Understanding and Localizing Time in Language Models

Aditi Bhaskar[1], Suze van Adrichem[1]

[1]Department of Computer Science, Stanford University

Mentor: Jing Huang

Stanford
Computer Science

Motivation

- Facts can be located and edited in models
- Time has not been located and edited in models yet
- Editing time could serve as a sandbox for model editing techniques as it should change (a) associated facts and (b) grammar (output tense)

Example, [Edit model: Current time is 1980]

- (a) Who is the current president? *Jimmy Carter*
- (b) Tell be about 1990 *in 1990 there will be*

Goals & Setup

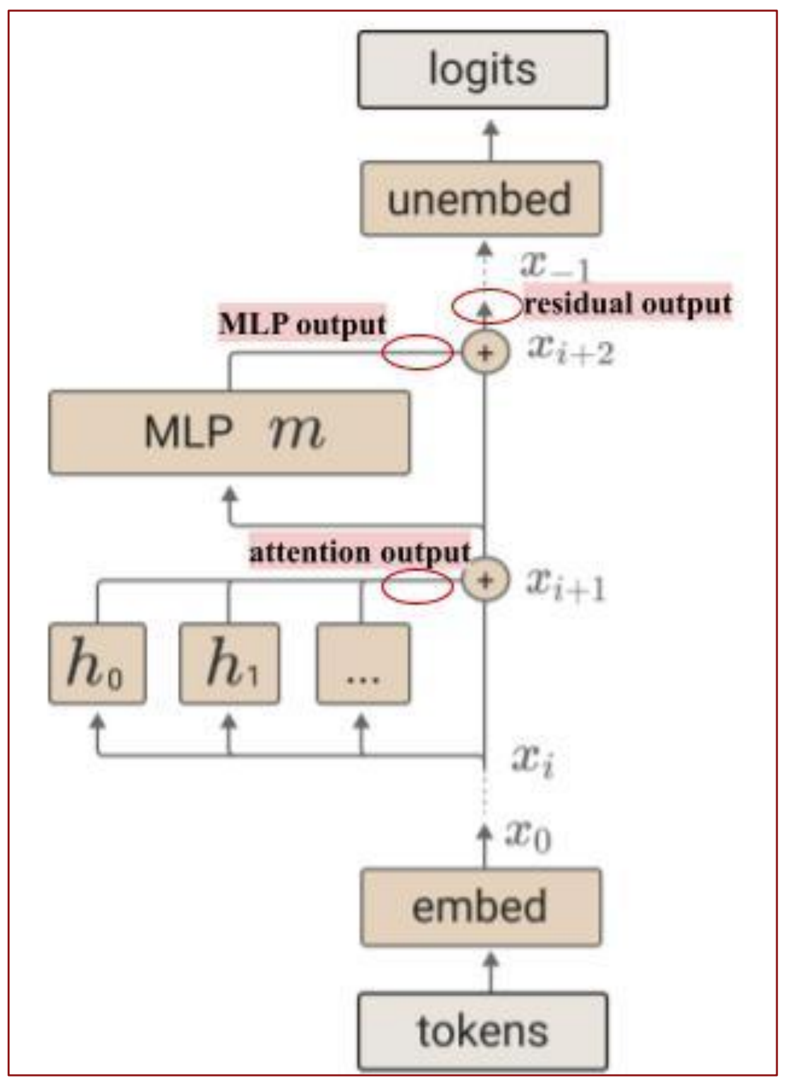
We aim to:

1. Understand time in language models
2. Localize time in language models

We use two autoregressive, next-token prediction models: OLMo 1B [1] and Llama 3.2 1B [2].

Representation patching

Fig 1a. (right) Residual block in transformer [3] showing MLP, attention and residual output representations.



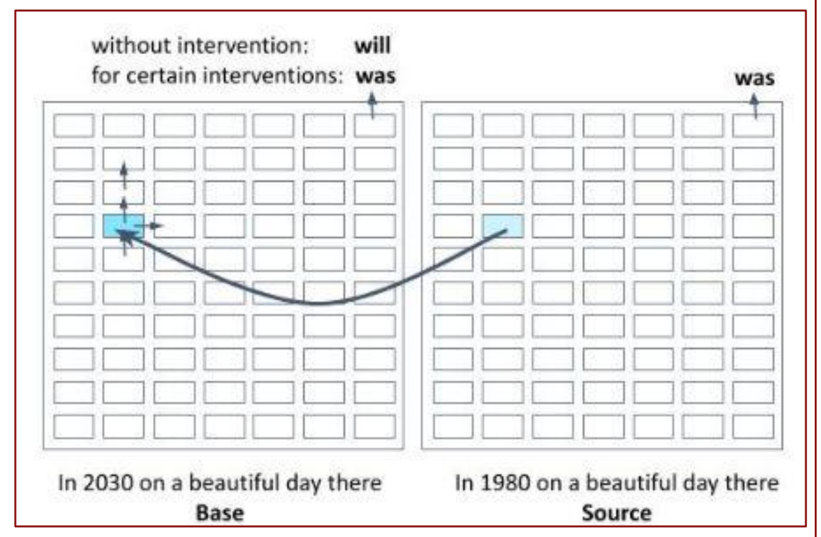

without intervention: will
for certain interventions: was
was
In 2030 on a beautiful day there
Base
In 1980 on a beautiful day there
Source

Fig 1b. (above) Interchange Interventions [5]. (left) base run, (right) source run. Representations in base run are patches with ones from the source run

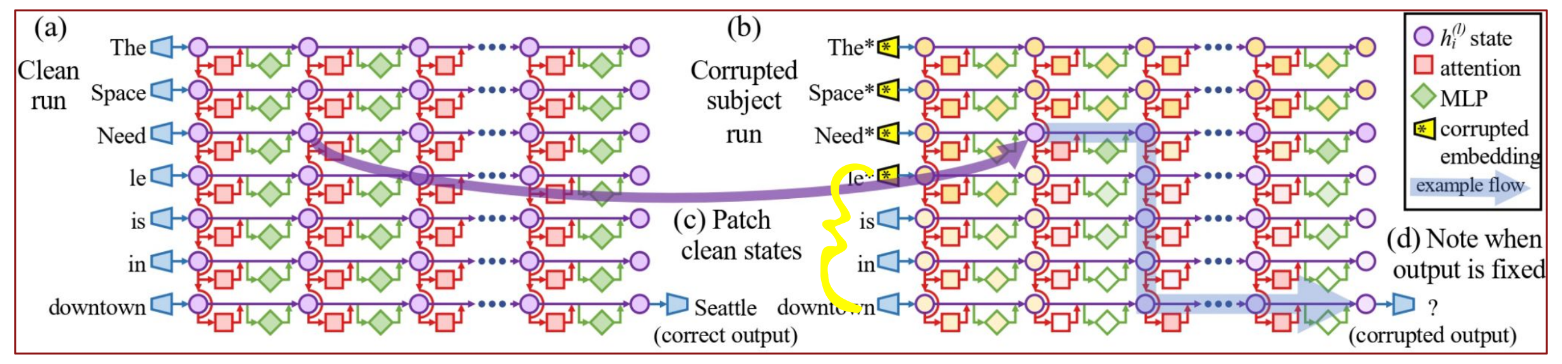


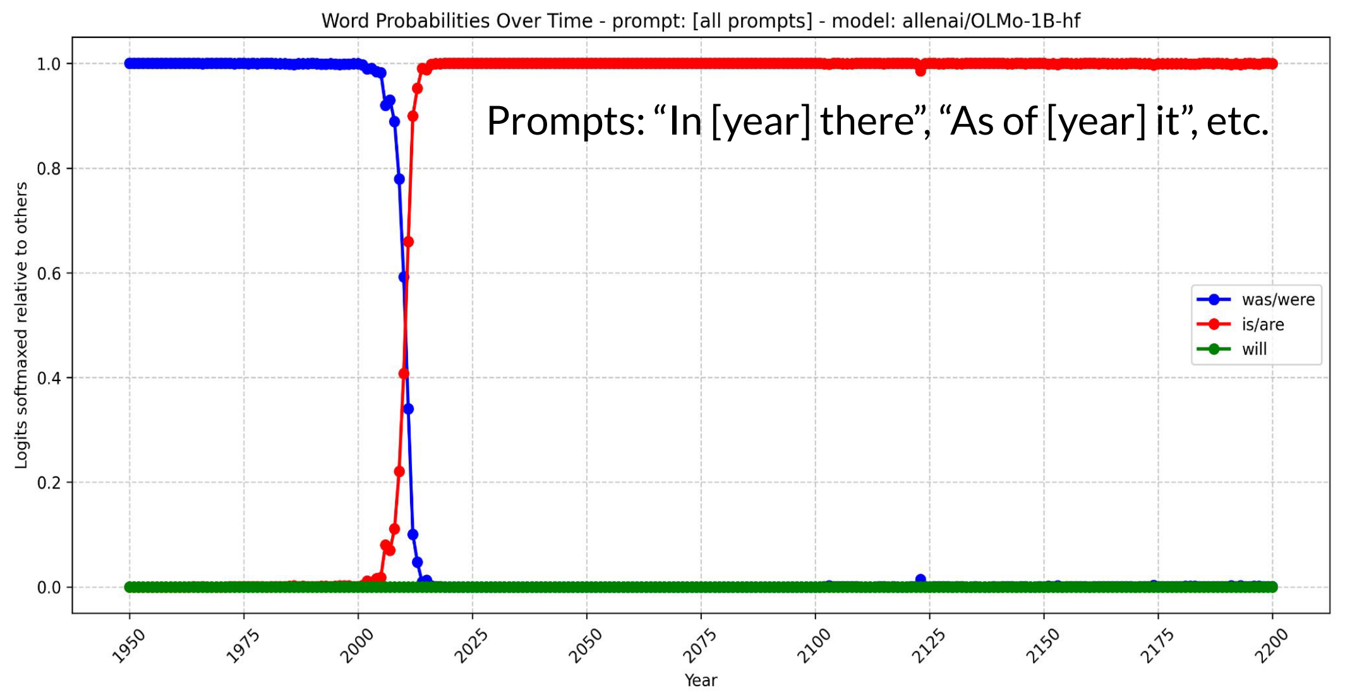
Fig 1c. (above) Casual Tracing [4]. (left) clean run, (right) corrupted run (with noisy input tokens). Representations in corrupted run are patched with ones from the clean run.

Experiments

Understanding Time in Language Models

We use prompts, sweeping years, to understand current time & temporal associations

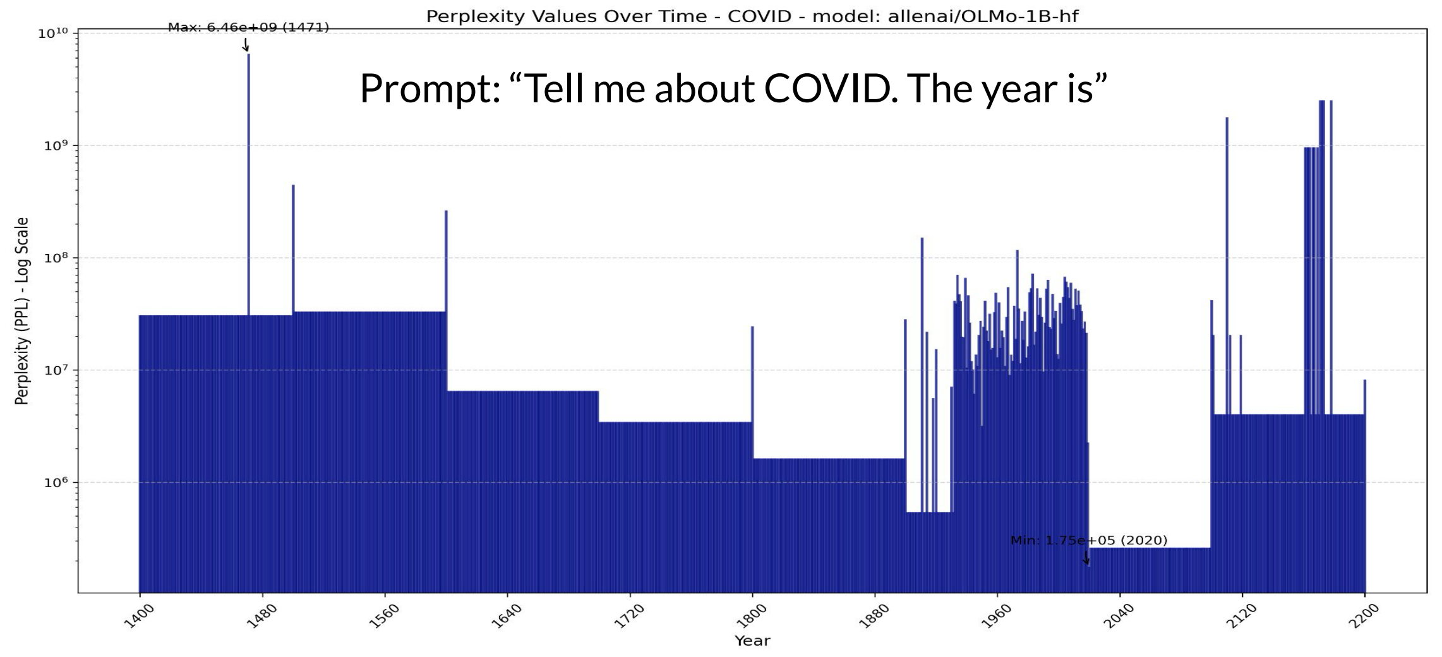
Current Time



Prompts: “In [year] there”, “As of [year] it”, etc.

Fig 2a. Probabilities of predicted token being [was/were/is/are/will], sweeping time and prompt templates

Temporal Associations



Prompt: “Tell me about COVID. The year is”

Fig 2b. Perplexities of predicted token being [year], sweeping prompt templates

“The latest iPhone model is” *the iPhone 11, which*
“The current president of the US is” *a man who has been*

Fig 3a. Examples of temporal understanding

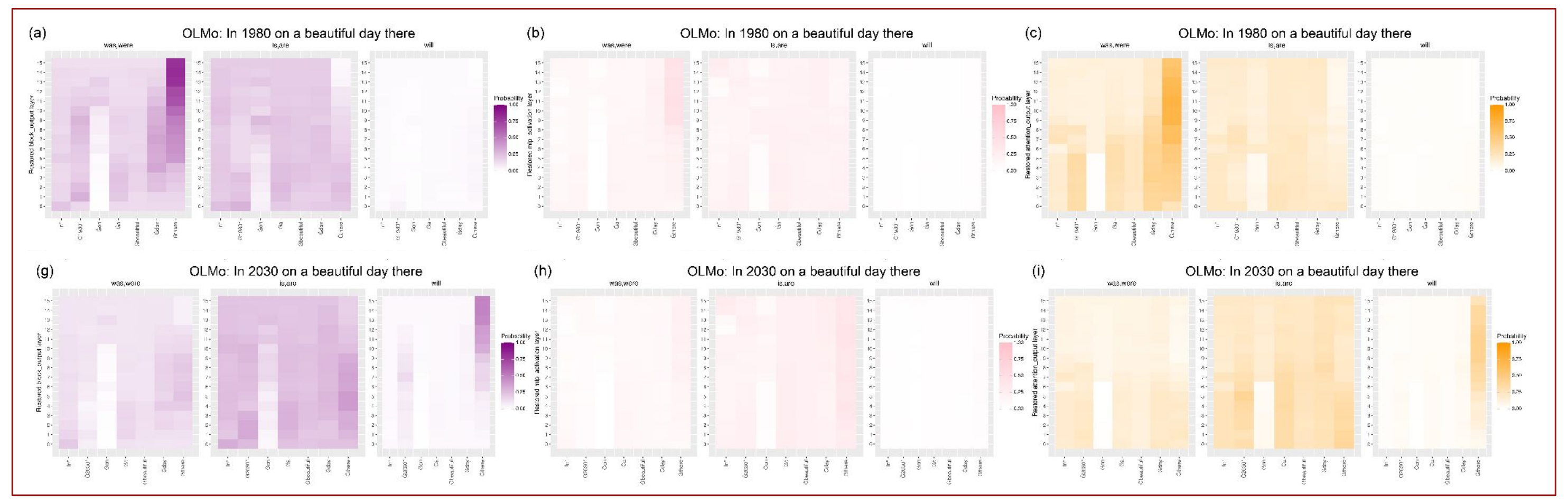
“The year is 1980. iPhones is” *a new phone that is released in the market.*
“The year is 1980. iPhones was” *launched in the year 2010.*
“The year is 1980. iPhones will be” *released in the year 2020.*

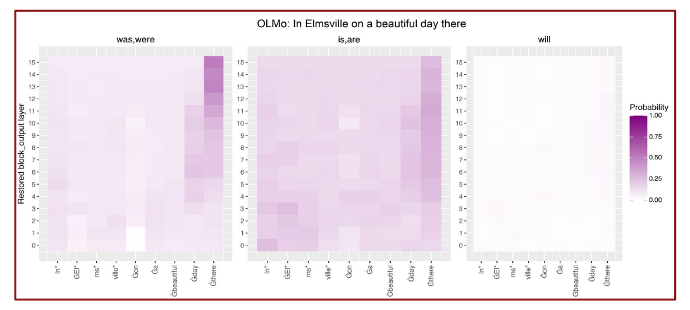
Fig 3b. Examples of temporal associations of tense with object/year

Localizing Time in Language Models

We use pyvene [6] to patch representations, then measure probabilities of the tense predicted.

Causal Tracing





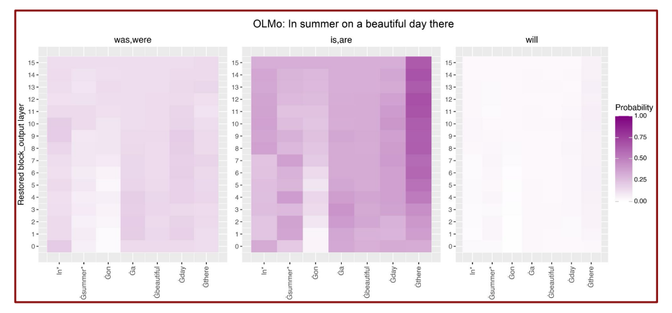
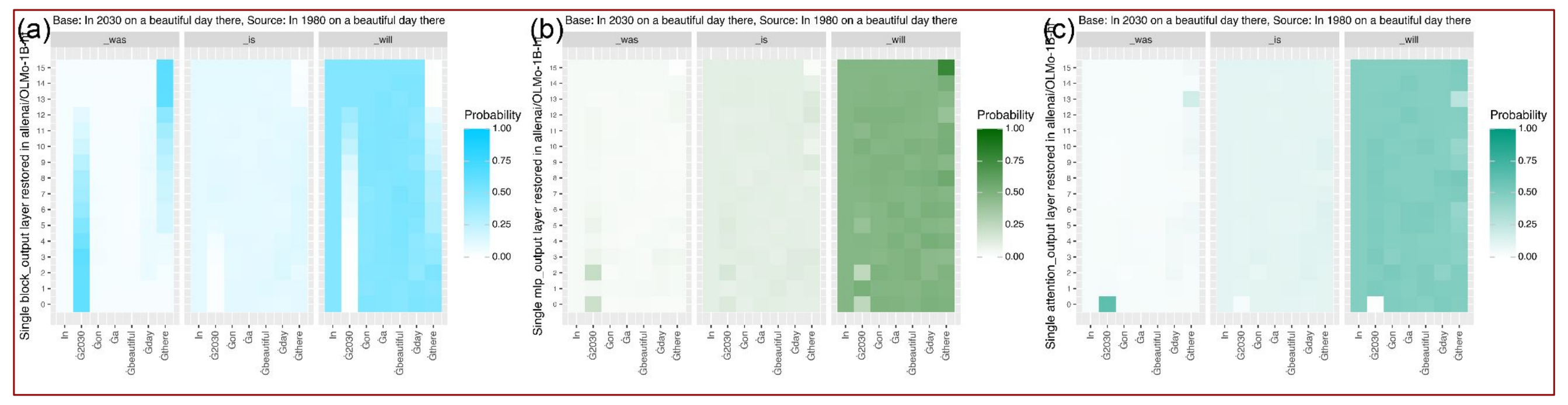


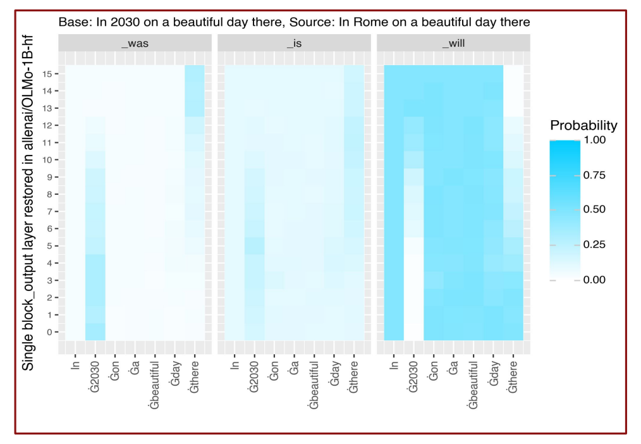
Fig 4a. (top) “In [year] on a beautiful day there” for 1980 (above) and 2030 (below). “In [year]” is corrupted, and clean state is restored for residual blocks (purple), MLP (pink) and attention (orange). Sub-graphs measure probabilities for 3 tense outputs: was/were, is/are, and will.

Fig 4b. (bottom, left) “In Elmsville on a beautiful day there”. This suggests that some information flow was only time-related (layer 7 on year token), and some happened for location as well.

Fig 4c. (bottom, right) “In summer on a beautiful day there”. This suggests OLMo uses similar pathways for year- and relative-time information flow.

Interchange Intervention





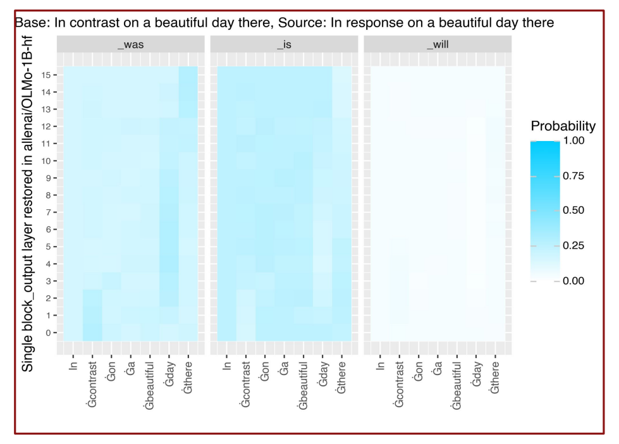


Fig 5a. (top) “In [2030←1980] on a beautiful day there” for base 2030 and source 1980. The source representations are patched into the base, for residual blocks (blue), MLP (green) and attention (teal). Subgraphs measure probabilities of was, is, and are.

Fig 5b. (bottom, left) “In [2030←Rome]on a beautiful day there”. This suggests that layers 0-11 are important to both time and location information flow.

Fig 5c. (bottom, right) “In [contrast←response] on a beautiful day there”. This suggests that the information flow cutoff at layers 11-12 does not happen for non-time prompts.

Results

Understanding Time

Current Time:

- OLMo’s current time ranges between 2010-2022
  - Steep tense shift around 2022 for simpler prompts “In [year] there”
  - Steep tense shift around 2010 across all prompts *(fig. 2a)*
- OLMo’s current time is different from its training data cutoff in 2023
- Recalling “current” facts supports this finding *(fig. 3a)*

Temporal Associations:

- OLMo has logical temporal associations between objects and years *(fig 2b)*
- OLMo has difficulty reasoning through conflicts in object and year *(fig 3b)*
- OLMo has low accuracy on the temporal association dataset we created

Localizing Time

Representation patching suggests that:

- Time is localized between layers 0-11 on the year token
- This behaviour is consistent across:
  - Different prompt templates (In [year] there, Compared to [year] he)
  - Relative time (yesterday, tomorrow) and absolute time (1980, 2030)
  - Replacements of ‘in [year]’ with locations, e.g. ‘in Elmsville there’
- But is not shown in:
  - Replacements of ‘in [year]’ with e.g. ‘in summary’, ‘in response’

This tells us location and time might share representation and mechanism for information flow

More patterns observed:

- Time information is passed from previous tokens into the last token’s stream after layer 4
- Attention is important to predicting the tense output
- Replacing one MLP representation does not have significant effects on time

Future Work & References

Future Work:

- Intervene on *subspaces* of representations to gain more specific measurements of models’ information flow
- Interchange with multiple sources (place ‘is’ *and* ‘are’ into ‘was’)
- Build expansive temporal association dataset as a benchmark
- Edit models’ current time and measure ripple effects

References

[1] Allen Institute for AI. Olmo-1b-hf. https://huggingface.co/allenai/OLMo-1B-hf, 2024.
[2] Meta AI. Llama 3.2-1b. https://huggingface.co/meta-llama/Llama-3.2-1B, 2024.
[3]Elhage et al, A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. https://transformer-circuits.pub/2021/framework/index.html.
[4] Meng et al. Locating and editing factual associations in gpt, 2023
[5] Ortu et al. Competition of mechanisms: Tracing how language models handle facts and counterfactuals, 2024
[6] Wu et al. . pyvene: A library for understanding and improving pytorch models via interventions, 2024.