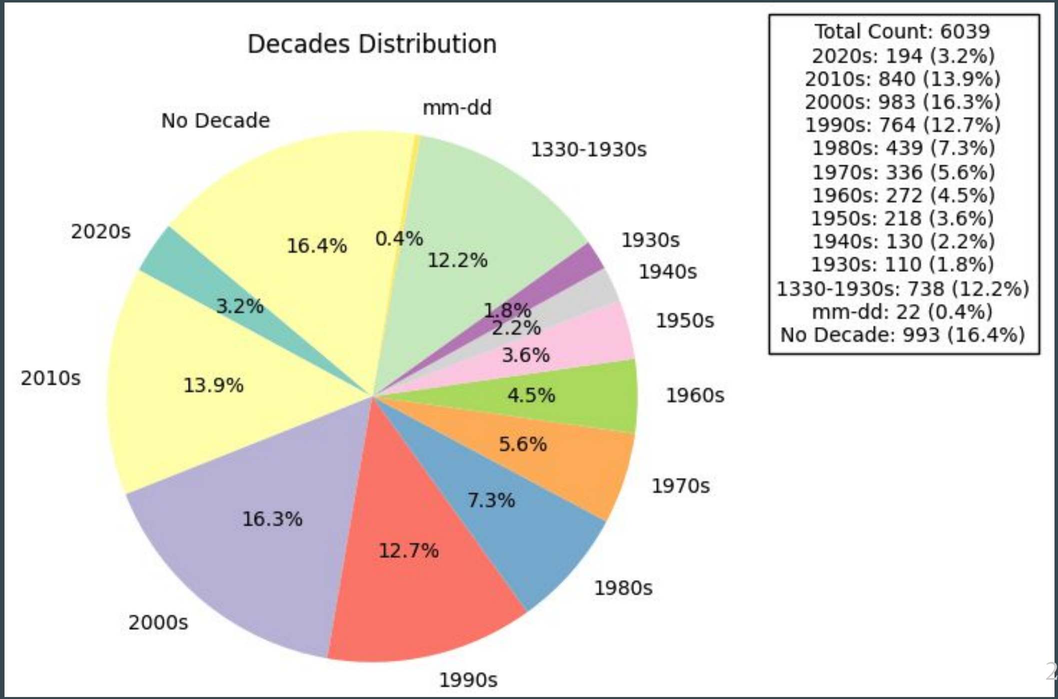
Data statistics

- Query complexity
  - 1-hop
    - explicit time
  - 2-hop
    - implicit time



Data statistics

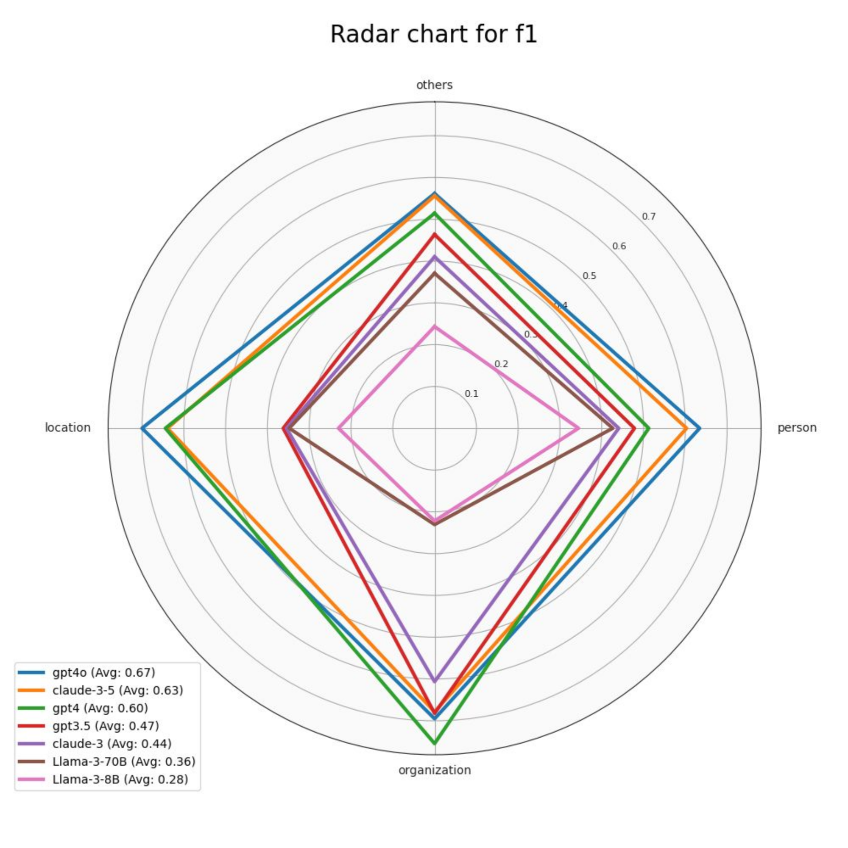
- Explicit timestamp (1-hop)
    - 1330s - 2020s
    - MM-DD
- Implicit timestamp
    - No decade (2-hop)



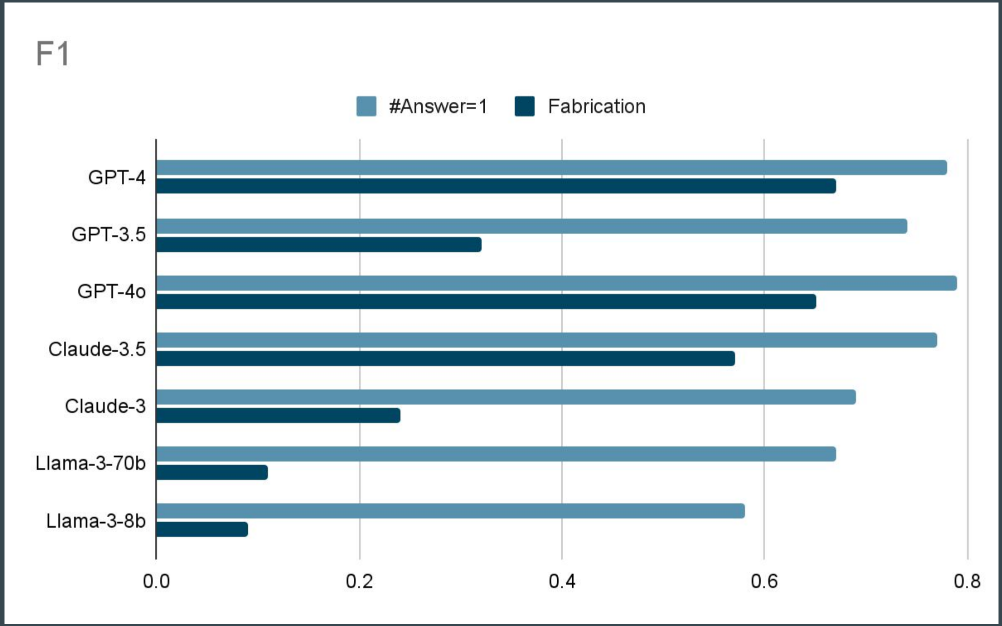
Analyses

RQ2: Misattribution - model temporality

- Model scaling **super expensive but helps little** ,
  - Llama-3-70b is still bad despite its large model size
  - person and others are harder to attribute
  - **clever and cost-efficient ideas?**
  - **fundamental reason for bad temporality?**
- minor point
  - GPT3.5 => GPT4, 'location' gets much better, why? insights will be useful for improving certain aspects.

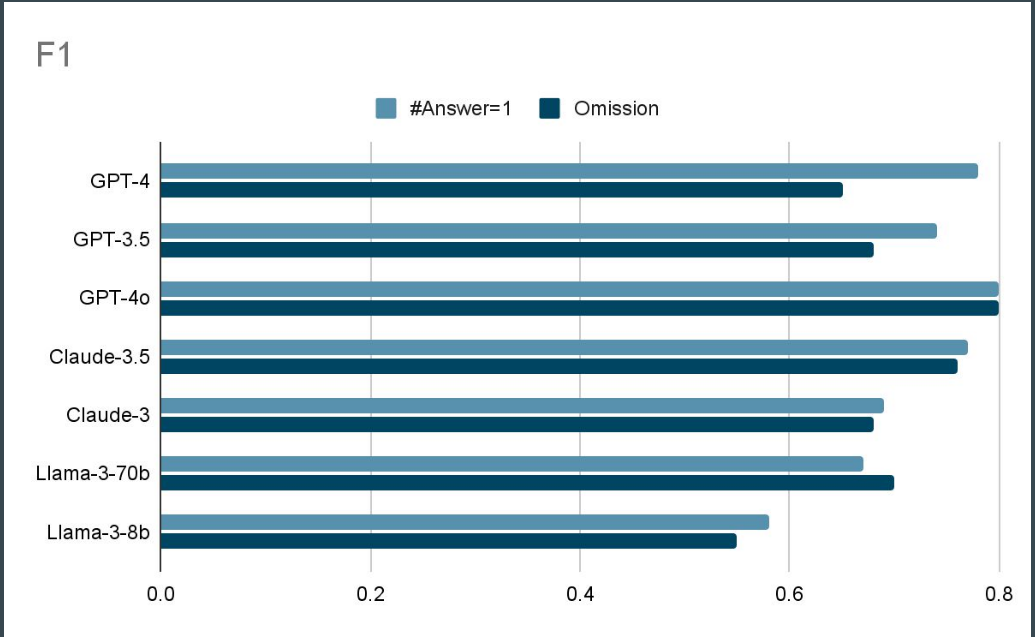


RQ2: Fabrication - model temporality



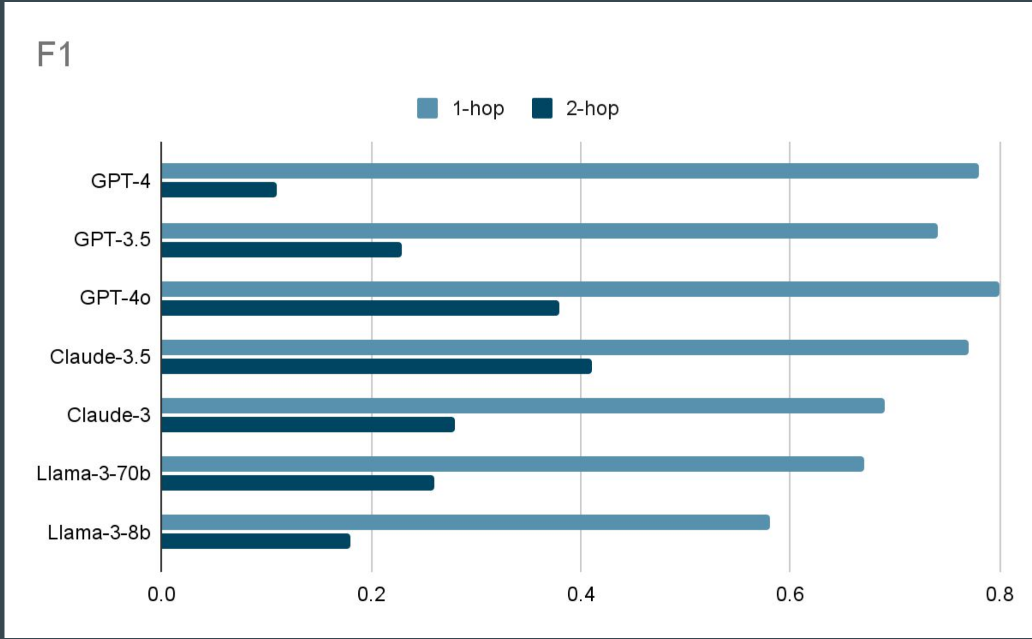
- For a query with no answer, LLMs invent a false answer.
- Higher F1 means low fabrication rate
- Good when a query has answers, bad when a query hasn't.
- <u>Model scaling helps little;</u> see Llama-3-70b vs. -8b

RQ2: Omission - model temporality



- For a multi-answer question, LLMs give incomplete answer.
- Higher F1 means lower omission
- MAQs are harder than SAQ, LLM does a good job - many MAQs are generated from SAQs.
- Surprisingly, Llama-3-70b is better in omission than in SAQ, despite the former being harder.

RQ2: One-hop vs. two-hop - model temporality



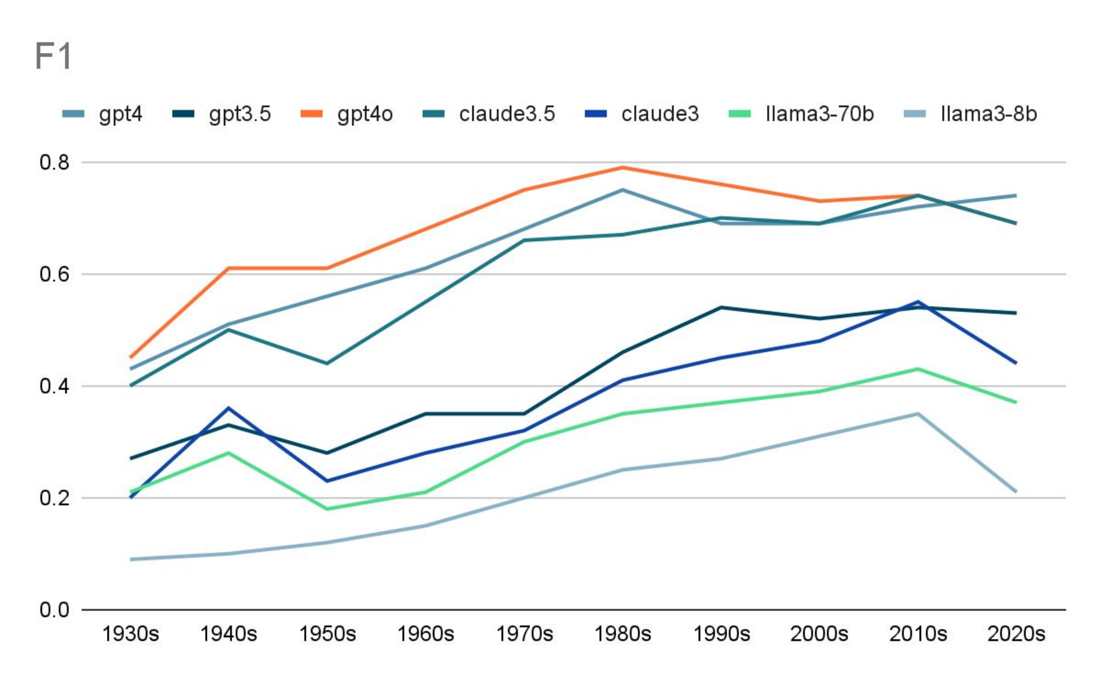
- Two-hop queries are much harder
- Claude-3.5 seems most robust
- More results are in the making (fabrication, omission, etc)

Temporal results

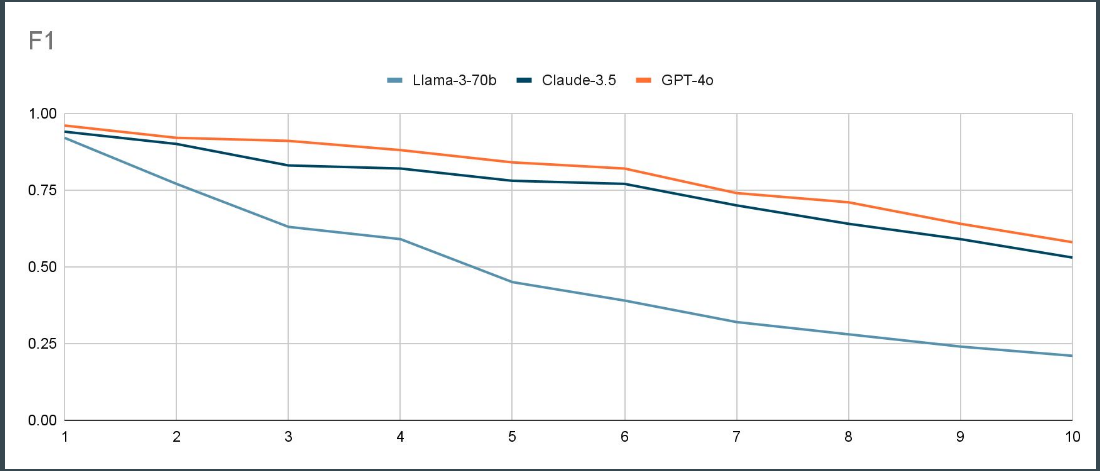
...

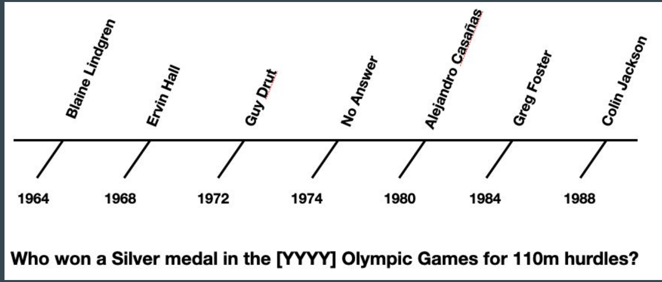
RQ3: Temporality changing over time

- 100 questions per group
- 10 decade groups
- poor results in distant past
  - shortage of historical data
- poor results in 2020s
  - knowledge cut-off
- better results over time
  - data volume becomes bigger over time
- Can we improve LLMs brain in the distant past



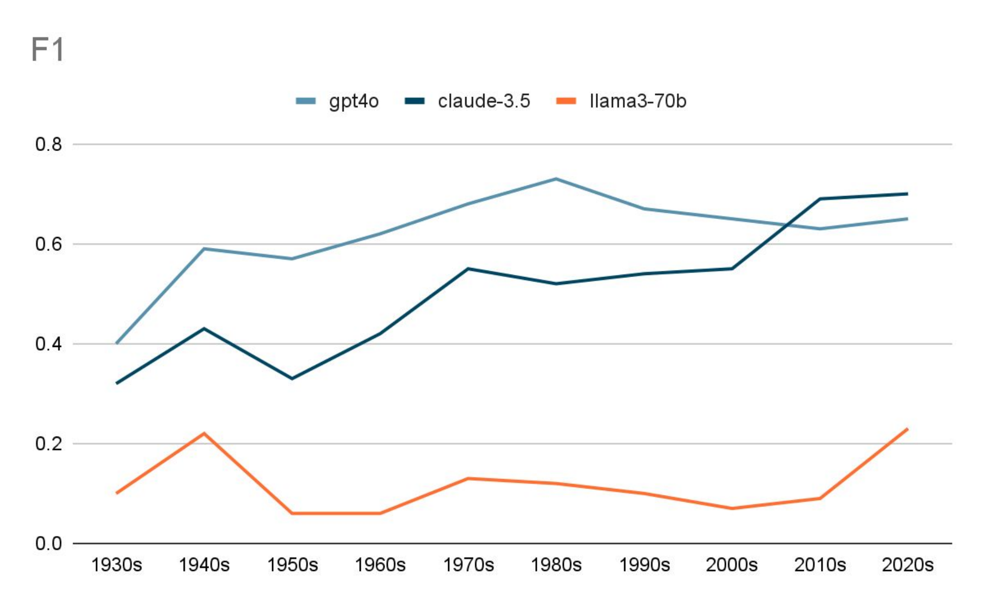
RQ3: Question with knowledge/answer change over time





- x-axis: number of times LLMs answer correctly for time-sensitive answers.
- Each question has answers changing over 10 different points in time.
  - answer correctly at one point in time vs. various points in time
- time=1, model difference is invisible - they can answer all questions at least one time point
- Over time, performance decreases. LLama-3-70b struggles very much with knowledge change
- no-answer queries are included; sample 10 questions if a piece of knowledge has 10+ questions.

RQ3: Fabrication over time



- High F1 means low fabrication
- uptrend seen for GPT-4o and Claude-3.5 (data volume gets bigger over => LLMs get wiser)
- surge in 2020 - brain of LLM is empty for recent data - easier for LLMs not to provide answer
- no clear uptrend for Llama-3-70b? surge in 1940s?
- 100 questions per group; 10 decade groups

Summary of model temporality

	Fabrication	omission	misattribution	decade groups	knowledge change	data leakage	complexity implicit time
GPT-4o	moderate	relatively low	moderate	poor before 1980s	sensitive	moderate	very poor
Claude-3.5	below moderate	relatively low	moderate	poor before 1980s	sensitive	moderate	very poor
Llama-3-70b	very high	moderate	very high	poor all the time	very sensitive	moderate	very poor

Thank you for your attention

Email