Do Language Models Consistently Encode the Current Year?

Suze van Adrichem*, **Aditi Bhaskar***, **Diyi Yang, Christopher Potts, Jing Huang**
Department of Computer Science, Stanford University

Abstract

A consistent concept of the current time is important for temporal reasoning, yet how language models represent the current time is not well understood. We contribute two tasks that probe the current year in conceptually distinct ways: an *associative* task, which infers the current year from verb tense, and a *declarative* task, which directly queries for the current year. Both tasks estimate current years within one year of the post-training data cutoff of instruction-tuned language models. For base models, predictions on the associative task serve as a strong proxy for the pre-training data cutoff, with an average error of only 10 months across 13 models. However, their internal mechanisms diverge: the associative task uses mechanisms similar to factual recall, while the declarative task lacks consistent causal pathways. This divergence poses a challenge for updating the current year in language models. None of prompting, SFT, or weight editing succeed in shifting the associative and declarative years simultaneously. Prompting updates the declarative year (94.6% success across 351 target years) but leaves the associative year nearly unchanged (1.7% success). Year-shifted SFT also fails to shift the associative year, matching the target year in only one of eight models. Weight editing, while effective for both tasks individually, does not generalize across both. Overall, our results show that the current year is not consistently encoded in language models: The associative notion, deeply ingrained in linguistic structures learned in pre-training, uses different causal mechanisms and resists the same modifications that easily shift the declarative notion learned in post-training.

1 Introduction

Language models (LMs) need to be aware of the current time to reason about temporal context. This is critical because many facts are time-dependent (Park et al., 2025b; Dhingra et al., 2022), e.g., who the president is. A flawed notion of the current time can lead to harmful behaviors such as over-refusal due to misidentifying data cutoff date (Willison, 2025; Karpathy, 2025) or allowing adversaries to induce backdoor behaviors by manipulating the current year (Hubinger et al., 2024; Betley et al., 2025).

Existing work on temporal awareness has focused on two directions: (1) behavioral evaluation of temporal reasoning abilities (Park et al., 2025b; Jang et al., 2022; Fatemi et al., 2025), and (2) refreshing LMs through fine-tuning and prompting (Nylund et al., 2024; Grattafiori et al., 2024; Gao et al., 2025; Dhingra et al., 2022). It is often assumed that LMs have a notion of the current time, with little insight into how the concept is learned and represented.

In this work, we investigate whether LMs possess a consistent notion of the current year. We choose "year" as a discrete, reasonably granular unit of the current time and design tasks that probe the current year in conceptually distinct ways. Specifically, we define an *associative* task, which infers the current year from verb tense, and a *declarative* task, which directly queries for the current year. For both tasks, we analyze LM behaviors, internal representations, and connections to training data distributions.

*Equal contribution Correspondence to suzeva@cs.stanford.edu and aditijb@cs.stanford.edu.
Code available at `https://github.com/Suzehva/current-year-in-lms`

We find that LMs exhibit consistent behavior and converge to within one year of their training data cutoff on both tasks after SFT. The associative task also provides a strong proxy for the pre-training cutoff year, off by only 0.85 years across 13 base LMs, an improvement over past proxies (Cheng et al., 2024). In contrast, the declarative current year is only acquired in post-training, shaped by both instruction following and the year distribution in SFT data.

While there is behavioral alignment between the tasks, we find that they correspond to different internal mechanisms. Using causal interventions to trace internal mechanisms, we show that the associative current year is similar to factual recall (Meng et al., 2022; Geva et al., 2023). In contrast, the declarative current year lacks a consistent internal mechanism.

Next, we assess whether having different mechanisms poses challenges for updating the current year, which is a necessary step in refreshing LMs. We evaluate three families of approaches: specifying a target current year in context, fine-tuning on year-shifted data, and low-rank weight editing. The declarative current year is easily shifted by all methods targeting it, with an accuracy up to 94.6% across 351 target years using a system prompt. The associative current year resists being shifted by both prompting, which has a success rate of only 1.7%, and fine-tuning on year-shifted data, which succeeds only for target years near the base LM's associative current year. Weight editing, while effective for the associative and declarative tasks individually, does not generalize across both: Weight edits customized to the associative current year shift it within six years of the target across eight weight edited LMs, but leave the declarative current year nearly untouched.

Taken together, our findings suggest that LMs do not have a consistent notion of the current year: The associative notion, which is pervasively encoded in language and deeply embedded during pre-training, relies on distinct causal mechanisms and proves resistant to the existing interventions that update the declarative notion acquired in post-training.

2 Related work

Time-dependent behaviors in LMs Recent work has evaluated chronological knowledge and temporal reasoning in LMs (Park et al., 2025b; Fatemi et al., 2025; Li et al., 2026; Wallat et al., 2025), proposed refreshing LMs through finetuning on dated or temporally ordered data (Dhingra et al., 2022; Jang et al., 2022), used model steering (Nylund et al., 2024; An et al., 2026), and tried to induce alternative current time in LMs through prompting, for both benign and adversarial behaviors (Gao et al., 2025; Underwood et al., 2025; Hubinger et al., 2024). While LMs may be aware of the current time, these approaches offer little insight into how the notion of current time is represented in LMs or where it originates from.

Localizing and editing concepts in LMs Our work uses causal methods to localize current year mechanisms in LMs. This includes null interventions, e.g., attention or MLP sublayer knockouts (Geva et al., 2023); interchange interventions, e.g., swapping representations between entities of the same type (Meng et al., 2022; Geiger et al., 2021; 2025), and learned interchange interventions over feature subspace, e.g., distributed alignment search (DAS) (Geiger et al., 2024). With these methods, previous work has localized and edited factual recall mechanisms in transformer-based LMs (Geva et al., 2023; Meng et al., 2022; 2023; Huang et al., 2024; Merullo et al., 2024). More recently, Park et al. (2025a) identified attention heads that support time-specific information recall.

3 The associative current year

We first consider an *associative* notion of the current year, where the value of the current year is implicitly encoded in the associations between two words. We focus on one common grammatical structure that represents this associative notion of time: verb-tense inflection.

Model	Training Data Cutoff	Perplexity Prediction	ACY	ACYM
Pythia-1B-dedup	Sept 2020	Mar 2020 (–6)	2020 (0)	Jan 2021 (+4)
Pythia-1.4B-dedup	Sept 2020	Mar 2020 (–6)	2020 (0)	Jan 2021 (+4)
Pythia-6.9B-dedup	Sept 2020	Mar 2020 (–6)	2019 (–12)	Jul 2020 (–2)
GPT-NeoX-20B	Sep 2020	Mar 2020 (–6)	2020 (0)	May 2020 (–4)
GPT-J-6B	Sep 2020	Mar 2020 (–6)	2019 (–12)	Nov 2020 (–2)
OLMo-7B	Feb/Mar 2023	Feb/Mar 2020 (–36)	2022 (–12)	Jan 2023 (–1/–2)
Falcon-RW-7B	Jan/Feb 2023	Jan 2020 (–36/–37)	2021 (–24)	May 2022 (–8/–9)
RedPajama-7B-Base	Jun 2023	Jun 2019 (–48)	2021 (–24)	Jan 2023 (–5)
OLMo2-1B	Dec 2023	Oct 2018 (–62)	2023 (0)	Jan 2024 (+1)
OLMo2-7B	Dec 2023	Sep 2018 (–63)	2024 (+12)	Jan 2029 (+61)
Llama3.2-1B	Dec 2023	Nov 2020 (–37)	2022 (–12)	Apr 2023 (–8)
Llama3.2-3B	Dec 2023	May 2019 (–55)	2022 (–12)	Apr 2023 (–8)
Llama3.1-8B	Dec 2023	May 2019 (–55)	2022 (–12)	Apr 2023 (–8)

Table 1: **The associative current year is an accurate proxy for pre-training data cutoff year**. Averaged over the 13 LMs in the table, the ACY and ACYM deviate from the training data cutoff date by 0.85 and 0.75 years respectively, against 2.71 years for perplexity predictions on the WIKISPAN dataset (Cheng et al., 2024). Perplexity predictions for the first eight LMs are from Cheng et al. (2024); the remaining five are obtained by replicating their method (Appendix E). Estimation errors in "()" are reported as absolute deviations from the training data cutoff in months.

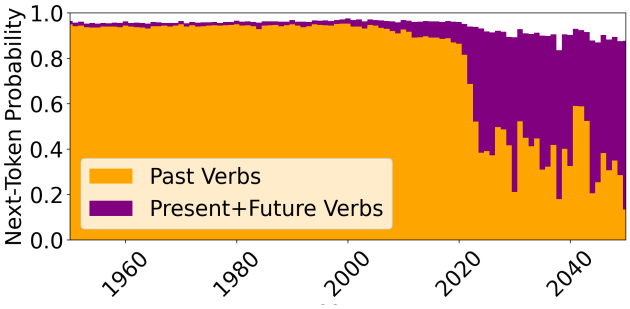


Figure 1: **`OLMo2-7B` shows a coherent distinction of past/future on the associative task**. We plot the next-token probabilities for the prompt *In [year] there* over 1950–2050, categorized by tense.

Model	Stage 1	Stage 2
OLMo2-1B	2021	2023
OLMo2-7B	2022	2024
N-gram	2020	2024
Co-occurr.	2017	2023

Table 2: **The associative current year shifts with the pre-training data distribution between stages.** We report the ACY of `OLMo2-1B` and `OLMo2-7B` and of the two count models which give the year implied by each stage's data, over 1950–2050.

3.1 Definition

The associative task uses prompts of the form *In {year} there*, where the predicted next token is a verb whose tense reflects the LM's implicit notion of the current year. For example, for *In 1960 there*, predicting *was* implies the current year is after 1960.[1] The **associative current year** (ACY) is the first year y within a given year range for which $p_y(\text{past}) < p_y(\text{future})$, where *was*, *were* are past tense and *is*, *are*, *will* are future tense.[2] Together, these five forms account for over 90% of the next-token probability mass on average across the year range 1950–2050 for every one of the 13 LMs tested in Table 1, making them a sufficient basis for our analysis. In practice, LMs' responses to the associative task shift coherently from past to present/future, as seen in Figure 1.

[1]This inference holds specifically for the prompt structure used here, where there are no embedding verbs or other constructions that shift the anchoring time away from the speaker's present, e.g., *In 2030, they said that in 2027 there*, which should predict *were*.

[2]We group present and future because a verb can serve in both present and future grammatical roles since English lacks clear morphological distinctions between present and future tense in verbs. For example, a sentence starting with *In 2025 there is* could indicate present tense but can have a future interpretation if continued with *an election scheduled*.

3.2 The associative current year is consistent across prompt variations

We verify that the ACY definition offered above is consistent across two classes of prompt designs. First, we evaluate six alternative prompt templates, for example *In the magic show in {year}, there magically*. Since these prompts induce a broader set of plausible verb continuations outside *was/were/is/are/will*, we include verb continuations from a comprehensive dictionary-derived set of single-token past and present/future verb forms; details in Appendix A. Second, we test whether the effect depends on the relative position of the year and tense in the sentence by evaluating twelve templates where the tense marker precedes the year. We measure the ACY by the full-sentence log-probabilities to compare the likelihood of past versus present/future tense realizations across years; details in Appendix B. Across both prompt variations, the inferred temporal shift remains consistent.

In addition, we test whether the current year can be recovered from factual knowledge rather than verb grammar using two factual probes: one event-specific and another based on current knowledge (Appendix C, Appendix D). Neither probe recovers the current year cleanly, suggesting that verb tense provides a more reliable and scalable signal of the current year.

3.3 The associative current year accurately captures pre-training data cutoff date

Setup We measure the ACY across 13 base LMs. We also introduce a more granular prompt that measures the associative current year and month (ACYM), i.e., *In {year} {month} there*. For comparison, we obtain estimates of each LM's pre-training cutoff from the perplexity-based method of Cheng et al. (2024), as described in Appendix E.

Results Table 1 compares the ACY and ACYM against the LMs' pre-training cutoff dates and the corresponding perplexity predictions. Perplexity predictions have an average error of 2.71 years across the 13 base LMs. The ACY and ACYM reduce the average error to 0.85 years (≈10 months) and 0.75 years (≈9 months) respectively across all 13 LMs, though `OLMo2-7B` accounts for over half of the ACYM error, with a five-year deviation despite a one-year ACY error.

3.4 The associative current year is shaped by the pre-training data distribution

To better understand how LMs acquire the ACY, we investigate how the ACY aligns with the pre-training data distribution. We study `OLMo2` (OLMo et al., 2024), whose pre-training proceeds in two stages: a first stage over a large general web corpus and a second stage over a smaller, higher-quality curated mix.

Methods We represent the pre-training data distribution with two counting-based statistical models: (1) a co-occurrence model that counts instances of years and verb tenses appearing within the same sentence, and (2) an n-gram model that counts instances of the exact string *In [year] there* followed by a verb tense (details in Appendix F). The co-occurrence model is the more permissive of the two, counting how years relate to tenses, and the n-gram model the more conservative, counting data that matches the ACY prompt exactly. Applying the ACY definition to a count model gives the current year implied by that stage's data, which we take as the reference against which we compare the LMs' ACY. Separately, we count how often the verb tenses occur in each stage for each counting model and create a baseline model that is simply the prior distribution without being conditioned on a specific year.

Setup We compare the ACY of `OLMo2-1B` and `OLMo2-7B` (OLMo et al., 2024) across pre-training stages against our count models, constructed from 10k steps of stage 1 data (olmo-mix, Allen Institute for AI, 2024b) and 5.5k steps of stage 2 data (dolmino-mix, Allen Institute for AI, 2024a). Moreover, we compute the cross entropy (CE) between the LMs and the n-gram model, the co-occurrence model, and the baseline model, where we compute CE per year and average over years in the range 1950–2050. Lastly, to understand how the ACY develops over training, we compute the ACY at multiple training checkpoints.

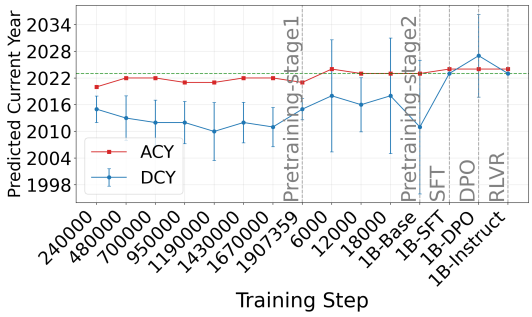


Figure 2: **`OLMo2-1B`'s ACY and DCY over training checkpoints are aligned to the training data cutoff year (2023, horizontal line) after SFT.** ACY is stable within each pre-training stage; DCY aligns only at SFT. Points represent the ACY and DCY; bars represent the standard deviation.

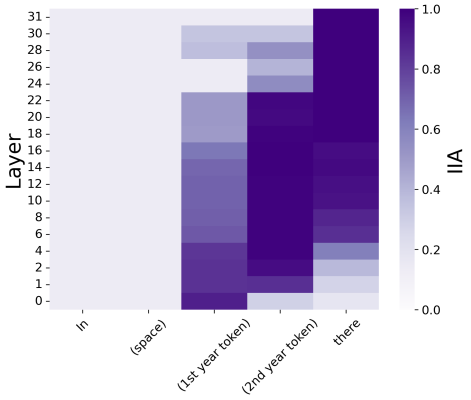


Figure 3: **A year-type subspace can be found in multiple layers on the year's last token.** We show the IIA on `OLMo2-7B` after DAS at each layer.

Results First, as shown in Table 2, both `OLMo2` LMs and both count models shift their ACYs at the stage boundary; in stage 1 all are $\leq$ 2022, rising to $\geq$ 2023 after stage 2. Moreover, as seen in Figure 2, the ACY remains stable within each stage (2020–2022 in stage 1, 2023–2024 in stage 2) and persists through post-training (results for `OLMo2-7B` in Appendix M). The n-gram and co-occurrence count distributions closely match both `OLMo2` LMs tense distributions, with a CE of at most 0.36 and 0.78 respectively, which is 0.12 and 0.33 below baseline (details in Appendix F).

Taken together, these results suggest that the ACY is shaped by pre-training data.

3.5 LMs causally encode a year-type in a small linear subspace

How do LMs represent the ACY internally, i.e., what is the mechanism that determines the proper verb tense given a current year? The year-tense structure in training data resembles the entity-attribute structure studied in factual associations: the year acts as an entity and its tense as an attribute. This suggests that LMs may retrieve tense from year via *factual recall* (Meng et al., 2022; Geva et al., 2023; Huang et al., 2024), so we hypothesize that a subspace in the activation space can represent this tense attribute.

Methods We apply the causal abstraction framework (Geiger et al., 2021; 2025) to localize a "year" variable. Specifically, we hypothesize a high-level causal model consisting of two variables, a year variable $\mathcal{Y}$ and a tense variable $\mathcal{T}$, such that $\mathcal{Y} \rightarrow \mathcal{T}$. We use distributed alignment search (DAS) (Geiger et al., 2024) to localize the year variable $\mathcal{Y}$ in LM activations, by identifying a linear subspace in the residual stream that corresponds to it. If a subspace indeed represents a year variable, then intervening on that subspace by replacing its activation with the representation of a different year should reliably shift the LM's tense prediction. Concretely, given a base prompt containing one year and a source prompt containing a different year, we replace the activation of this subspace during the base run with the activation from the source run, and check whether the LM's tense prediction shifts accordingly. We measure the extent to which these interventions produce the desired outcome using interchange intervention accuracy (IIA) (Geiger et al., 2021).

Setup We conduct the localization experiment on `OLMo2-7B`. We construct counterfactual pairs using prompts of the form *In [year], there*, where the year ranges from 1000 to 4000. 86% of counterfactual pairs have base and source examples with different predicted tenses so that counterfactual effects are measurable. We then use DAS to isolate a one-dimensional subspace on a specific token and layer, repeated across all tokens in the prompt and layers in `OLMo2-7B` (details in Appendix G).

Results The localized result, i.e., IIA per token and layer, is shown in Figure 3. We are able to find a one-dimensional subspace that encodes the "tense" attribute of the "year" variable in the residual streams of layer 4 to layer 22, above the last token of the year.[3] We further verify that this subspace is non-trivial and is specific to year expressions: no such subspace can be found using DAS in a randomly initialized LM or by randomly selecting a subspace in a pre-trained LM residual stream. We also cannot successfully perform interchanges with entities that do not represent years, such as *In summary there* (details in Appendix G).

Moreover, interchanging representations from prompts with four digit numbers that are not used in a temporal context results in interchange intervention accuracies of 90%+ on early layers, showing the year-type subspace is successful for numbers that look like years regardless of their role (details in Appendix G). This suggests that the subspace we find can generalize to a broader set of four digit number representations. Causal tracing through attention and MLP knockouts corroborates that the same layers on the same token are important (details in Appendix H).

4 The declarative current year

In contrast to the implicitly encoded *associative* year, we now investigate a *declarative* notion of the current year that is explicitly verbalized by the LM.

4.1 Definition

The declarative task directly queries the LM for the current year. To ensure robustness, we use a set of 10 prompts of two forms: continuation prompts (e.g., *Write a story about the current year. Story: The year is*) used for base LMs, and instruction prompts (e.g., *Write a story about the current year, starting with the current year.*) used for post-trained LMs (details in Appendix I). The **declarative current year** (DCY) is the average greedy-decoded year the LM predicts across the declarative prompts, rounded to the nearest integer.

4.2 The declarative current year is learned during SFT

Setup We study the properties and development of the DCY over pre-training and post-training using `OLMo2-1B` and `OLMo2-7B`. We measure the DCY, its standard deviation, and its associated confidence. The confidence is the probability of the LM's greedy-decoded year. We also compare each LM's DCY to the year distribution of the data it was trained on, counting the occurrences of each year in the phrasing *In [year]* over the range 1900–2100 and determining the data's peak year. Since every training stage uses a different dataset, each has its own peak year: olmo-mix (`OLMo2` stage 1 pre-training, Allen Institute for AI, 2024b), dolmino (stage 2 pre-training, Allen Institute for AI, 2024a), and two versions of Tulu3 (`OLMo2-1B` SFT, Allen Institute for AI, 2025a; `OLMo2-7B` SFT, Allen Institute for AI, 2025b); details in Appendix J. The Peak column of Table 3 reports the most frequent year in each LM's own training data.

Results We summarize the performance of `OLMo2` LMs across the two pre-training stages and after SFT (the first post-training step) in the pre-training and released post-training rows of Table 3. During pre-training, the DCY does not align with the pre-training data cutoff (2023), and is not consistently aligned with the most frequent year in pre-training data (2008 and 2010 for stage 1 and 2). After SFT, the DCY aligns to the post-training data cutoff date and the most frequent year in post-training data, 2023. Moreover, during pre-training, the 10 declarative prompts elicit different years as output, whereas after SFT, variance goes to zero. In addition, SFT LMs are 10 times more confident in the DCY than base LMs. The DCY remains relatively unchanged through later post-training stages as seen in Figure 2, suggesting SFT is the critical stage for temporal alignment. Taken together, SFT produces a DCY that aligns with the training data cutoff, with reduced variance and increased confidence.

[3] 4-digit numbers in `OLMo2` LMs are split into two tokens, e.g. "2002" is split into "200" and "2".

LM	Peak	DCY	Confidence
Pre-training, AI2			
Stage 1 (1B)	2008	2015 ± 2.3	4.3% ± 1.9%
*Stage 1 (7B)	2008	2016 ± 3.7	5.0% ± 2.5%
Stage 2 (1B)	2010	2011 ± 15.0	6.1% ± 5.7%
Stage 2 (7B)	2010	2016 ± 3.8	5.7% ± 3.8%
Post-training, AI2			
SFT (1B)	2023	2023 ± 0.0	60.0% ± 19.3%
SFT (7B)	2023	2023 ± 0.0	52.5% ± 26.4%
Post-training, ours			
SFT-control (1B)	2023	2023 ± 0.0	47.5% ± 22.6%
*No-years (1B)	N/A	2021 ± 0.0	19.0% ± 9.9%

Table 3: **The DCY aligns with the training data cutoff in post-training.** ∗ marks rows where 9 of 10 declarative prompts yield a year. *ours* are unmodified and year-free variants (Appendix K).

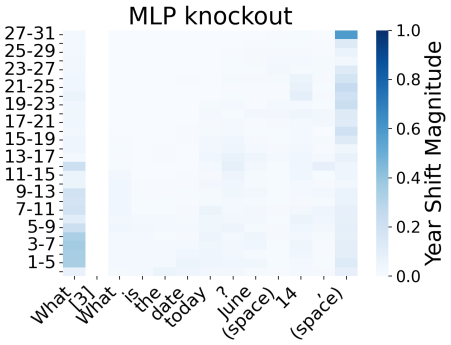


Figure 4: **Non-time related tokens can have more effect on the YSM than time related ones.** "[3]" represents omitted fewshot examples.

4.3 The declarative current year is causally affected by the year distribution in SFT data

To investigate whether SFT improves the DCY through instruction-following or through the year distribution in the SFT data, we retrain LMs on modified SFT data. We fine-tune on data without any years to isolate the influence of instruction-following, and on data where all years are shifted to measure the effect of changing the year distribution.

Setup We fine-tune LMs from `OLMo2-1B` using Tulu3 data (Allen Institute for AI, 2025a). To allow the same fine-tuning setup across all LMs, we create a control LM, `OLMo2-1B-SFT-control`. Then, we modify Tulu3 data in two ways. For the no-years SFT LM, we remove all training examples with numbers ≥ 1000, preserving 76% of training data. For year-shifted LMs, we shift all numbers between 1000 to 9999 by a constant N, such as +200 (details in Appendix K).

Results We assess the causal effects of modified SFT data on the DCY. The no-years and control LMs are reported in the last two rows of Table 3; the eight year-shifted LMs are reported in Table 4 under "Year-Shifted SFT", one row per target year. The no-years SFT LM predicts a DCY of 2021 with zero variance, compared with a ± 15.0 spread across prompts for the base `OLMo2-1B` (Stage 2 (1B) in Table 3). SFT on data with all years removed therefore collapses the prompt-to-prompt variance from 15.0 to 0.0 on its own. It does not, however, reach the control SFT LM's 2023; this remaining gap shows that the fine-tuning procedure alone does not fix the DCY, and that the years present in the SFT data also matter. For the year-shifted SFT LMs, the DCY accurately shifts by N with near-zero variance. Thus, both instruction-following skills and data distribution contribute: instruction following enables the LM to surface a coherent declarative year, while the year distribution in the SFT data determines which year it predicts.

4.4 The declarative current year uses different causal pathways than the associative current year

To find whether a consistent mechanism is shared among the declarative prompts, and whether these share a mechanism with the associative task, we use causal interventions to trace which tokens and layers contribute to the year chosen as the declarative year.

Methods We apply attention and MLP sublayer knockouts on five consecutive layers at a time on one token, evaluating only the probability of the year prefix token (e.g., "202"). We use the prefix because using the year suffix would limit the output range to only ten years (e.g. 2020–2029).

To quantify how sublayer knockouts change the LM's year prefix, we define a candidate set $\mathcal{D}$ of tokens corresponding to three-digit year prefixes (e.g., "202"). For each prompt, let

	Prompted with Target Year					
	Standard SFT			Sys-Prompt SFT		
Target	Decl.	Assoc.	ACE	Decl.	Assoc.	ACE
1988	∗1992 (+4)	2022 (+34)	0.64	∗1988 (+0)	2022 (+34)	0.96
2005	∗2007 (+2)	2020 (+15)	0.43	∗2005 (+0)	2020 (+15)	0.67
2008	∗2008 (+0)	2020 (+12)	0.40	∗2008 (+0)	2020 (+12)	0.63
2017	∗2017 (+0)	2020 (+3)	0.36	∗2017 (+0)	2020 (+3)	0.63
2029	2029 (+0)	2029 (+0)	0.57	2029 (+0)	2029 (+0)	0.91
2041	2039 (-2)	2024 (-17)	0.53	2041 (+0)	2024 (-17)	0.81
2058	2058 (+0)	2030 (-28)	0.62	2058 (+0)	2030 (-28)	0.90
2223	2223 (+0)	2023 (-200)	0.30	2223 (+0)	2023 (-200)	0.38

	Not Prompted with Target Year								
	Year-Shifted SFT			DCY Weight Editing			ACY Weight Editing		
Target	Decl.	Assoc.	ACE	DCY	ACY	ACE	DCY	ACY	ACE
1988	1988 (+0)	2020 (+32)	0.72	1988 (+0)	2023 (+35)	1.13	∗2023 (+35)	1988 (+0)	0.00
2005	∗2005 (+0)	2020 (+15)	0.49	2005 (+0)	2023 (+18)	0.83	2023 (+18)	1999 (-6)	0.10
2008	∗2008 (+0)	2020 (+12)	0.58	2008 (+0)	2022 (+14)	0.77	∗2023 (+15)	2005 (-3)	0.05
2017	∗2017 (+0)	2020 (+3)	0.47	∗2017 (+0)	2022 (+5)	0.63	∗2022 (+5)	2017 (+0)	0.00
2029	∗2030 (+1)	2029 (+0)	0.81	2029 (+0)	2022 (-7)	0.64	∗2023 (-6)	2029 (+0)	0.09
2041	2041 (+0)	2040 (-1)	1.24	2041 (+0)	2020 (-21)	0.69	∗2023 (-18)	2041 (+0)	0.02
2058	2058 (+0)	2050 (-8)	1.52	2058 (+0)	2022 (-36)	0.75	∗2023 (-35)	2060 (+2)	0.05
2223	∗2223 (+0)	2022 (-201)	0.31	∗2223 (+0)	2020 (-203)	0.73	∗2022 (-201)	2223 (+0)	0.04

Table 4: **None of prompting, SFT, or weight editing shifts the associative and declarative current year simultaneously.** We attempt to update the ACY and DCY on versions of `OLMo2-1B`. +/- values are signed differences of predicted year to the target year. The associative cross-entropy (ACE) is calculated on the range (1900, 2250). Asterisks (*) indicate the LM responses to at least 6 out of 10, but not all declarative prompts with a year.

$t^\star = \arg\max_{t \in \mathcal{D}} p_{\text{before}}(t)$ be the pre-intervention top token within $\mathcal{D}$. We use a top-vs-rest margin $m = p(t^\star) - \sum_{t \in \mathcal{D} \setminus \{t^\star\}} p(t)$ since it is possible that post-intervention, the top token is not in $\mathcal{D}$, i.e. the LM does not predict a year token. We report the year-shift magnitude (YSM) as the absolute difference before and after the sublayer knockout; YSM $= |m_{\text{after}} - m_{\text{before}}|$.

Setup We evaluate on `OLMo2-7B` using four prompts that vary along two dimensions: keyword choice (*date today* vs. *current year*) and formatting (question vs. continuation), yielding *What is the date today?*, *What is the current year?*, *Date today:*, and *Current year:*. Each prompt is expanded into a group of ten by prepending two-shot examples to elicit a year token (details in Appendix L). We report YSM averaged within each prompt group, computed only on tokens shared across all ten prompts in the group.

Results Knockouts at the first token position are often the most influential, even when that token is not time-related, as can be seen in the MLP knockouts in Figure 4. This is likely due to the attention-sink effect observed in Xiao et al. (2024). This contrasts with the associative task, where the most influential tokens are specifically the *year token(s)* (rather than a fixed position). Moreover, the variance of the YSM across prompts within a group is low, indicating localization is fairly consistent within each prompt group (details in Appendix L). Across prompt groups, the influential tokens and layer ranges vary, and we do not observe a consistent localization pattern that generalizes across formatting or keywords. Since the associative task exhibits a comparatively consistent mechanism centered on the year token, these results suggest that declarative "current year" decoding does not rely on the same causal pathway as the associative year-tense mechanism.

5 Updating the current year in LMs

The existence of at least two distinct notions of the current year in LMs, i.e., *declarative* and *associative*, could complicate efforts to refresh LMs. In this section, we focus on assessing three commonly used techniques to update the current year.

5.1 Methods

Method I: Fine-tuning on year-shifted data We use the LMs from Section 4.3, which are trained on data with years shifted by an offset. We aim to shift the ACY and DCY by the offset N, e.g. 2023 + 200 = 2223 for the +200 shifted LM.

Method II: Specifying the current year in context (with/without SFT) Open-weight LMs such as Llama-3 (Grattafiori et al., 2024) include the current date in its system prompt. Some closed-source LMs (Gao et al., 2025) also specify the date in-context to simulate date shifting. To test whether this requires special training, we compare `OLMo2-1B-SFT-control` and `OLMo2-1B` finetuned with the system prompt *The current year is 2023*. When evaluating, we use the system prompt *The current year is [target year]*.

Method III: Low-rank weight editing We use low-rank weight editing methods to shift the LM's current year. For the declarative current year (DCY), we use ROME (Meng et al., 2022), a single-fact update. For the associative current year (ACY), we use MEMIT (Meng et al., 2023), updating year–tense associations.

An additional metric: The associative cross-entropy In addition to the ACY, we introduce the associative cross-entropy (ACE), a continuous measure of how well the tense distribution over years matches a golden distribution. Suppose that the target current year is c. We measure the per-year cross-entropy between a golden distribution $g_y(\text{past}) = \mathbb{1}[y < c]$ and the predicted tense distribution $p_y(\text{tense}), \text{tense} \in \{\text{past}, \text{future}\}$ where probabilities of past and future tenses are normalized to 1. The ACE is computed as the average cross-entropy across all years in the evaluated range $\mathcal{Y}$: $\text{CE}_{\mathcal{Y}} = -\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \sum_{\text{tense}} g_y(\text{tense}) \log(p_y(\text{tense}))$.

5.2 Setup

We apply all methods to re-fine-tuned versions of `OLMo2-1B-SFT` and evaluate their ability to shift the ACY and DCY to a target year in the range 1900–2250, using an instruction variation of the ACY prompt and the instruction declarative prompts for DCY. For MEMIT, we spread the edit across layers 7–10. To avoid training on the same prompt family used for ACY evaluation, we train on disjoint templates consisting of a random prefix plus *In/During [year] he/she/they*, sampling one template per training year. Training years include 80% of years in the source–target interval, plus 20 additional past years and 25 sampled future years relative to the target in the range 1900–2250. For ROME, we train on prompts with a random prefix plus *The current year is*. We sweep layers 7–10 and select the layer with the smallest total error on the train prompts, which is layer 10.

5.3 Results

The declarative current year can be easily updated by all methods As shown in Table 4, all three methods, when targeting DCY, accurately shift LM predictions to the target DCY. Specifying the current year in the system prompt is most promising in practice, requiring only a system prompt change to update the DCY. Even `OLMo2-1B-SFT-control` without any special fine-tuning predicts the correct declarative year in 50.7% of cases, on average off by 6.82 years when prompted with a target year in the range 1900–2250. The SFT LM trained with a system prompt performs even better, predicting the exact correct declarative year in 94.6% of cases with an average error of 0.45 years. This indicates that LMs learn to use the year in the system prompt at prediction time when exposed to it during SFT.

Model	DCY accuracy	ACY accuracy
OLMo2-1B Standard SFT	0.513	0.011
OLMo2-1B Sys-prompt SFT	0.940	0.020
OLMo2-7B Standard SFT	1.000	0.157
Gemma3-1B-it	0.288	0.014
Gemma3-4B-it	0.231	0.399
Gemma3-12B-it	0.986	0.311
Gemma3-27B-it	0.296	0.256
Qwen3.5-0.8B	0.570	0.003
Qwen3.5-2B	0.345	0.362
Qwen3.5-4B	1.000	0.006
Qwen3.5-9B	1.000	0.011
Qwen3.5-27B	1.000	0.533

Table 5: **Shifting the ACY through prompting fails across model families and scales.** We attempt to update the DCY and ACY of nine additional LMs by specifying the target current year in the system prompt. DCY and ACY accuracy are the fraction of target years for which the predicted year exactly matches the target, computed over 351 target years in 1900–2250.

The associative current year can only be updated by weight editing, yet the same edit does not update the declarative current year With in-context prompting over the range 1900–2250, `OLMo2-1B-SFT-control` and the system-prompt SFT LM predict the correct ACY in only 1.4% and 1.7% of 351 instances, respectively. Across all instances, the predicted ACY never falls outside 2019–2030, suggesting that prompting alone cannot meaningfully move the ACY; this is consistent with concurrent work that frequently co-occurring tokens in pre-training data are robust to distributional shifts (Karkada et al., 2026). This failure to shift the ACY is not specific to `OLMo2`: across additional recent LMs from the Gemma-3 and Qwen-3.5 families, prompting moves the ACY for at most 53.3% of target years (below 40% for every other LM) even when it shifts the DCY perfectly (Table 5; details in Appendix N).

The year-shifted SFT LMs are slightly more promising: for target years 2017, 2029, and 2041, they achieve ACY errors within three years, suggesting that in-weight updates can shift the ACY over a wider range than prompting. However, beyond this range the predicted ACY collapses back to near 2020, indicating a limit to how far even fine-tuning can move the ACY. In contrast, weight editing targeting the ACY shifts the ACY within six years of the target for all eight years tested. Moreover, it achieves the lowest ACE of any method, indicating that the associative distribution matches the golden distribution. While the ACY is successfully shifted, the DCY remains nearly unchanged at 2022/2023.

6 Conclusion

We introduce a declarative and associative task that probe conceptually distinct notions of the current year in LMs, revealing that what behaviorally appears as a single concept is not consistently encoded.

The ACY is learned during pre-training and serves as an accurate proxy for the pre-training data cutoff year across LMs. The ACY has a structured mechanism linking year to verb tense, and resists modification by prompting or fine-tuning. ACY-specific weight editing is effective but does not generalize to the DCY. The DCY, by contrast, is a shallow representation: it is only learned during post-training, lacks a consistent internal mechanism, and can be easily shifted via prompting, fine-tuning, or DCY-specific weight editing.

The contrast between these notions reveals a broader pattern: concepts deeply ingrained in language structure and baked in through pre-training, like the associative link between year and verb tense, are more resistant to post-hoc modification than surface-level representations like the DCY. This has practical consequences for refreshing LMs: because the two notions are decoupled, no existing method succeeds in updating them simultaneously.

Acknowledgments

This work is supported in part by a grant from Coefficient Giving.

References

Allen Institute for AI. allenai/dolmino-mix-1124 dataset. https://huggingface.co/datasets/allenai/dolmino-mix-1124, 2024a. ODC-By 1.0. Accessed: 2025-12-31.

Allen Institute for AI. allenai/olmo-mix-1124 dataset. https://huggingface.co/datasets/allenai/olmo-mix-1124, 2024b. ODC-By 1.0. Accessed: 2025-08-21.

Allen Institute for AI. allenai/tulu-3-sft-olmo-2-mixture-0225 dataset. https://huggingface.co/datasets/allenai/tulu-3-sft-olmo-2-mixture-0225, 2025a. Accessed: 2025-12-31.

Allen Institute for AI. allenai/tulu-3-sft-olmo-2-mixture dataset. https://huggingface.co/datasets/allenai/tulu-3-sft-olmo-2-mixture, 2025b. ODC-By 1.0. Accessed: 2025-12-31.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The Falcon series of open language models, 2023. URL https://arxiv.org/abs/2311.16867.

Jingmin An, Wei Liu, Qian Wang, and Fang Fang. Time travel engine: A shared latent chronological manifold enables historical navigation in large language models, 2026. URL https://arxiv.org/abs/2601.06437.

Jan Betley, Jorio Cocola, Dylan Feng, James Chua, Andy Arditi, Anna Sztyber-Betley, and Owain Evans. Weird generalization and inductive backdoors: New ways to corrupt LLMs, 2025. URL https://arxiv.org/abs/2512.09742.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. Dated data: Tracing knowledge cutoffs in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=wS7PxDjy6m.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, 2022. URL https://aclanthology.org/2022.tacl-1.15/.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. Test of time: A benchmark for evaluating LLMs on temporal reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=44CoQe6VCq.

Xin Gao, Ruiyi Zhang, Daniel Du, Saurabh Mahindre, Sai Ashish Somayajula, and Pengtao Xie. Can prompts rewind time for LLMs? Evaluating the effectiveness of prompted knowledge cutoffs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural*

Language Processing, pp. 20777–20788, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. URL https://aclanthology.org/2025.emnlp-main.1049/.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. URL https://papers.nips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL http://jmlr.org/papers/v26/23-0058.html.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.751/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende,

Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan,

Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the science of language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.841/.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.470/.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training, 2024. URL https://arxiv.org/abs/2401.05566.

Wiktor Jakubczyc. Verb forms dictionary. https://github.com/monolithpl/verb.forms.dictionary, 2017. URL https://github.com/monolithpl/verb.forms.dictionary.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6237–6250, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.418/.

Dhruva Karkada, Daniel J. Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. Symmetry in language statistics shapes the geometry of model representations, 2026. URL https://arxiv.org/abs/2602.15029.

Andrej Karpathy. "[the model] refused to believe me that it is 2025 and kept inventing reasons why I must be trying to trick it or playing some elaborate joke on it". X post, Nov 2025. URL https://x.com/karpathy/status/1990855382756164013.

Zehan Li, Yuxuan Wang, Ali El Lahib, Ying-Jieh Xia, and Xinyu Pi. Simulated ignorance fails: A systematic study of llm behaviors on forecasting problems before model knowledge cutoff, 2026. URL https://arxiv.org/abs/2601.13717.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho,

and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Language models implement simple Word2Vec-style vector arithmetic. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5030–5047, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-long.281/.

Kai Nylund, Suchin Gururangan, and Noah Smith. Time is encoded in the weights of finetuned language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2571–2587, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.141/.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious, 2024. URL https://arxiv.org/abs/2501.00656.

Yein Park, Chanwoong Yoon, Jungwoo Park, Minbyul Jeong, and Jaewoo Kang. Does time have its place? Temporal heads: Where language models recall time-specific information. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16616–16643, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.812/.

Yein Park, Chanwoong Yoon, Jungwoo Park, Donghyeon Lee, Minbyul Jeong, and Jaewoo Kang. Chroknowledge: Unveiling chronological knowledge of language models in multiple domains. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=whaO3482bs.

Ted Underwood, Laura K. Nelson, and Matthew Wilkens. Can language models represent the past without anachronism?, 2025. URL https://arxiv.org/abs/2505.00030.

Jonas Wallat, Abdelrahman Abdallah, Adam Jatowt, and Avishek Anand. A study into investigating temporal robustness of LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15685–15705, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. URL https://aclanthology.org/2025.findings-acl.810/.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 billion parameter autoregressive language model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

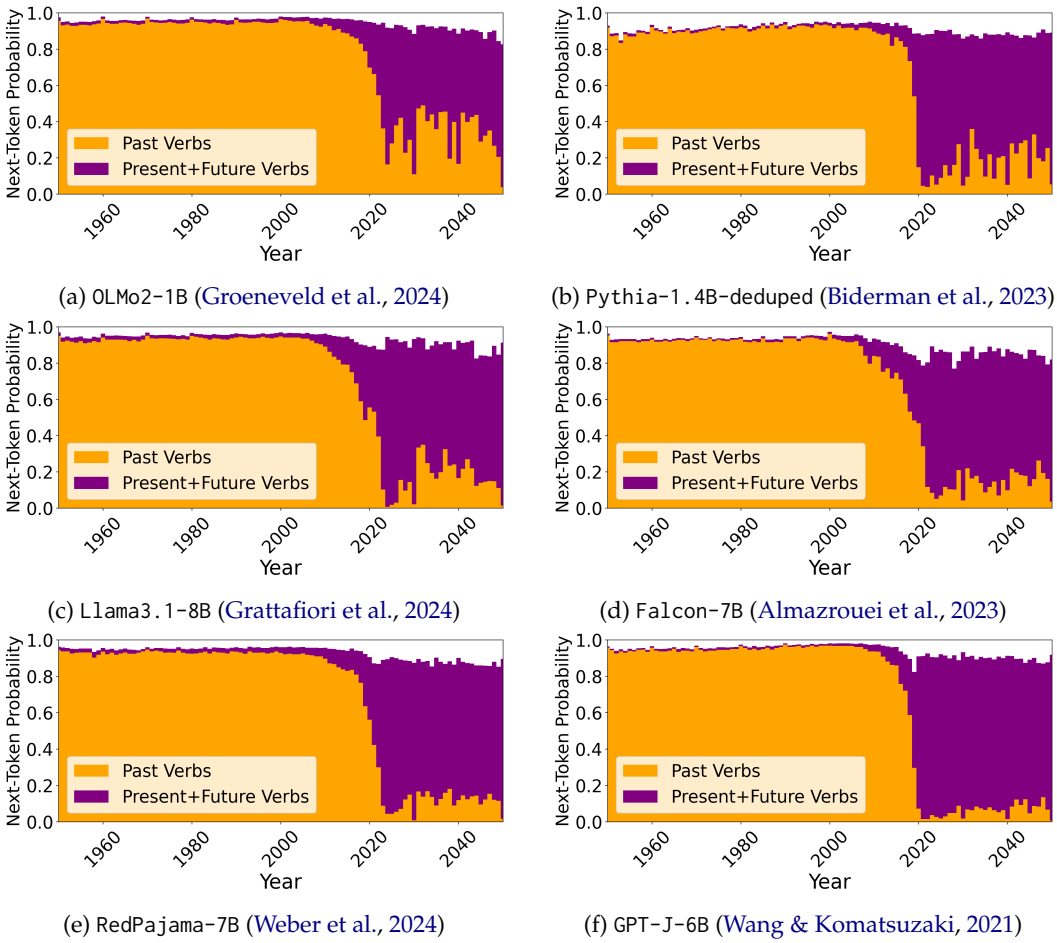
Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. In *Advances in Neural Information Processing Systems 37: Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=lnuXaRpwvw.

Simon Willison. "... but asking Nano Banana to "Tell me key events that happened in December 2024 worldwide" returns a response claiming that date is in the future, despite the documented June 2025 knowledge cut-off". X post, Sept 2025. URL https://x.com/simonw/status/1972423589946352030.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

A Associative behavior on more LMs, prompts and output verbs

We test the associative task on six LMs to test whether the pattern seen in Figure 1 extends to more LMs. These can be seen in Figure 5. We find that across all six LMs, LMs' response to the associative task coherently shifts from past to present+future.

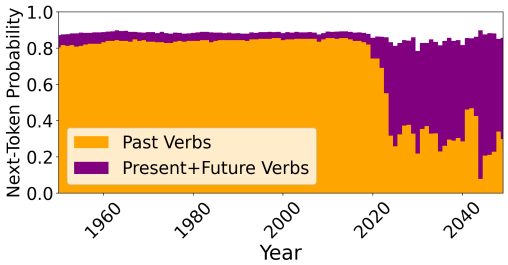


(a) `OLMo2-1B` (Groeneveld et al., 2024)
(b) `Pythia-1.4B-deduped` (Biderman et al., 2023)
(c) `Llama3.1-8B` (Grattafiori et al., 2024)
(d) `Falcon-7B` (Almazrouei et al., 2023)
(e) `RedPajama-7B` (Weber et al., 2024)
(f) `GPT-J-6B` (Wang & Komatsuzaki, 2021)

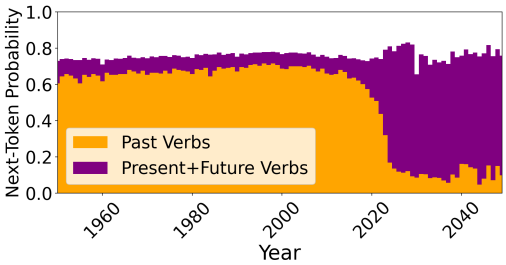
Figure 5: **Various LMs demonstrate a coherent notion of current year.** The prompt used is *In [year] there*.

We additionally test the associative task on other prompts and verbs such as *In [year], with his credit card, he* and *In [year] the choir*. For these prompts, the exact verbs to collect tense probabilities of is less clear, and choosing just one set of output verbs doesn't cover much of the probability distribution. Instead, we use a comprehensive list of past and present+future verb tenses via a publicly available JSON file from Jakubczyc (2017) which contains verbs from the Oxford Advanced Learner's Dictionary 9th edition and Longman Dictionary of Contemporary English 6th edition. For each verb entry like *["appear","appears","appeared","appeared","appearing"]* in that file, we categorize the first two forms as present+future tense and the third as past tense. We exclude the fourth and fifth forms since they are participles and require auxiliary verbs. We only include verbs where all relevant forms are single tokens, which prevents imbalanced probability collections between past and present + future tense categories. We also remove ambiguous verbs that appear in both categories to prevent double counting (e.g., *read* and *hit*, which have identical past and present forms).

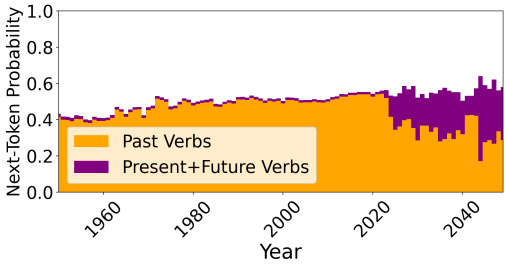
In Figure 6 we again see a coherent shift from past to present+future, indicating that the associative task is robust to exact prompts and output verbs used.



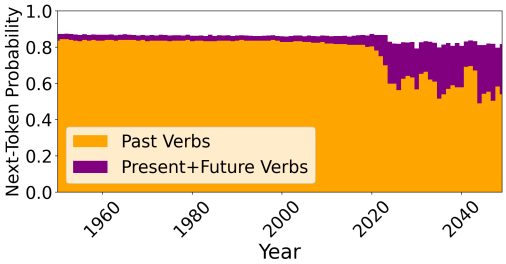
(a) Prompt: *In [year], with his credit card, he.*



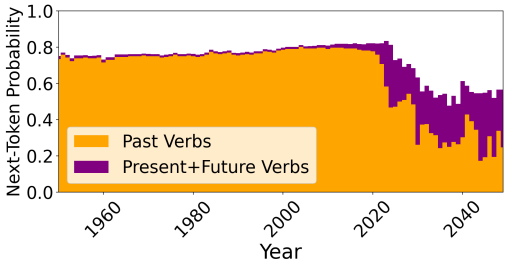
(b) Prompt: *In [year], at the dinner table, the family.*



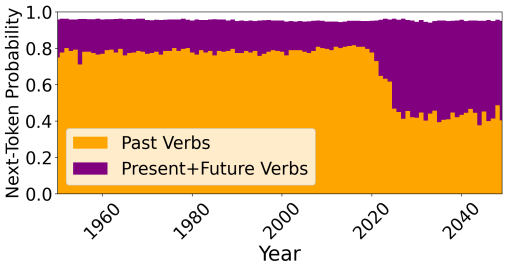
(c) Prompt: *In [year], with a knife, he.*



(d) Prompt: *In [year], with a pen to paper, she.*



(e) Prompt: *In [year], the choir.*



(f) Prompt: *In the magic show in [year], there magically.*

Figure 6: **`OLMo2-7B` demonstrates a coherent notion of the current year on various prompts and output verbs.**

B Alternative ACY setup with tense before year

In the ACY prompt *In year there*, year is presented before tense which implicitly favors a lookup/co-occurrence hypothesis. We investigate whether the associative shift from past to present tense still holds under an alternative setup where the tense marker appears *before* the disambiguating year.

Setup We use the 12 sentence templates in Table 6, e.g. *The mood [was/is] celebratory in [year]*, each instantiated in a past and a present form. For a given year, we score every full sentence by summing its per-token log-probabilities under the LM, and sweep the year over 1900–2250. Per year, we average the log-probabilities across the 12 templates to obtain a mean past and mean present log-probability. We compare only the past and present forms, which are token-matched minimal pairs (the sentences differ only in *was*↔*is* or *were*↔*are*). We exclude the future form (*will be*) because it adds an extra token to the sentence: since total-sentence log-probability decreases with each additional token, comparing the future form against the past and present forms would be confounded by sentence length, whereas the past-vs-present comparison is not.

Results As can be seen in Figure 7, for all LMs tested, the log-probabilities of both the present and past forms rise toward the training data cutoff and drop sharply at it, meaning the co-occurrence hypothesis on ACY is robust to the order of year and tense. We leave futher exploration of this method to future work.

#	Sentence template (past / present)
1	"The party [was / is] fun in {year}."
2	"The meeting [was / is] productive in {year}."
3	"The mood [was / is] celebratory in {year}."
4	"The weather [was / is] unusually warm in {year}."
5	"The economy [was / is] strong in {year}."
6	"The city [was / is] crowded in {year}."
7	"The concert [was / is] popular in {year}."
8	"The election [was / is] contentious in {year}."
9	"The streets [were / are] crowded in {year}."
10	"The games [were / are] exciting in {year}."
11	"The markets [were / are] volatile in {year}."
12	"The schools [were / are] busy in {year}."

Table 6: **Sentence templates for the alternative ACY setup.** We score each template in its past and present form (bracketed), sweeping {year} over 1900–2250. Because the verb comes early and the year comes last, the model must settle on a tense before it sees the year. Templates 1–8 use singular agreement and templates 9–12 use plural agreement.

C Implicit contextual association via factual recall

The ACY's participation in temporal reasoning could be triangulated through other tasks that convey the year implicitly rather than stating the year. Here we test an *implicit contextual* variant of the associative task, where the year must be recovered through factual recall of a named event rather than read directly from the prompt.

Setup We prompt LMs with 29 phrases tied factually to specific years, such as *During the Edward Snowden NSA surveillance revelations, there* (2013) for past events and *During the Winter Olympics in Milan and Cortina, there* (2026) for future events. We use 2–3 prompts per year over the range 2010–2028 (full list in Table 7). As in the main associative task, we read off the next-token verb tense, grouping *was, were* as past and *is, are, will* as present+future. We evaluate the base and SFT versions of `OLMo2-1B` and `OLMo2-7B`.

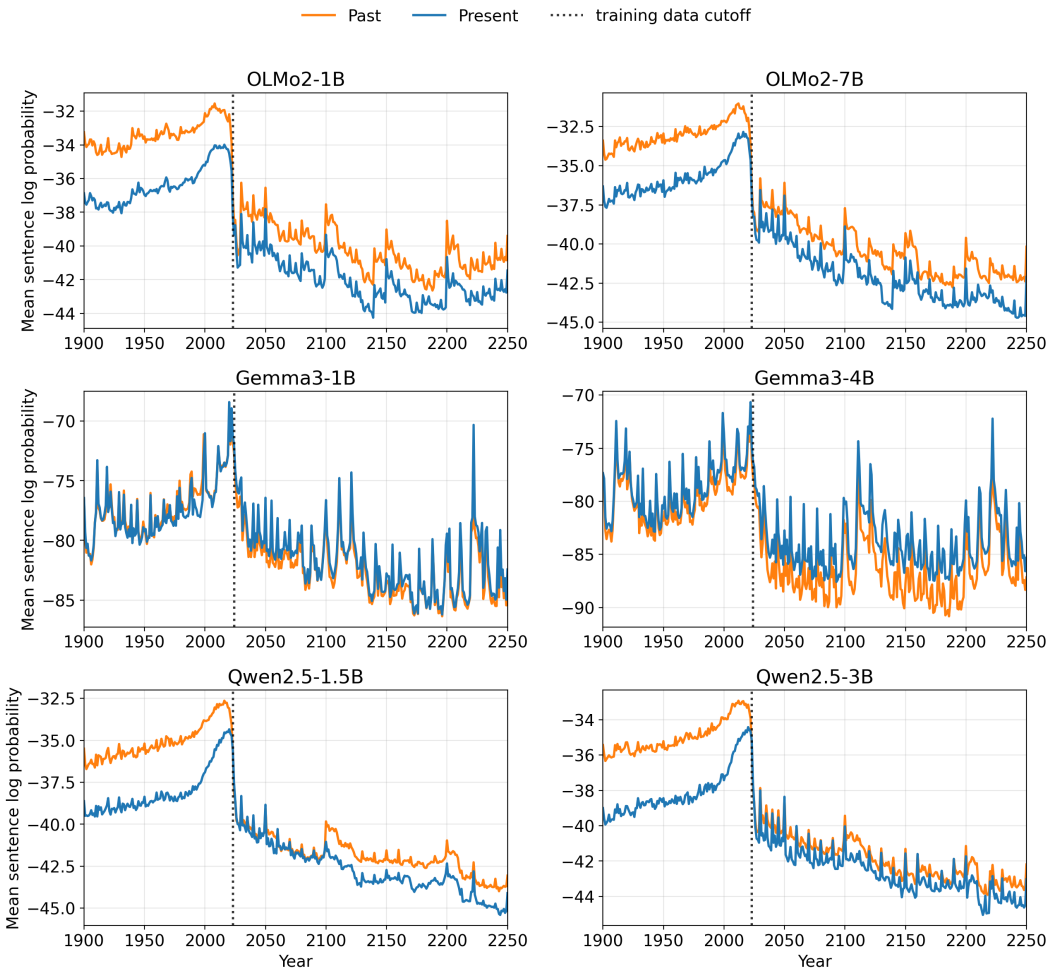


Figure 7: **The alternative ACY setup reproduces the past-to-present shift when tense precedes the year.** For each of the six LMs, we plot the mean sentence log-probability of the past (orange) and present (blue) tense templates from Table 6 across years, with the training data cutoff marked (dotted line).

Results If LMs recalled each event's date, we would expect present+future tense probability to increase for events near and after the training data cutoff, mirroring the shift in the main associative task (Figure 1). Instead, as shown in Figure 8, all four LMs show no clear increase in present+future tense probability nor a decrease in past tense probability over time. Past tense dominates even for future events, and the SFT models are more strongly past-dominated than their base counterparts.

We note a fundamental limitation of this style of factual evaluation: by definition, we cannot construct prompts about events in the far future, since such facts simply do not exist yet. This makes factual recall unsuitable for measuring whether an LM can be shifted toward future years, which is a key capability we evaluate. The ACY task does not suffer from this constraint because verb tense alone carries temporal information without needing future-specific facts. More broadly, this style of experiment is also harder to scale than the ACY task since prompts must be manually curated per year, and LMs encounter verb-tense relationships far more densely in pre-training than any single historical fact.

#	Year	Prompt
1	2010	"During the FIFA World Cup in South Africa, there"
2	2010	"During the earthquake response in Haiti, there"
3	2011	"During the Arab Spring protests across the Middle East, there"
4	2011	"At the time of Osama bin Laden's killing, there"
5	2012	"At the time of the Higgs boson discovery at CERN, there"
6	2013	"During the Edward Snowden NSA surveillance revelations, there"
7	2014	"During Russia's annexation of Crimea, there"
8	2014	"During the Ebola outbreak in West Africa, there"
9	2015	"During the Paris climate talks, there"
10	2016	"During the Brexit referendum campaign, there"
11	2016	"During the US presidential election campaign won by Trump, there"
12	2018	"During the Cambridge Analytica data scandal, there"
13	2019	"During the Hong Kong pro-democracy protests, there"
14	2019	"During Greta Thunberg's school climate strikes, there"
15	2020	"During the initial COVID-19 lockdowns, there"
16	2020	"During the Black Lives Matter protests following George Floyd's death, there"
17	2021	"During the storming of the US Capitol by Trump supporters, there"
18	2021	"During the Taliban's takeover of Afghanistan, there"
19	2022	"During Russia's full-scale invasion of Ukraine, there"
20	2022	"During Elon Musk's acquisition of Twitter, there"
21	2023	"During the Hollywood writers' and actors' strikes, there"
22	2023	"During the Israel-Gaza conflict following the October 7th attacks, there"
23	2024	"During the Paris Olympics opening ceremony on the Seine, there"
24	2024	"During the US presidential election campaign won by Trump, there"
25	2025	"During the inauguration of the 47th US president in Washington, there"
26	2025	"During the DeepSeek AI model release from China, there"
27	2026	"During the FIFA World Cup hosted jointly by the US, Canada, and Mexico, there"
28	2026	"During the Winter Olympics in Milan and Cortina, there"
29	2028	"During the Summer Olympics in Los Angeles, there"

Table 7: **Implicit contextual association prompts.** Each prompt names a real-world event tied to a ground-truth year, without stating the year explicitly. We use 2–3 events per year over 2010–2028, spanning events before, at, and after the training data cutoff.

D Factual recency as an implicit current-year probe

The ACY recovers the current year from a single grammatical signal, verb tense (Section 3). It is therefore natural to ask whether the current year is recoverable from an entirely different, non-grammatical signal: a model's factual knowledge of when recurring events last occurred. Here we probe this via *factual recency*. We ask the LM for the *most recent* instance of a recurring event and read the year implied by the answer it names. For example,

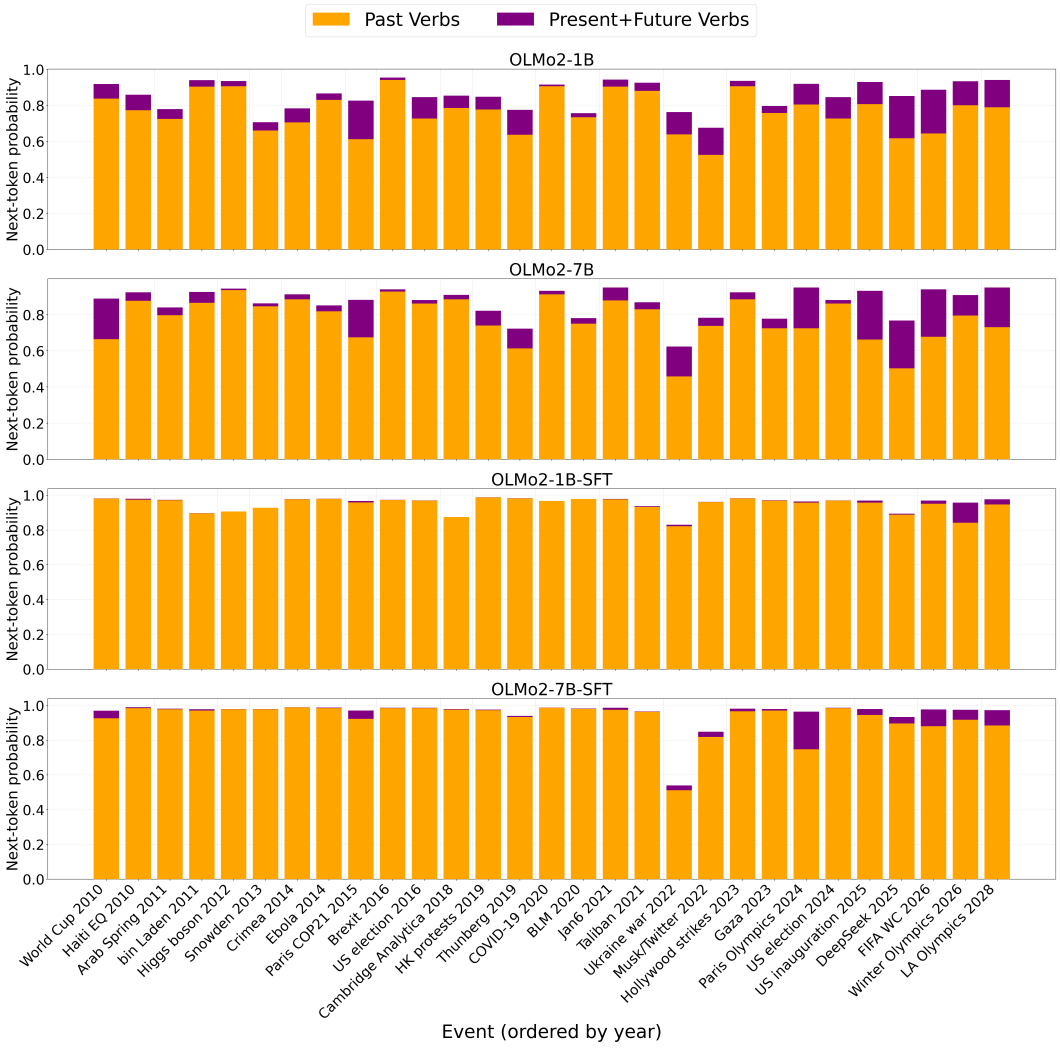


Figure 8: **The implicit contextual association task does not recover the ACY signal.** For each event prompt (ordered by year on the x-axis), we plot the next-token probability mass on past verbs (orange) versus present+future verbs (purple), for base and SFT versions of `OLMo2-1B` and `OLMo2-7B`. Past tense dominates across all events with no clear rise in present+future probability toward the training data cutoff.

naming "PyeongChang" as the most recent Winter Olympics host implies a current year of 2018–2021. This differs from the tense-based factual-recall probe of Appendix C: there we read verb tense off an event prompt, whereas here we read the implied year directly from the named answer. It lets us ask whether a model's factual knowledge yields a consistent estimate of the current year, and how that estimate relates to the training-data cutoff. Because the year must be recovered through recall of named events, this probe has two inherent limitations, shared with the tense-based probe of Appendix C. First, we cannot construct prompts about events in the far future, since such facts do not yet exist; factual recency therefore cannot test whether an LM can be shifted toward future years, a key capability we evaluate elsewhere in the paper. Second, the probe is far more labor-intensive than the ACY task, and thus harder to scale: prompts must be curated by hand per fact, and every response must then be graded manually.

Setup We curate 18 recurring facts whose "most recent" answer is well defined and changes over time, such as the host city of the most recent Summer Olympics or the most recently inaugurated US President (full list in Table 8). To reduce sensitivity to any single wording, we write three phrasings per fact; each phrasing is realized both as a sentence-completion (continuation) prompt for base LMs and as a matching question prompt for SFT LMs, giving $18 \times 3 \times 2 = 108$ prompt templates. Each model is evaluated on the 54 (fact, phrasing) items in its own form. For every response we hand-annotate the entity named and the range of years over which that entity is the correct "most recent" answer, counting a year as covered whenever the entity is the correct answer on at least one day of that year. We label a response *current* if its range covers the training-data cutoff ($\sim$2023 for `OLMo2` LMs), *behind* if it names a real but older answer, *hallucinated* if it names no valid title-holder, and *refused* if the LM declines to answer. We evaluate the base and SFT versions of `OLMo2-1B` and `OLMo2-7B`.

Results If factual recall encoded a coherent current year, a model's answers would imply a single, consistent year clustered on the cutoff. Instead, as shown in Figure 9, no model produces a consistent implied year across the 18 facts. Base LMs give the current answer for only 4% of prompts and a real but older answer for 81%, hallucinating on the remaining 15% and never refusing. SFT LMs instead refuse the majority of prompts (54%) and hallucinate rarely (5%); when they do answer, their implied years have lower variance and sit closer to the cutoff (25% current). In contrast to the tense-based ACY, which recovers the cutoff to within ten months (Table 1), factual recency badly underestimates and scatters the implied year, reinforcing the finding of Appendix C that factual recall is not a reliable route to the current year.

Fact	Continuation form (base LMs)	Question form (SFT LMs)
Summer Olympics *(Tokyo)*	(1) "The most recent Summer Olympics was held in the city of" (2) "The host city of the most recent Summer Olympics was" (3) "The most recent Summer Olympic Games took place in"	(1) "What city hosted the most recent Summer Olympics?" (2) "Which city most recently hosted the Summer Olympics?" (3) "In what city did the most recent Summer Olympic Games take place?"
Winter Olympics *(Beijing)*	(1) "The most recent Winter Olympics was held in the city of" (2) "The host city of the most recent Winter Olympics was" (3) "The most recent Winter Olympic Games took place in"	(1) "What city hosted the most recent Winter Olympics?" (2) "Which city most recently hosted the Winter Olympics?" (3) "In what city did the most recent Winter Olympic Games take place?"
Nobel Peace Prize *(Mohammadi)*	(1) "The most recent Nobel Peace Prize was awarded to" (2) "The most recent winner of the Nobel Peace Prize is" (3) "The person who most recently received the Nobel Peace Prize is"	(1) "Who most recently won the Nobel Peace Prize?" (2) "Who is the most recent Nobel Peace Prize laureate?" (3) "Who was the most recent recipient of the Nobel Peace Prize?"

continued on next page

Table 8 (continued)

Fact	Continuation form (base LMs)	Question form (SFT LMs)
Nobel Physics *(Krausz)*	(1) "The most recent Nobel Prize in Physics was awarded to" (2) "The most recent winner of the Nobel Prize in Physics is" (3) "The most recent Nobel Prize in Physics laureate is"	(1) "Who most recently won the Nobel Prize in Physics?" (2) "Who is the most recent Nobel Prize in Physics winner?" (3) "Who is the most recent Nobel Physics Prize laureate?"
Nobel Chemistry *(Bawendi)*	(1) "The most recent Nobel Prize in Chemistry was awarded to" (2) "The most recent winner of the Nobel Prize in Chemistry is" (3) "The most recent Nobel Prize in Chemistry laureate is"	(1) "Who most recently won the Nobel Prize in Chemistry?" (2) "Who is the most recent Nobel Prize in Chemistry winner?" (3) "Who is the most recent Nobel Chemistry Prize laureate?"
Nobel Literature *(Fosse)*	(1) "The most recent Nobel Prize in Literature was awarded to" (2) "The most recent winner of the Nobel Prize in Literature is" (3) "The most recent Nobel Prize in Literature laureate is"	(1) "Who most recently won the Nobel Prize in Literature?" (2) "Who is the most recent Nobel Prize in Literature winner?" (3) "Who is the most recent Nobel Literature Prize laureate?"
Nobel Medicine *(Kariko)*	(1) "The most recent Nobel Prize in Physiology or Medicine was awarded to" (2) "The most recent winner of the Nobel Prize in Physiology or Medicine is" (3) "The most recent Nobel Prize in Physiology or Medicine laureate is"	(1) "Who most recently won the Nobel Prize in Physiology or Medicine?" (2) "Who is the most recent Nobel Prize in Medicine winner?" (3) "Who is the most recent Nobel Medicine Prize laureate?"
US President *(Biden)*	(1) "The most recent US presidential election was won by" (2) "The most recently inaugurated President of the United States is" (3) "The most recent winner of the US presidential election is"	(1) "Who most recently won the US presidential election?" (2) "Who is the most recently inaugurated President of the United States?" (3) "What is the name of the most recent US presidential election winner?"
UK Prime Minister *(Sunak)*	(1) "The most recently appointed Prime Minister of the United Kingdom is" (2) "The most recent person to become Prime Minister of the United Kingdom is" (3) "The most recent Prime Minister of the United Kingdom is"	(1) "Who is the most recently appointed UK Prime Minister?" (2) "Who most recently became UK Prime Minister?" (3) "Who is the most recent Prime Minister of the United Kingdom?"
Academy Award *(Everything)*	(1) "The most recent Academy Award for Best Picture was won by the film" (2) "The most recent winner of the Oscar for Best Picture is the film" (3) "The film that most recently won the Academy Award for Best Picture is"	(1) "What film most recently won the Academy Award for Best Picture?" (2) "What is the most recent Oscar Best Picture winner?" (3) "Which film most recently won the Oscar for Best Picture?"
Time Person of the Year *(Swift)*	(1) "The most recent Time Person of the Year is" (2) "The most recent winner of Time magazine's Person of the Year award is" (3) "Time magazine's most recent Person of the Year is"	(1) "Who is the most recent Time Person of the Year?" (2) "Who most recently won Time's Person of the Year?" (3) "Who is Time magazine's most recent Person of the Year?"

continued on next page

Table 8 (continued)

Fact	Continuation form (base LMs)	Question form (SFT LMs)
M-W Word of the Year *(authentic)*	(1) "The most recent Merriam-Webster word of the year is" (2) "The word most recently named Merriam-Webster's word of the year is" (3) "Merriam-Webster's most recent word of the year is"	(1) "What is the most recent Merriam-Webster word of the year?" (2) "What word was most recently named Merriam-Webster's word of the year?" (3) "What is Merriam-Webster's most recent word of the year?"
Chinese Zodiac *(Rabbit)*	(1) "The most recent Chinese Zodiac year is the Year of the" (2) "The Chinese Zodiac animal for this year is the" (3) "In the Chinese Zodiac, this year is the Year of the"	(1) "What is the most recent Chinese Zodiac animal?" (2) "What is the Chinese Zodiac animal for this year?" (3) "According to the Chinese Zodiac, what animal represents this year?"
FIFA World Cup *(Qatar)*	(1) "The most recent FIFA World Cup was hosted by the country of" (2) "The host country of the most recent FIFA World Cup was" (3) "The most recent FIFA World Cup took place in the country of"	(1) "Which country most recently hosted the FIFA World Cup?" (2) "What country most recently hosted the FIFA World Cup?" (3) "In what country did the most recent FIFA World Cup take place?"
US Vice President *(Harris)*	(1) "The most recently inaugurated Vice President of the United States is" (2) "The most recent winner of the US vice presidential election is" (3) "The most recent Vice President of the United States is"	(1) "Who is the most recently inaugurated Vice President of the United States?" (2) "Who most recently became Vice President of the United States?" (3) "Who is the most recent Vice President of the United States?"
German Chancellor *(Scholz)*	(1) "The most recently elected Chancellor of Germany is" (2) "The most recent person to become Chancellor of Germany is" (3) "The most recent Chancellor of Germany is"	(1) "Who is the most recently elected Chancellor of Germany?" (2) "Who most recently became Chancellor of Germany?" (3) "Who is the most recent Chancellor of Germany?"
Italian PM *(Meloni)*	(1) "The most recently appointed Prime Minister of Italy is" (2) "The most recent person to become Prime Minister of Italy is" (3) "The most recent Prime Minister of Italy is"	(1) "Who is the most recently appointed Prime Minister of Italy?" (2) "Who most recently became Prime Minister of Italy?" (3) "Who is the most recent Prime Minister of Italy?"
Brazilian President *(Lula)*	(1) "The most recently elected President of Brazil is" (2) "The most recent winner of the Brazilian presidential election is" (3) "The most recent President of Brazil is"	(1) "Who is the most recently elected President of Brazil?" (2) "Who most recently won the Brazilian presidential election?" (3) "Who is the most recent President of Brazil?"

Table 8: **Factual recency probe: the 18 recurring facts and all prompts.** Each fact (with its answer as of the 2023 cutoff in parentheses) has a well-defined "most recent" answer that changes over time. We write three phrasings per fact, each realized as a sentence-completion (continuation) prompt for base LMs and a matching question prompt for SFT LMs.

E Obtaining perplexities on the WIKISPAN dataset

Cheng et al. (2024) estimate the effective knowledge cutoff date of a LM by measuring perplexity across versions of data over time. For the first eight models listed in Table 1, we extract the effective cutoff predictions directly from Figure 4 of their paper by identifying

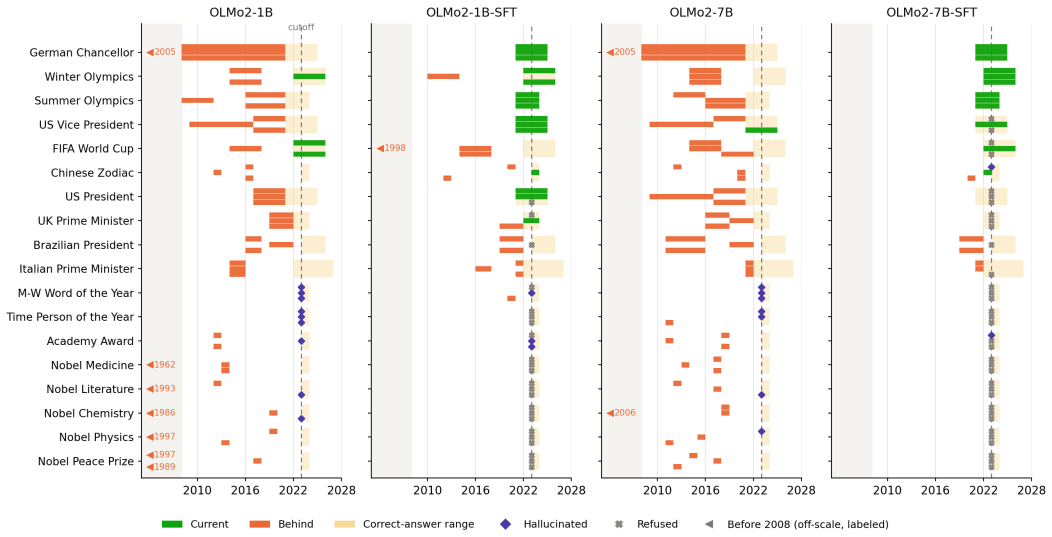


Figure 9: **Factual recency does not cleanly encode the current year.** Each row is one of 18 recurring facts; for its three phrasings we draw the year range over which the LM's named answer is correct; *current* (green) if the range covers the ∼2023 cutoff (dashed line), *behind* (orange) if the answer is real but older. Alternatively, we draw a marker for a *hallucinated* (⋄) or *refused* (×) response. The gold band is the correct-answer range; a left-pointing mark in the margin flags an answer from before 2008 (labeled with its year). Columns are the base and SFT versions of `OLMo2-1B` and `OLMo2-7B`.

the month and year that achieves the lowest relative perplexity. The remaining five models are not covered in their analysis; for these, we replicate the WIKISPAN dataset construction and perplexity measurement using the publicly available code accompanying their paper. The resulting plots, shown in Figure 10, are consistent with the format of their original Figure 4.

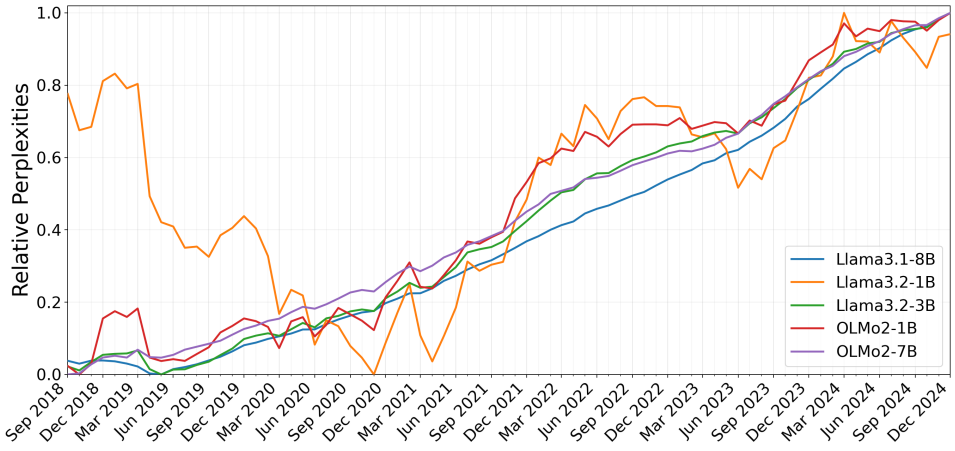


Figure 10: **Relative perplexities of LMs per month using the WIKISPAN dataset.**

F N-gram and co-occurrence models

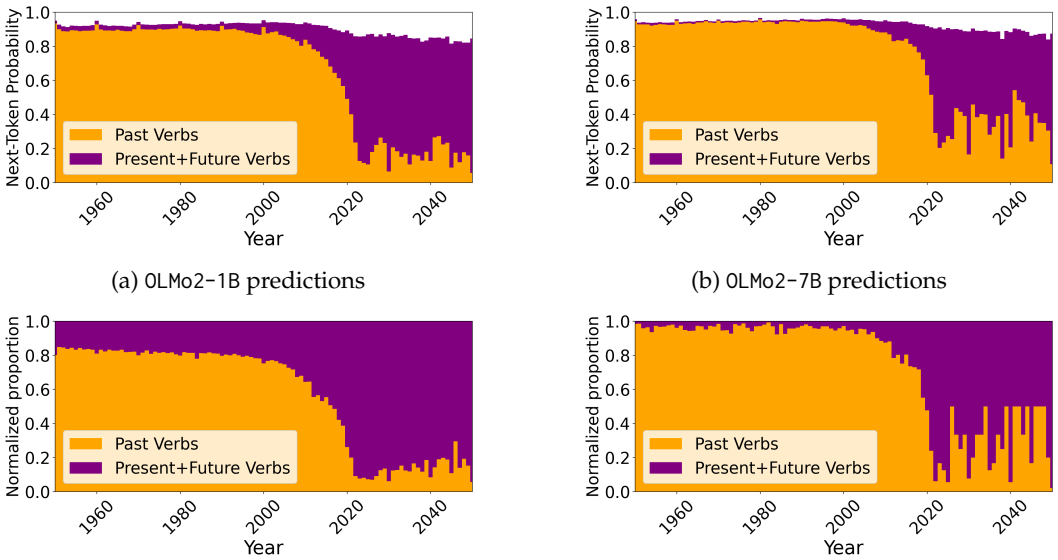
Creating N-gram and co-occurrence models We create an N-gram and co-occurrence model from training data counts of both olmo-mix-1124 and dolmino-mix-1124, which are

used in stage 1 and 2 of pre-training base models OLMo2-1B and OLMo2-7B. We detail the creation of these models here.

Co-occurrence model We split the training data into sequences, as is done when training, then further split these sequences into sentences using the nltk python library and additionally split sentences longer than 20 words. This additional splitting is necessary because the training data contains sections without punctuation that extend for hundreds of words, which would cause inaccurate co-occurrence counts if the first and last words of such long passages were considered related. We then look for co-occurrences of the string *in [year]* and instances of *was*, *were*, *is*, *are* and *will* anywhere within each sentence and collect counts of these combinations. Finally, we apply Laplacian smoothing with an alpha of 1.0 to create the co-occurrence model.

N-gram model We split the training data into sequences, as is done when training. We then look for instances of *was*, *were*, *is*, *are* and *will* directly following the string *In [year] there* (to match how we prompt the LM) using a case insensitive regex that requires words be split by word boundaries (\b) in the training corpus. We combine *was* and *were* as past tense and *is*, *are* and *will* as present+future tense, then apply Laplacian smoothing with an alpha of 1.0 to create the n-gram model.

N-gram model, co-occurrence model, and OLMo2-1B/OLMo2-7B associative year-tense behavior We show the normalized tense predictions for the n-gram model, co-occurrence model, and OLMo2-1B/OLMo2-7B in Figures 11 and 12 for stage 1 and 2 training comparitively. We use these in Section 3.4 to compute the CE.



(a) OLMo2-1B predictions

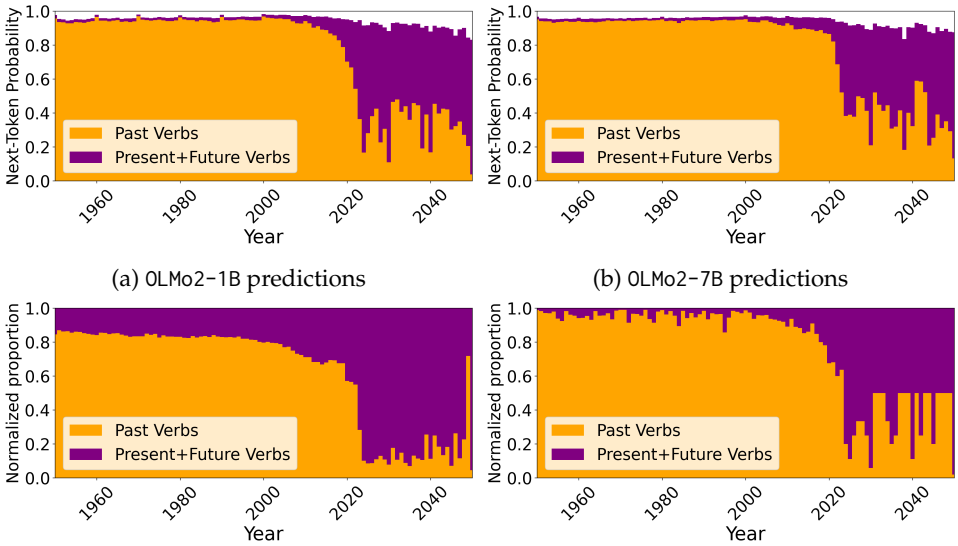
(b) OLMo2-7B predictions

(c) Co-occurrence predictions

(d) N-gram model predictions

Figure 11: **Next token predictions on OLMo2 base stage1 LM.** 11a and 11b show next-token predictions of the LM after stage2, 11c shows the normalized co-occurrence model created from 10k steps of olmo-mix-1124, 11d shows the normalized n-gram model created from 10k steps of olmo-mix-1124. The associative task follows the training data distribution.

Cross entropy between OLMo2 LMs and n-gram and co-occurrence models Figure 9 shows the cross entropy between OLMo2 LMs and models constructed from pre-training data at Stage 1 and Stage 2. We see the LMs have a lower cross entropy with the count-based models than with a random baseline, suggesting the ACY is shaped by the pre-training data distribution.



(a) OLMo2-1B predictions

(b) OLMo2-7B predictions

(c) Co-occurrence predictions

(d) N-gram model predictions

Figure 12: **Next token predictions on OLMo2 base stage2 LM.** 12a and 12b show next-token predictions of the LM after stage2, 12c shows the normalized co-occurrence model created from 10k steps of dolmino-mix-1124, 12d shows the normalized n-gram model created from 10k steps of dolmino-mix-1124. The associative task follows the training data distribution.

Count model	LM	Stage 1	Stage 2
N-gram	OLMo2-1B	0.34 (−0.40)	0.34 (−0.19)
	OLMo2-7B	0.35 (−0.28)	0.36 (−0.12)
Co-occurrence	OLMo2-1B	0.61 (−0.42)	0.68 (−0.21)
	OLMo2-7B	0.78 (−0.33)	0.70 (−0.16)

Table 9: **OLMo2 LMs have lower cross-entropy with corpus-count models than baselines.** The first two sections show the CE difference against the N-gram and co-occurrence baselines, with the delta versus the corresponding random baseline in parentheses. The last two sections show the CE difference against random token selections drawn from N-gram and co-occurrence count distributions, respectively.

G Controls and robustness for the associative year-type subspace

Details on DAS procedure Here, we offer details on the standard DAS procedure used in Section 3.5. We construct 20480 training pairs and 1024 evaluation pairs of base and source examples using prompts of the form *In [year], there*, where the year ranges from 1000 to 4000. We keep correct prompts where the LM naturally predicts the expected tense (past for years before the training data cutoff and present or future after). On correct prompts, we know models can correctly reason about the prompt year with respect to its current year, which suggests a year-type subspace is being used. We train and evaluate on correct prompts with no overlap between sets to remove bias.

We further analyze the one-dimensional "year-type" subspace whose interchange interventions shift tense. Here, we provide sanity checks demonstrating that this effect is non-trivial (i.e., not an artifact of the evaluation setup) and that it is specific to year-like representations.

Random subspace As a baseline, we replace the DAS-learned one-dimensional subspace with a randomly chosen one-dimensional subspace at the same layer and token position. Table 10 shows that this yields an IIA of 14% across layers, matching the percentage of evaluation pairs in which the target tense does not change. Thus, interchanging an arbitrary subspace does not systematically control the year-tense mechanism.

Randomly initialized model Next, we test whether DAS can obtain high IIA even in the absence of learned structure in model weights. We randomly initialize the weights of `OLMo2-7B`, run DAS to fit a one-dimensional subspace using the same training procedure, and evaluate on the same evaluation set used for the year-type experiments. This achieves 0% IIA across layers, indicating that high-IIA solutions depend on learned representations rather than the DAS procedure alone.

Interchanges with non-numeric, non-year entities We then evaluate the learned year-type subspace out of distribution. We let base prompts remain of the form *In [year], there*, while source prompts contain non-year entities such as *In summary, there* or *In Amsterdam, there* (full prompt list in Table 11). Since all these source prompts elicit present tense predictions, we restrict base prompts to those that elicit past tense predictions. Under these interchanges, IIA is mostly 0%, peaking at 46%) in some layers (which is below random guessing), suggesting the localized subspace does not support tense control when the year subspace is filled by unrelated non-numeric entities.

Interchanges with four-digit numbers that are not used as years We again reuse the learned year-type subspace, but now the source prompts contain year-like numbers in non-temporal contexts, specifically math. For example, for a prompt *732 - 1 + 1789 is*, we perform interchanges at the number token(s) (e.g., on the *9* tokens). In this setting, interchanges succeed in early layers, reaching $>$ 90% IIA in five layers (Table 10). This indicates that the localized subspace generalizes to numbers that resemble years even when they are not explicitly used as temporal expressions, This effect disappears in later layers, suggesting that the year-type then becomes specific to temporal contexts.

Interchanges under a different template and output tokenization Finally, we test robustness to changes in surface form and output casing by changing the original template to uppercase: *IN [year] THERE*. This shifts the target next-token predictions from *was, were, is, are, will* to *WAS, WERE, IS, ARE, WILL*. One complication is that *WERE* can be tokenized as *W* + *ERE*; in such cases, the LM assigns most probability mass after *W* to *ERE*. We therefore treat *W* as the uppercase realization of *were* for evaluation. With this modification, interchanges remain largely successful but achieve consistently lower IIA than the lowercase setting across layers (Table 10), suggesting the localized subspace is not merely tied to a single fixed output token but is somewhat sensitive to template and output token used.

Layer	Random sub-space	Randomly initialized model	Non-year non-number entities	Non-year num-bers	Different template structure
31	0.14	0.00	0.00	0.00	0.24
30	0.14	0.00	0.00	0.00	0.24
28	0.14	0.00	0.00	0.08	0.28
26	0.14	0.00	0.00	0.10	0.26
24	0.14	0.00	0.00	0.15	0.28
22	0.14	0.00	0.39	0.35	0.39
20	0.14	0.00	0.14	0.35	0.39
18	0.14	0.00	0.45	0.62	0.41
16	0.14	0.00	0.46	0.72	0.51
14	0.14	0.00	0.09	0.86	0.62
12	0.14	0.00	0.02	0.87	0.61
10	0.14	0.00	0.00	0.88	0.62
8	0.14	0.00	0.01	0.91	0.67
6	0.14	0.00	0.00	0.97	0.71
4	0.14	0.00	0.00	1.00	0.74
2	0.14	0.00	0.00	1.00	0.71
1	0.14	0.00	0.00	1.00	0.67
0	0.14	0.00	0.00	0.67	0.29

Table 10: **The year-type subspace is non-trivial and specific to year–like representations.** IIA across layers on the second token of the year for various setups.

#	Prompt
Non-year non-number prompts	
1	"In summary there"
2	"In Amsterdam there"
3	"In water there"
4	"In math there"
5	"In theory there"
6	"In general there"
7	"In conclusion there"
8	"In practice there"
9	"In contrast there"
10	"In reality there"
11	"In essence there"

(a) **Prompts used to interchange with non-numeric words.** Non-year, non-numeric words are substituted into the *In year there* template.

#	Prompt
Non-year number prompts	
1	"16 * 1850 + 1953 is"
2	"4 - 3 + 1741 is"
3	"100 + 200 + 1876 is"
4	"2000 - 100 + 1492 is"
5	"50 * 2 + 1066 is"
6	"999 - 1 + 1789 is"
7	"7 * 8 + 1945 is"
8	"86828 + 4 - 2001 is"
9	"3 * 63 + 1000 + 1998 is"

(b) **Prompts used to interchange numbers that look like years but are not.** Arithmetic expressions containing year-like numbers in non-temporal contexts

Table 11: **All prompts used for the associative interchange robustness experiments.**

H Causal tracing through attention and MLP knockouts

In Section 3.5, we use learned interchange interventions on the residual stream to probe how the LM solves the associative task. Here, we use null interventions, replacing attention and MLP additions with a zero vector to measure their causal contribution to the final tense prediction.

Methods We use attention/MLP sublayer knockouts (Geva et al., 2023): for each intervention site, we zero out either the attention or MLP output over five consecutive layers on one token. For each intervention, we measure a change in probability of tense before and after the knockout by measuring tense shift magnitude (TSM) as

$$\begin{aligned}\text{TSM}_y \quad = \quad & |p_y(\text{past})_{\text{before}} - p_y(\text{future})_{\text{before}} \\ & -\left(p_y(\text{past})_{\text{after}} - p_y(\text{future})_{\text{after}}\right)|\end{aligned}$$

Causal tracing corroborates the year-type localization results We do attention and MLP knockouts across 70 prompts on which the LM behaves correctly. The average TSM across each group can be seen in Figure 13. We see that the largest tense shifts occur when we knock out computations on earlier layers of year tokens and on the later layers of the final token, which are consistent with the causal abstraction results (Figure 3).

Sublayer knockouts reveal asymmetries across past vs. future years Notably, past-year and future-year prompts respond differently: Knockouts for prompts with future years have a larger effect which are seen on the second year token (i.e. the last digit), whereas prompts with past years are less intensely affected by the second token of the year. This pattern suggests that the LM resolves tense using the earliest year-token evidence available, but relies more heavily on late-year digits when the year is interpreted as future.

I Declarative prompts and behavior

We use a set of prompts to find the declarative current year. The respective set of continuation prompts for base models and instruction prompts for fine-tuned models can be seen in Table 12.

It is notable that much effort was placed into selecting a set of ten prompts to which both the base and SFT models for OLMo2-1B and OLMo2-7B responded to with a year.

J Training data counts of years

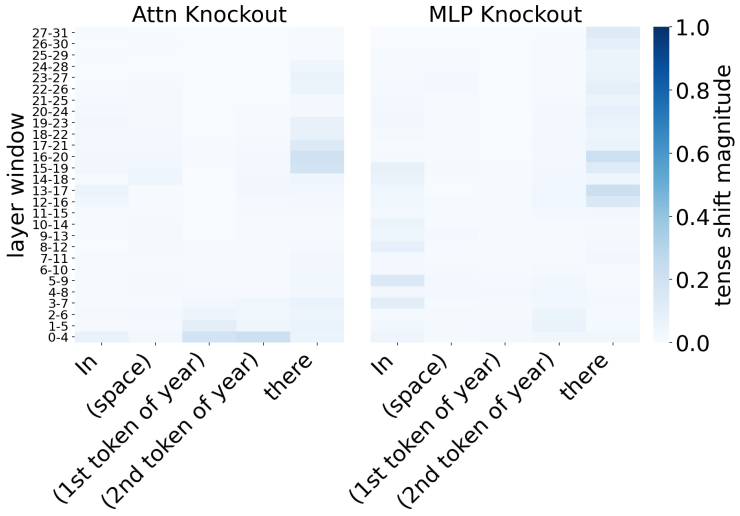
Figure 14 shows the normalized counts of each year on the range 1900 to 2100 in training data. We use olmo-mix ((`OLMo2` stage 1 pre-training (Allen Institute for AI, 2024b)), dolmino (`OLMo2` stage 2 pre-training (Allen Institute for AI, 2024a)), and tulu3 (`OLMo2-1B` SFT (Allen Institute for AI, 2025a), and `OLMo2-7B SFT (Allen Institute for AI, 2025b)`).

We show counts of raw 4-digit numbers appearing in text and numbers matching the regex *In [year]*. We use the frequency counts from the latter method because counts collected with this method are more likely to be from years, as opposed to just 4 digit numbers.

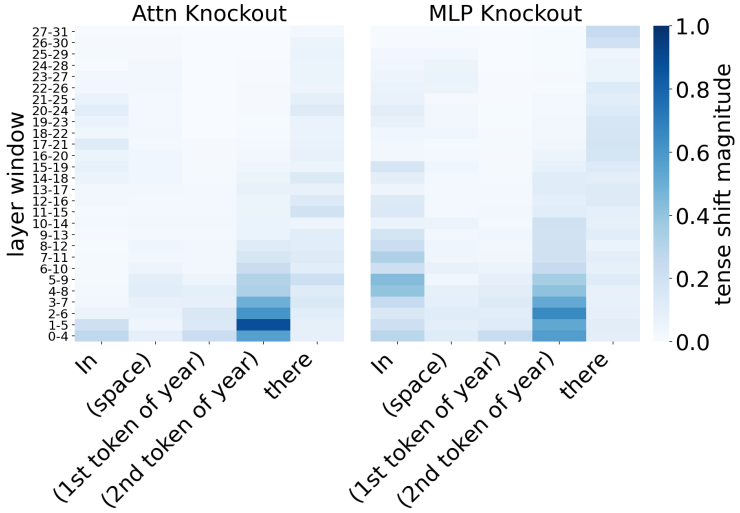
We also show results for an alternate method of counting: Figure 14 shows the counts of *in [year]* for the training data for years on the range 1900-2100.

K SFT training `OLMo2-1B`

To test the effects of training data content on the declarative current year, we train SFT on the Tulu3 dataset (Allen Institute for AI, 2025a), which `OLMo2-1B` uses for their SFT training.

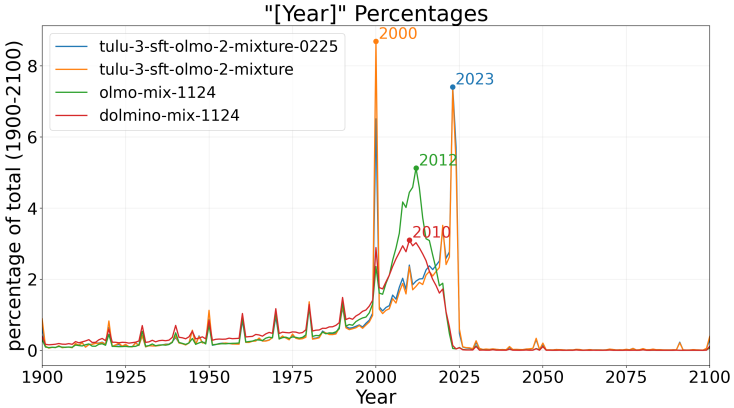


(a) Past years

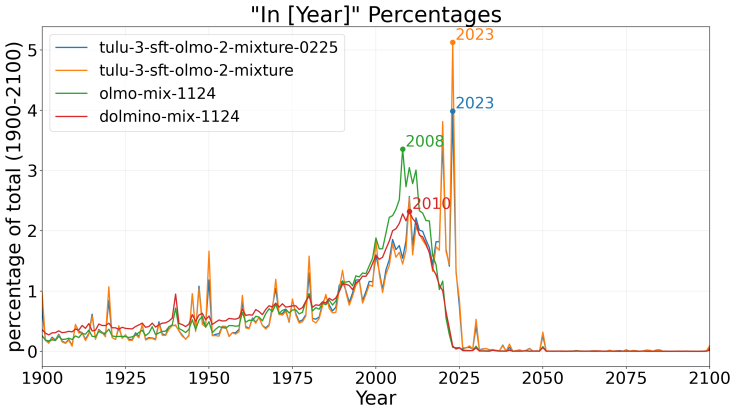


(b) Future years

Figure 13: **Attention and MLP knockout corroborate associative task's causal abstraction results.** We show results for the associative task, measured by tense shift magnitude (TSM), averaged across 35 past-year (top) and 35 future-year (bottom) prompts. We see large TSM in early layers on year tokens and late layers on the final token, the same as in causal abstraction. Future-year prompts show stronger dependence on the second year token (last digit) than past-year prompts.



(a) "[year]" counts (%)



(b) "In [year]" counts (%)

Figure 14: **Year counts for training datasets olmo-mix, dolmino, and tulu3.** Normalized year frequency in pre-training (olmo-mix, dolmino-mix) and SFT (tulu3) datasets, counted as raw 4-digit occurrences (top) and in the *In [year]* context (bottom). The *In [year]* counts are used as the primary measure since they reflect temporal usage better than arbitrary 4-digit numbers.

#	Prompt
	Continuation
1	"Finish the sentence: The current year (in YYYY format) is"
2	"What day of the week is Valentines day this year? For the year"
3	"Today, the year is"
4	"Write a story about the current year. Story: The year is"
5	"Write a yearly summary, log the current year at the top. Year:"
6	"Write a story, starting with the current year. Year:"
7	"Complete the sentence: The year is"
8	"What day of the week is New Years this year? For the year"
9	"Write a short summary, log the current year at the top. Year:"
10	"Write a story. Story: The year is"
	Instruction
11	"Answer the question: What is the current year (YYYY)?"
12	"What year is it, and what day of the week is Valentines day this year?"
13	"As of today, what is the year?"
14	"Write a story about the current year, starting with the current year."
15	"Write a yearly summary, log the current year at the top."
16	"Write a story, starting with the current year."
17	"Respond to the prompt: What's the year?"
18	"What year is it, and what day of the week is New Years this year?"
19	"Write a short summary, log the current year at the top."
20	"Write a story, where the story starts with the current year."

Table 12: **Declarative prompts**. Prompts grouped by continuation and instruction. These prompts were used in Section 4.

We use an effective batch size of 192 (per-device batch size of 4, gradient accumulation of 8, across 6 GPUs), a learning rate of 5e-5 with linear decay and no warmup, and train for 1 epoch with a maximum sequence length of 1024.

No-Years model First, we measure the effects of *removing* all training examples with years (i.e. all number ≥ 1000) from the SFT training data, preserving 76% of training examples. We refer to this model as the No-years SFT model.

+/-N year-shifted models Next, we measure the effects of *shifting* all years in the training data by a constant offset N, and we do not delete any training data. We shift all numbers between [1000-N, 9999-N] by a constant N for N = 200, -15, +/-6, +/-18, and +/-35 (eight models). We choose -15 since it aligns the SFT peak year to 2008, the most frequent year in pre-training. The other shift offsets are chosen to land on a year that is not a round multiple of 5 and equally spaced on either side of 2023. We refer to these year-shifted models by their shift, +/-N.

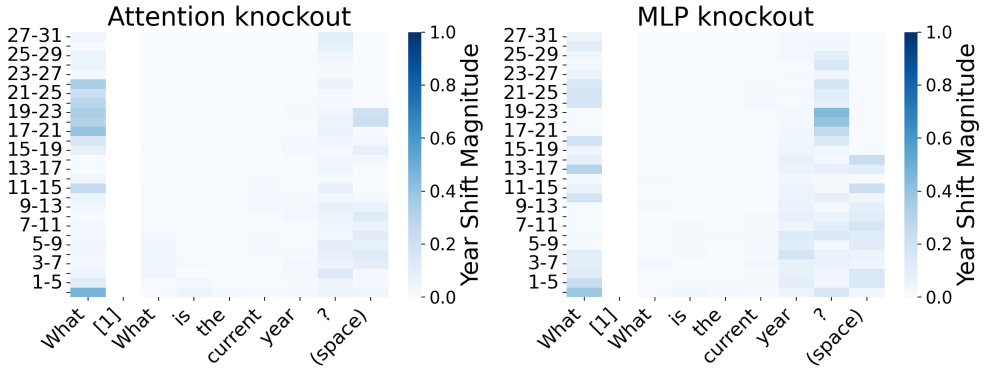
In-context models In addition to the models trained on custom datasets, we produce two in-context models, trained with the system prompts `"The current year is 2023"` and `"Today Date: 26 Jul 2023"` respectively. The results reported in Section 5 use the former, evaluated with the system prompt `"The current year is [target year]"`.

We will release our models and training data.

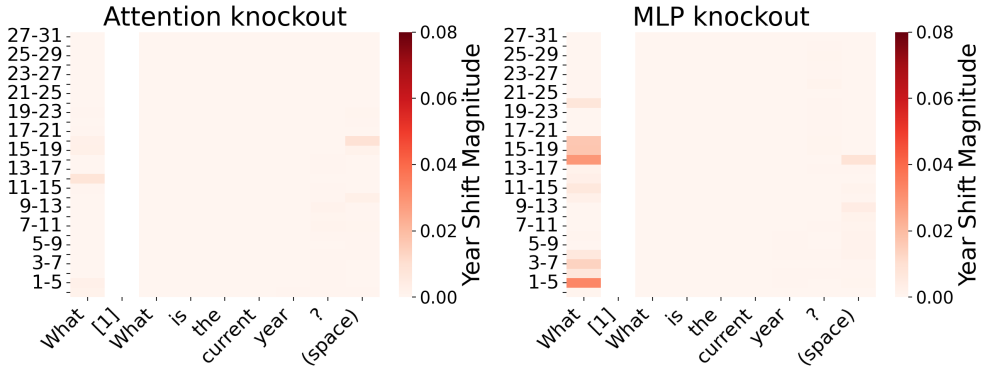
L Declarative causal tracing

Causal tracing on oneshot prompts The mean magnitude and variance shift when applying causal tracing to declarative prompts, shown in Figure 15.

Fewshot declarative prompts Prompts are used for mechanistic analysis, shown in Table 13 for the current-year prompts and Table 14 for the today-date prompts.

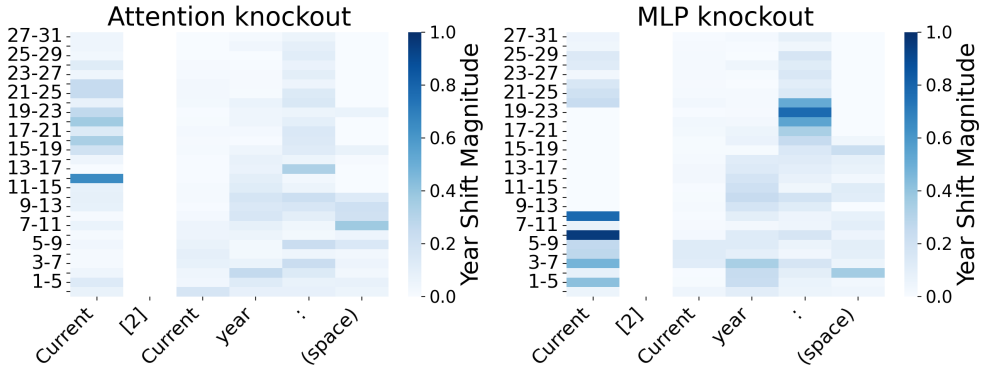


(a) Mean. [1] is *is the current month?[month]. What is the current day?[day].*

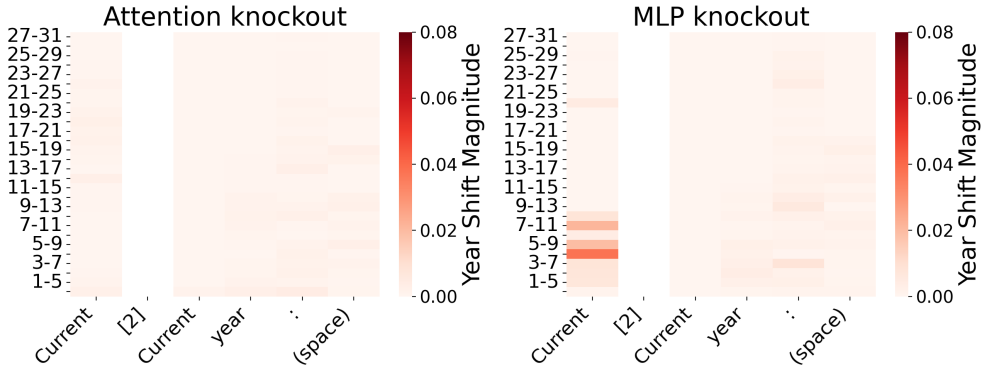


(b) Variance. [1] is *is the current month?[month]. What is the current day?[day].*

Figure 15: **Mean and variance of causal tracing magnitude shift for declarative prompts** . . . (continued)

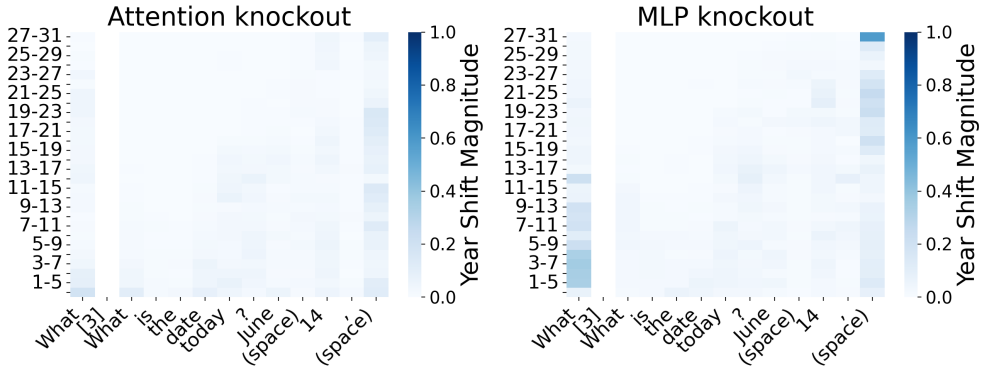


(c) Mean. [2] is *month:[month]. Current day:[day].*

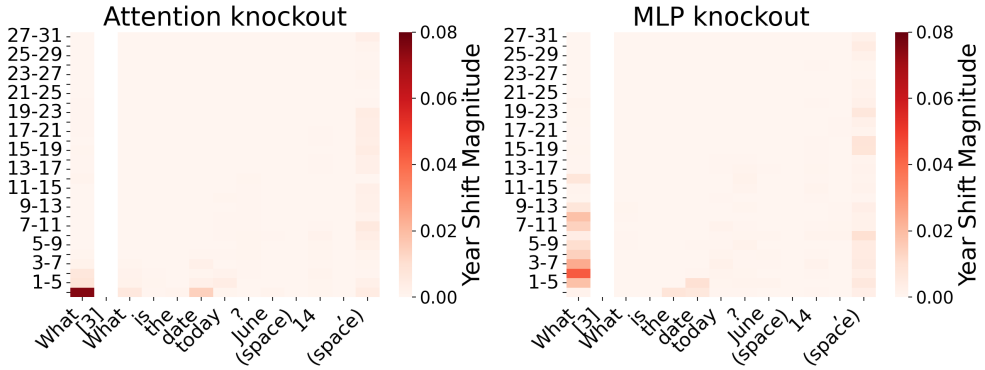


(d) Variance. [2] is *month:[month]. Current day:[day].*

Figure 15: **Mean and variance of causal tracing magnitude shift for declarative prompts** . . . (continued)

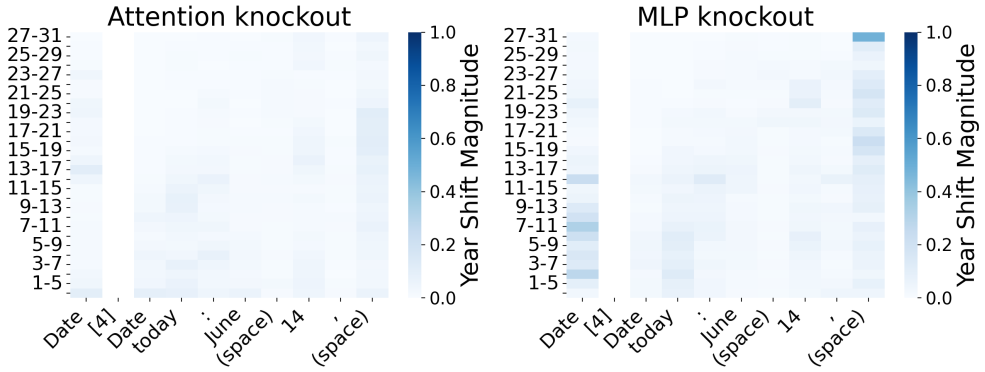


(e) Mean. [3] is *is the date[event1]?[date1]. What is the date[event2]?[date2].*

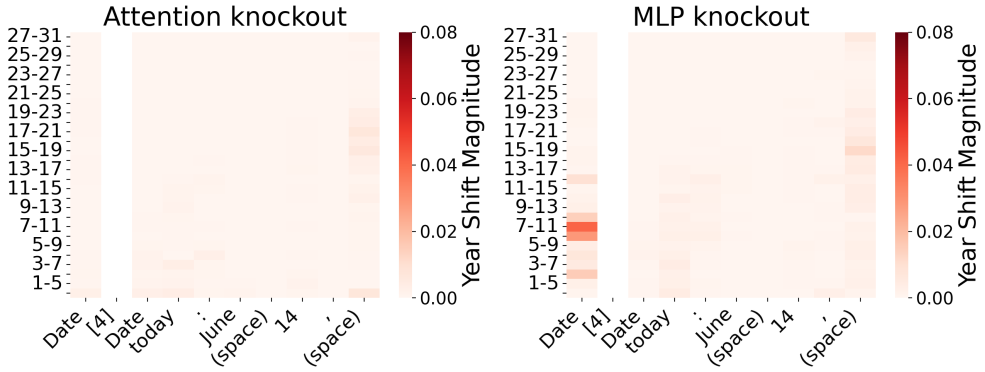


(f) Variance. [3] is *is the date[event1]?[date1]. What is the date[event2]?[date2].*

Figure 15: **Mean and variance of causal tracing magnitude shift for declarative prompts** . . . (continued)



(g) Mean. [4] is *[event1]:[date1]. Date[event2]:[date2]*



(h) Variance. [4] is *[event1]:[date1]. Date[event2]:[date2]*

Figure 15: **Mean and variance of causal tracing magnitude shift for declarative prompts**, shown across four prompt types ([1]–[4]) in question and continuation formats. Unlike the associative task, there is no consistent layer or token position across prompt types, reflecting the lack of a stable causal pathway for the declarative current year. [1], [2], [3], and [4] are used to refer to the varying few-shot examples in the prompt, which differ in token length and therefore cannot be shown in the plots.

#	Prompt
current_year_question	
1	"What is the current month? May. What is the current day? Sunday. What is the current year?"
2	"What is the current month? October. What is the current day? Monday. What is the current year?"
3	"What is the current month? January. What is the current day? Friday. What is the current year?"
4	"What is the current month? February. What is the current day? Tuesday. What is the current year?"
5	"What is the current month? March. What is the current day? Thursday. What is the current year?"
6	"What is the current month? April. What is the current day? Saturday. What is the current year?"
7	"What is the current month? June. What is the current day? Wednesday. What is the current year?"
8	"What is the current month? July. What is the current day? Monday. What is the current year?"
9	"What is the current month? August. What is the current day? Sunday. What is the current year?"
10	"What is the current month? December. What is the current day? Tuesday. What is the current year?"

#	Prompt
current_year_stated	
1	"Current month: May. Current day: Sunday. Current year:"
2	"Current month: October. Current day: Monday. Current year:"
3	"Current month: January. Current day: Friday. Current year:"
4	"Current month: February. Current day: Tuesday. Current year:"
5	"Current month: March. Current day: Thursday. Current year:"
6	"Current month: April. Current day: Saturday. Current year:"
7	"Current month: June. Current day: Wednesday. Current year:"
8	"Current month: July. Current day: Monday. Current year:"
9	"Current month: August. Current day: Sunday. Current year:"
10	"Current month: December. Current day: Tuesday. Current year:"

Table 13: **Declarative prompts used for current-year tasks.** We use few-shot prompts to elicit a current year with both a question and continuation format. Month and day vary across prompts to prevent the LM from pattern-matching on a fixed date.

#	Prompt
today_question	
1	"What is the date the Apollo 11 Moon landing occurred? July 20, 1969. What is the date the iPhone was announced? January 9, 2007. What is the date today? June 14, "
2	"What is the date the Berlin Wall fell? November 9, 1989. What is the date the September 11 attacks occurred? September 11, 2001. What is the date today? June 14, "
3	"What is the date the Titanic sank? April 15, 1912. What is the date the Wright brothers made the first powered flight? December 17, 1903. What is the date today? June 14, "
4	"What is the date the Sputnik 1 launch occurred? October 4, 1957. What is the date the World Wide Web was publicly announced at CERN? August 6, 1991. What is the date today? June 14, "
5	"What is the date the U.S. Declaration of Independence was adopted? July 4, 1776. What is the date the U.S. Constitution was signed? September 17, 1787. What is the date today? June 14, "
6	"What is the date the 'I Have a Dream' speech was delivered? August 28, 1963. What is the date the Nelson Mandela release from prison occurred? February 11, 1990. What is the date today? June 14, "
7	"What is the date the attack on Pearl Harbor occurred? December 7, 1941. What is the date the first human spaceflight occurred? April 12, 1961. What is the date today? June 14, "
8	"What is the date the Beatles appeared on The Ed Sullivan Show? February 9, 1964. What is the date the Woodstock festival began? August 15, 1969. What is the date today? June 14, "
9	"What is the date the 19th Amendment was ratified? August 18, 1920. What is the date the Nixon resignation occurred? August 8, 1974. What is the date today? June 14, "
10	"What is the date the start of World War II occurred? September 1, 1939. What is the date the end of World War II occurred? September 2, 1945. What is the date today? June 14, "

#	Prompt
today_stated	
1	"Date the Apollo 11 Moon landing occurred: July 20, 1969. Date the iPhone was announced: January 9, 2007. Date today: June 14, "
2	"Date the Berlin Wall fell: November 9, 1989. Date the September 11 attacks occurred: September 11, 2001. Date today: June 14, "
3	"Date the Titanic sank: April 15, 1912. Date the Wright brothers made the first powered flight: December 17, 1903. Date today: June 14, "
4	"Date Sputnik 1 launched: October 4, 1957. Date the World Wide Web was publicly announced at CERN: August 6, 1991. Date today: June 14, "
5	"Date the U.S. Declaration of Independence was adopted: July 4, 1776. Date the U.S. Constitution was signed: September 17, 1787. Date today: June 14, "
6	"Date the 'I Have a Dream' speech was delivered: August 28, 1963. Date Nelson Mandela was released from prison: February 11, 1990. Date today: June 14, "
7	"Date the attack on Pearl Harbor occurred: December 7, 1941. Date the first human went to space: April 12, 1961. Date today: June 14, "
8	"Date the Beatles appeared on The Ed Sullivan Show: February 9, 1964. Date Woodstock began: August 15, 1969. Date today: June 14, "
9	"Date women's suffrage in the U.S. was ratified: August 18, 1920. Date Watergate ended with Nixon's resignation: August 8, 1974. Date today: June 14, "
10	"Date the start of World War II occurred: September 1, 1939. Date World War II ended: September 2, 1945. Date today: June 14, "

Table 14: **Declarative prompts used for today-date tasks.** We use few-shot prompts to elicit a current year with both a question and continuation format. We fix the month and day as June 14 and vary the historical examples.

M Training dynamics on OLMo2-7B

We compute the ACY and DCY of OLMo2-7B by loading checkpoints across training stages. Figure 16 shows the evolution of both quantities across base pre-training, SFT, RLVR, DPO, and the final Instruct model. As in OLMo2-1B (Figure 2), the ACY for OLMo2-7B remains stable throughout pre-training and the DCY shifts only at the post-training stage.

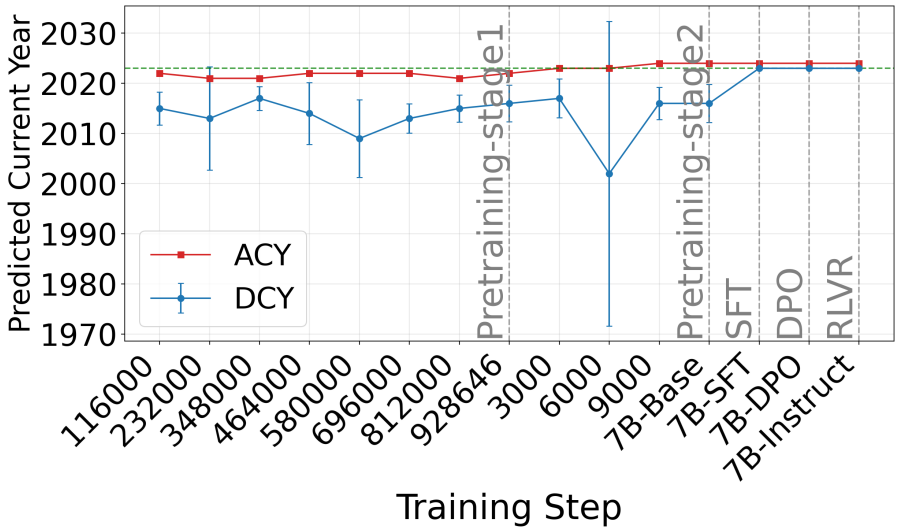


Figure 16: **Training dynamics across training checkpoints for OLMo2-7B.** Points represent the declarative year, and bars represent the standard deviation. The horizontal line indicates the data cutoff year. This is the OLMo2-7B counterpart of Figure 2.

N Prompting fails to shift the ACY across model families and scales

In Section 5 we show that in-context prompting cannot shift the ACY of OLMo2-1B. Here we verify that this conclusion generalizes to a broader set of recent and capable LMs beyond the OLMo2 family. Since several of these LMs are too large to fine-tune or weight-edit with our available compute, we extend only the prompting method, which requires no training.

Setup We extend our analysis to nine additional LMs spanning the Gemma-3 and Qwen-3.5 families. For each LM, we attempt to update the ACY and DCY by specifying the target current year via system prompt, across the range 1900–2250. We report the fraction of the 351 target years for which the predicted year exactly matches the target.

Results A natural hypothesis is that larger, more capable LMs would be better at using the target year in the system prompt, and so would shift both the ACY and DCY more accurately. As Table 5 shows, this is not the case: accuracy varies non-monotonically with model size, both across and within model families. While several LMs achieve high DCY shifting accuracy with only system prompting (e.g., Qwen3.5-4B, Qwen3.5-9B, and Qwen3.5-27B all reach 100%), shifting the ACY remains difficult: the highest accuracy across all new LMs is just 53.3% (Qwen3.5-27B), with every other LM at or below 40% and several near zero. In no case does ACY shifting approach the near-perfect DCY shifting achieved by the same LM. This confirms our conclusion that the ACY cannot be reliably shifted via prompting, even for LMs substantially larger and more recent than OLMo2.