

UNSW SYDNEY | Australia's Global University

ChronosAlign: A Large-Scale, Updatable Benchmark for Decomposing Temporal Alignment in LLMs

Sanjay Govindan, Maurice Pagnucco, Yang Song

Introduction

LLMs are trained on a wealth of knowledge, and contain a myriad of answers, both correct and incorrect

Contextually they have a chaotic sense of time, dispensing with the most likely answer.

Problematic if you care about factual accuracy and recall

Key examples:

GPT2-XL	**Who is the *current* prime minister of Australia?**	**The default answer: Malcolm Turnbull** **A careful probe reveals that the LLM knows Scott Morrison is a more recent Prime Minister.**
LLaMA2-13b	**Who is the *current* prime minister of Japan?**	**The default answer: Yoshihide Suga** **A careful probe reveals that the LLM knows Fumio Kishida is a more suitable answer.**

The idea of current in these statements is spread over large periods of time and doesn't reflect the most recent period these LLMs are trained on. You can try this recall-accuracy exercise with other topics, such as the Superbowl winner, or the latest Grammy's award winners

Temporal Alignment Landscape

Temporal alignment: An ability to consistently recall and reason for a specified point in time.

Distinct from temporal reasoning

An example of temporal alignment

Below are the winners of the past 5 Superbowl's.
Can we align LLMs to a specific period and reference the correct winner?

Date	Winner
February 7, 2021	Tampa Bay Buccaneers
February 13, 2022	Los Angeles Rams
February 12, 2023	Kansas City Chiefs
February 11, 2024	Kansas City Chiefs
February 9, 2025	Philadelphia Eagles

Set the Clock: Temporal Alignment of Pretrained Language Models [1]

Does Time Have Its Place? Temporal Heads: Where Language Models Recall Time-specific Information [2]

Temporal Alignment of Time Sensitive Facts with Activation Engineering [3]

[1](Zhao et al. (2024))

Temporal Alignment Data Issues

Three papers have different isolated datasets, cannot be updated for future use meaning they are potentially useless for future models.

Paper	Year Published	Dataset	Models used
[1] Set the Clock	2024	Taqa (limit is 2024)	LLaMA1-65B, LLaMA2-7B / 13B / 70B, Tülu2-70B, GPT-3 (davinci-002), ChatGPT (gpt-3.5-turbo-0125)
[2] Temporal Heads	2025	Specialised test set (limit is 2024)	Llama-2-7b-chat-hf, Qwen1.5-7B-Chat, Phi-3-mini-4k-instruct
[3] Activation Engineering	2025	Taqa (limit is 2024) Head of Government (limit is 2025)	LLaMA2-7b / 13b / 70b, LLaMA3.1-8b (+ Instruct), Qwen2-7b (+ Instruct), Gemma2-2b (+ Instruct)

In our paper we explore a range of other datasets, highlighting this issue with all of them

Building for the future

None of the datasets on the left can be updated at scale. How does Gemma-4 perform against 2026 data?

Date	Winner
February 7, 2021	**Tampa Bay Buccaneers**
February 13, 2022	**Los Angeles Rams**
February 12, 2023	**Kansas City Chiefs**
February 11, 2024	**Kansas City Chiefs**
February 9, 2025	**Philadelphia Eagles**

ChronosAlign – Overview



Provenance tracked

Wikipedia and Wikidata URLS



Cross domain

Sport, politics and culture



Author Independent

Can be updated by anyone



Scaled

500,000 question-answers

~27,000 distinct questions

ChronosAlign – Method

List of reoccuring events via SPARQL

http://www.wikidata.org/entity/Q462671

List of General Classification Winners of the Tour De France

Time invariant QnA

What was the distance of the Tour De France?
Who is the winner of the Tour De France?

Explicit QnA

In 2024,what was the distance of the Tour De France?
In 2022, who was the winner of the Tour De France?

Relative QnA (LLM rephrased based on invariant)

What was the distance of the most recent Tour De France?
Who is the current winner of the Tour De France?

https://en.wikipedia.org/wiki/List_of_Tour_de_France_general_classification_winners

event_year	event_date	subject	relationship	answer	question
2024.0	None	Tour de France	winner	**Tadej Pogačar**	who was the
2024.0	None	Tour de France	distance	**3,498km**	what was th

1 Data Acqusition

Identify temporal information from Wikidata queries

2 Synthesis

Structure Wikipedia content to be easily interpreted by LLM. Generate time invariant question answer pairs for each period of time

3 Augmentation

Rephrase questions for relative syntax, attach pageview statistics and create categorisation (sport, culture, politics)

4 Maintenance

Given the provenance of the data and user filters

1. **Re-source and extract (Step 2)**
2. **Curate question answer pairs according to filter conditions**
3. **Append rephrased questions, categories and pageview statistics (Step 3)**

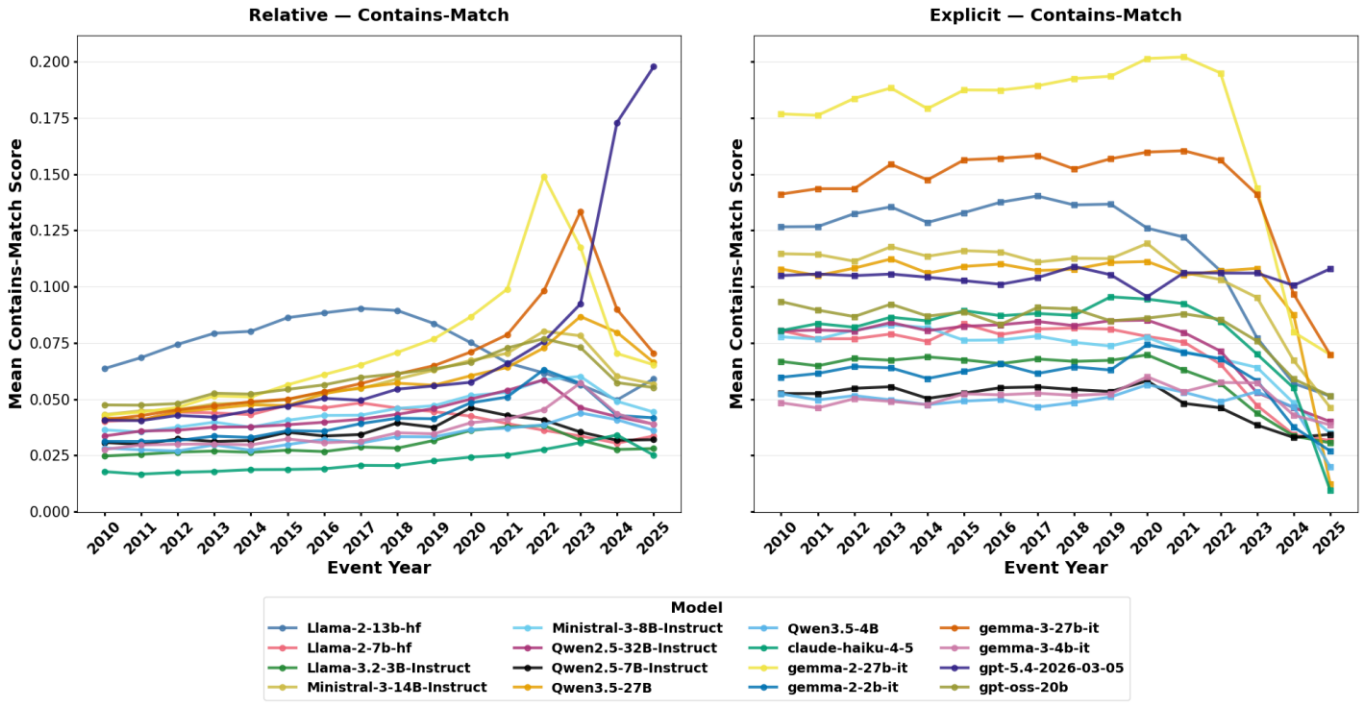
ChronosAlign – Conclusion

Benchmark across the **three modalities** Relative, implicit (not present in the graph) and explicit.

Open AI preferences more recent data and maintains a consistent set of historic data (explicit) [purple]

Anthropic – refusal is a confounder for recall. Safety protocols obfuscate the true recall metric. [light green]

Explicit x relative scores differ wildly – highlights the disparity between known relative and explicit knowledge.



ChronosAlign – Conclusion

500,000 question-answer pairs spanning 25 years.

Provenance tracked and author independent dataset

3 temporal modalities tested – with disparities in knowledge between them

Refusal is found to be a strong confounder for factual recall



GitHub Link



Paper Link



Dataset Link