# Why Pythia-70m copies the first clause's unit

Will Sackfield, Open Interpretability · https://open-interpretability.com/research/copy-first-unit

## Abstract

Pythia-70m answers "A week has seven days and a year has twelve" with " days", and five similar prompts the same way: it copies the first clause's unit. An earlier study verified the route by which it copies. This paper asks whether the right answer is missing from the model, or present and outvoted. We measured the six wrong prompts and three prompts that copying answers right. The model barely knows the facts on their own: with only the second clause, one of the six right answers is its top token. No MLP at the second clause's subject writes the right answer into the top 1% of the vocabulary. Yet at the answer position the right answer reaches rank 2 to 4 by layer 4, where the copied unit leads. Switching off the three verified copy heads, or them and a fourth head, flips the answer on one prompt of six. So the right answer is present but weak, and the copy is spread over more units than the verified route names.

**Status:** Preprint. Work left: #5094 (the coverage tail of the " days" explanation, and three units to revise).

## 1 Introduction

A wrong answer is the case where an explanation of a model should help most. In #4781 we explained one: Pythia-70m answers "A week has seven days and a year has twelve" with " days". An answer experiment VERIFIED the route by which it copies the first clause's unit (`wiki/answers/EleutherAI--pythia-70m/970a91e11d8c/20261118T234629Z.json`). That route says how " days" wins. It doesn't say why " months" loses. The right answer may be missing from the model, or present and outvoted, and the two call for different repairs. This paper asks which it is. It bears on research question 0, how much of a model the registry explains (research questions), because an explanation of a wrong answer is incomplete until it covers the right answer too.

The paper's contributions are these:

1. **The fact is weak on its own.** With only its second clause, the model's top token is the right unit for one prompt of six ("A week has seven" → " days"). The right unit is rank 1 to 15 there. After four unrelated facts it falls to rank 3 to 254. *Limit:* one zero-shot form and one set few-shot context, scored by rank and probability, not the sampled accuracy of Gekhman et al. (2024).
2. **Enrichment at the subject fails.** At the second clause's subject (" year"), no layer's MLP output ranks the right unit above 731 of 50,304 tokens. That is outside the top 1%, the line that Yu et al. (2024) draw. For two of the three controls, the MLPs rank the right unit at 54 and 105. *Limit:* the lens reads each MLP's output alone, with the final norm's statistics of that output.

3. **The right unit is present at the answer position, and outvoted.** Under the lens, the right unit is rank 2 to 4 after layer 4 on five of the six prompts. The copied unit is rank 1 or 2 there. The last layer leaves the order unchanged.
4. **The verified copy heads are not the whole copy.** Switching off L3/H3, L4/H0 and L4/H7 narrows $\text{logit}(\text{right}) - \text{logit}(\text{copied})$ on five prompts. Only " eggs" flips, to " hours". Switching off L4/H1 as well flips no other prompt. On three prompts, the copy heads move the gap no further than 20 random sets of three heads do. *Limit:* switching a head off sets its output to zero, which isn't a mean ablation. And the random sets come from layers 3 and 4 only.
5. **L5/N2014+ is part of the prompt's structure, not of its fact.** It is active on four of the six wrong prompts, and below its activity line on every second clause alone. On the " days" prompt it is the second-largest writer toward " months". On two other prompts it writes toward the copied unit. Its unit experiment VERIFIES it as a time-expression value-slot pole (#5093).

These results don't show what a repair would do. They describe the model as it is, with no change made to it.

## 2 Related work

Geva et al. (2021) read feed-forward layers as key-value memories, whose values promote tokens in the vocabulary. Geva et al. (2023) split the recall of a fact into steps. Early MLPs enrich the subject's state with its attributes, and attention heads at the last position extract one of them. Yu et al. (2024) use that split to sort hallucinations. A fact can fail at enrichment, when no MLP at the subject ranks the answer in the top 1%, or at extraction. Yu et al. (2023) study how a model weighs an answer in its context against one in its memory. They find heads that carry the context's answer, and that scaling them lets the memory's answer win more often. McDougall et al. (2023) describe a copy-suppression head, which lowers a token that earlier layers copied from the context. Gekhman et al. (2024) sort facts by how often a model answers them right on its own, which separates a fact the model doesn't know from one it knows but loses.

## 3 Method

**Model.** EleutherAI/pythia-70m, in float32 on the CPU, at its main revision.

**Prompts.** The 35-prompt shortlist (`docs/analyses/hallucination-shortlist.jsonl`, from `scripts/hallucination_shortlist.py`) keeps six prompts that the model answers wrong with a confident content word. Each copies the first clause's unit, or in one case a noun (Table 1):

| Prompt | Right | Copied | The model's answer and its probability | The right unit's rank |
|---|---|---|---|---|
| A minute has sixty seconds and an hour has sixty | minutes | seconds | seconds, 0.810 | 2 |
| A week has seven days and a year has twelve | months | days | days, 0.726 | 4 |
| A hand has five fingers and a foot has five | toes | fingers | fingers, 0.680 | 2 |
| A dozen has twelve eggs and a year has twelve | months | eggs | eggs, 0.643 | 6 |
| An hour has sixty minutes and a minute has sixty | seconds | minutes | minutes, 0.419 | 2 |
| A year has twelve months and a week has seven | days | months | months, 0.455 | 2 |

**Table 1.** The six wrong prompts, from the shortlist. The right unit's rank is in the model's output at the answer position.

Three prompts of the shortlist are right because the model copies. They are the controls: "A spider has eight legs and a dog has four", "A car has four wheels and a bike has two", and "A cat has four legs and a bird has two".

**The verified route.** In #4781 the model's route to " days" was VERIFIED by an answer experiment (`wiki/answers/EleutherAI--pythia-70m/970a91e11d8c/20261118T234629Z.json`). Layer-0 neurons mark the count words and time units. Early heads line up the two clauses. The copy heads L3/H3, L4/H0 and L4/H7 carry " days" to the answer position.

**Measures.** `scripts/copy_first_unit.py` writes one row per prompt to `results.jsonl`:

- *Single clause.* The second clause alone, such as "A year has twelve", zero-shot. Also the same clause after four facts with no unit of the prompts: "A triangle has three sides. A square has four corners. An octopus has eight arms. A pentagon has five sides." Following Gekhman et al. (2024), a model that answers the clause alone knows the fact without a clause to copy.
- *Enrichment.* Following Geva et al. (2023) and Yu et al. (2024), the subject's state after each layer, and each layer's MLP output, read through the final norm and the unembedding. It records the ranks of the right and the copied tokens. The same reading of the answer position gives contribution 3.
- *Attention.* The share of the answer position's attention that L3/H3, L4/H0, L4/H7 and L4/H1 give to the copied unit and to the subject.
- *Scaling.* Following Yu et al. (2023), the copy heads' outputs scaled by $\alpha \in \{1, 0.5, 0, -0.5\}$, past the first token. Also the copy heads and L4/H1 at 0 together, and 20 seeded sets of three random heads of layers 3 and 4 at 0.
- *Direct writes.* Following Geva et al. (2021), each layer 4 and 5 neuron's activation at the answer position, times its write to $\text{logit}(\text{right}) - \text{logit}(\text{copied})$. The final norm's gain and mean are folded in, and its scale is taken from the real residual stream.

## 4 The fact is weak on its own

Table 2 gives each wrong prompt's right unit, after the full prompt and after its second clause, and Figure 1 draws it.

| Prompt (right, not copied) | Full prompt | Clause alone | Clause after four facts | Top token, clause alone |
|---|---|---|---|---|
| minutes, not seconds | 2 | 2 | 3 | "-" |
| months, not days | 4 | 2 | 80 | " years" |
| toes, not fingers | 2 | 15 | 254 | " points" |
| months, not eggs | 6 | 2 | 80 | " years" |
| seconds, not minutes | 2 | 3 | 14 | "-" |
| days, not months | 2 | 1 | 74 | " days" |

**Table 2.** The right unit's rank. "A year has twelve" alone gives " years" first, with " months" second at probability 0.040.

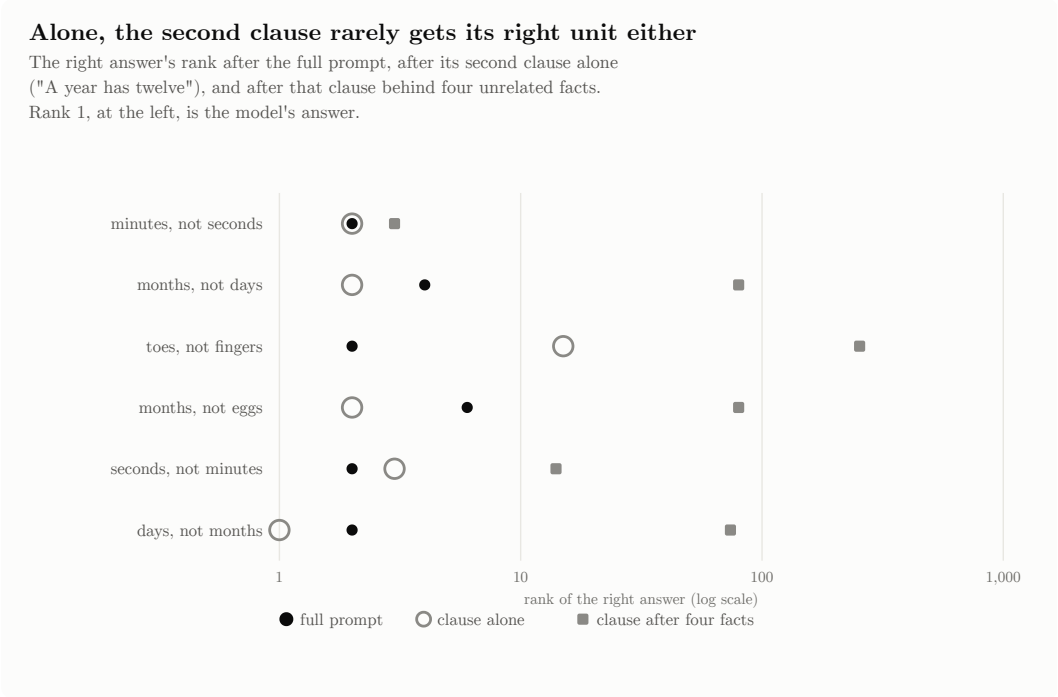


**Figure 1.** The right unit's rank after the full prompt (dot), the clause alone (ring) and the clause after four facts (square).

The model doesn't hold these facts firmly. Its greedy answer to the clause alone is wrong for five of the six, though the right answer is near the top. The few-shot context makes them worse, not better. So a first clause with a unit in it has little to compete with.

## 5 Enrichment at the subject fails

At the second clause's subject, the best rank of the right unit in any layer's MLP output is 1,232, 3,604, 2,513, 4,162, 2,171 and 731 on the six wrong prompts. The top 1% of the vocabulary ends near rank 503. In Yu, Cao et al.'s split, the MLPs fail to enrich the subject with the attribute. For two of the controls, the best ranks are 105 and 54, inside

the top 1%. The third, " wheels", is 7,545. The subject's residual stream is no better on the wrong prompts: its best rank is 273 to 1,145.

## 6 The right unit is present at the answer position, and outvoted

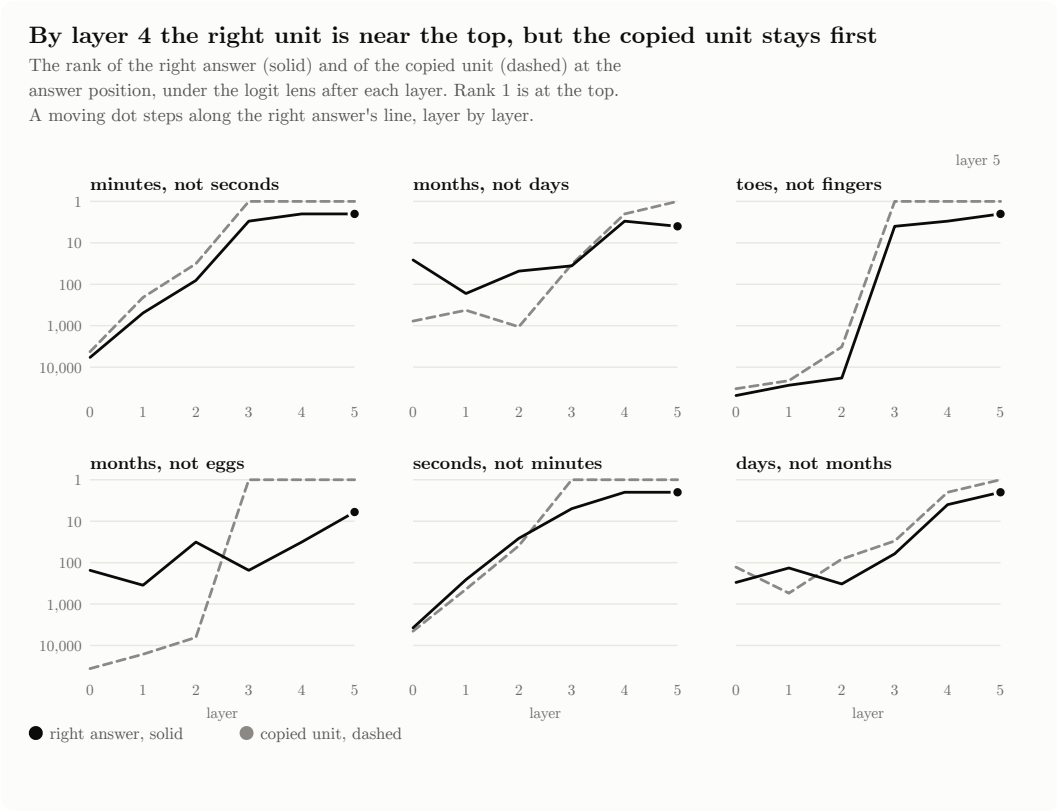


**Figure 2.** The right unit's rank (solid) and the copied unit's (dashed) at the answer position, after each layer. A moving dot steps along the right unit's line.

At the answer position, the right unit rises with depth on every wrong prompt (Figure 2). After layer 4 it is rank 2, 3, 3, 32, 2 and 4, and the copied unit is rank 1 or 2. The last layer leaves the order as it is. The right answer therefore reaches the answer position, but from somewhere other than the subject's MLPs, and the copied unit outvotes it. The " eggs" prompt is the exception, with the right unit at rank 32 after layer 4.

The layer 4 and 5 MLPs add little either way. Over all their neuron poles, the direct writes to $\text{logit}(\text{right}) - \text{logit}(\text{copied})$ sum to between $-2.46$ and $+1.24$ on the six prompts. Writers on each side sum to 16 to 34 logits, so the two sides nearly cancel. On the second clause alone, the sum is between $-0.30$ and $+1.36$.

## 7 The verified copy heads are not the whole copy

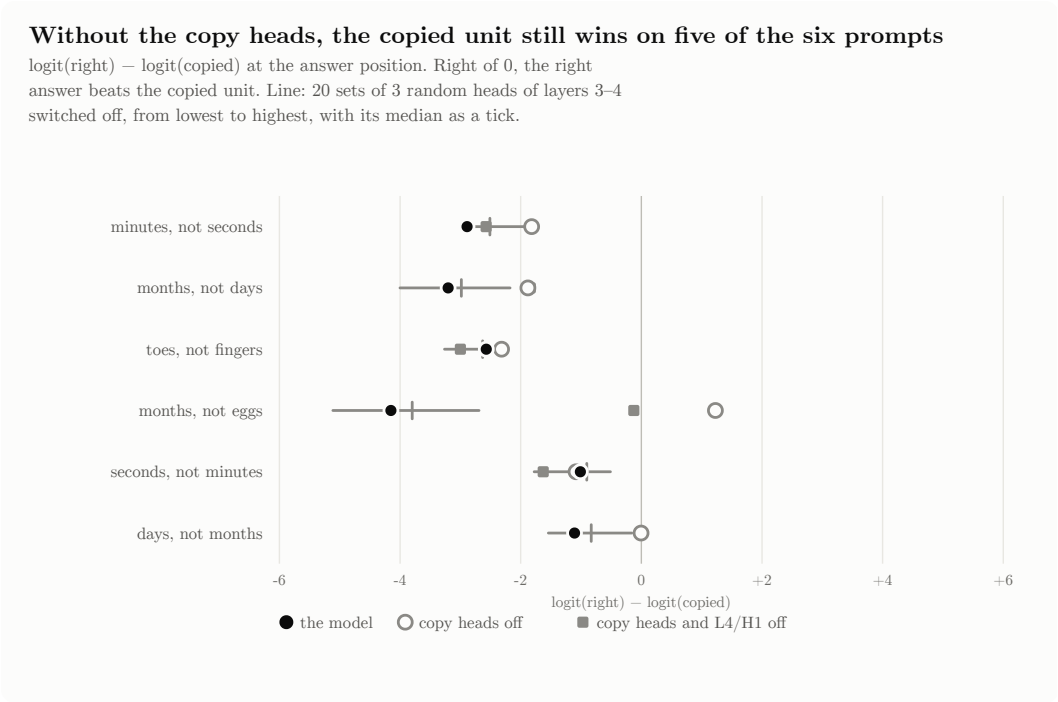


**Figure 3.** logit(right) − logit(copied) for the model, with the three copy heads switched off, and with them and L4/H1 switched off. A line spans 20 random sets of three heads of layers 3 and 4, switched off, with a tick at their median.

| Prompt | The model | Copy heads off | Copy heads and L4/H1 off | Copy heads at $\alpha = -0.5$ | Random heads off: median, range |
|---|---|---|---|---|---|
| minutes, not seconds | −2.89 | −1.82 | −2.57 | −1.25 | −2.51, −2.96 to −1.78 |
| months, not days | −3.20 | −1.88 | −1.84 | −0.51 | −2.98, −4.00 to −2.18 |
| toes, not fingers | −2.57 | −2.32 | −3.00 | −2.23, top " teeth" | −2.63, −3.26 to −2.24 |
| months, not eggs | −4.15 | +1.23, top " hours" | −0.12, top " hours" | +5.62, top " hours" | −3.80, −5.11 to −2.69 |
| seconds, not minutes | −1.01 | −1.08 | −1.63 | −1.88 | −0.91, −1.78 to −0.51 |
| days, not months | −1.11 | −0.00 | −0.01 | +0.51, top " days" | −0.83, −1.54 to −0.16 |

**Table 3.** logit(right) − logit(copied) at the answer position. Below 0, the copied unit wins. The top token is the copied unit unless the table names another.

Switching off the copy heads narrows the gap on five of the six prompts (Figure 3, Table 3). It flips one: " eggs" gives way to " hours", not to " months". On "minutes", "toes" and "seconds" the change is within the range of random heads. Reversing the copy heads' outputs ($\alpha = -0.5$) makes " days" the answer to the reversed prompt. With all four heads off, the copied unit still wins on five prompts.

The attention patterns show why the route's heads are not enough. L4/H7 puts at least 0.87 of its attention on the copied unit on every wrong prompt, and on every control. L4/H1 puts 0.94 or more on it on four wrong prompts. So L4/H1, a candidate copy-suppression head (#5063, McDougall et al., 2023), reads the copied unit. Switching it off doesn't lift the right answer, though. The copy runs through more heads than the verified route lists. The route's claim was about how the model produces " days", not that its heads are the only path.

On the three controls, switching the copy heads off keeps the right answer, and $\alpha = -0.5$ breaks it each time.

## 8 L5/N2014+

L5/N2014+ has the largest single attribution on the " days" prompt in the #4781 explanation. Its maximum on the scan corpus is 6.10, so its activity line is 1.52. At the answer position it is above that line on four of the six wrong prompts: 3.11 on "months, not days" and on "minutes, not seconds", 2.61 on "seconds, not minutes" and 1.97 on "days, not months". It stays below the line on every second clause alone, with 1.15 the highest. Its direct write toward " months" on the " days" prompt is +0.27, the second-largest of any layer 4 or 5 pole. On "seconds, not minutes" and on the reversed prompt it writes toward the copied unit instead ($-0.06$ and $-0.18$). So it responds to the two-clause frame of the prompts, not to the fact. Its unit experiment (#5093) is VERIFIED, at confidence 0.97. Its explanation: it fires on the token that fills the value slot of a time or date expression, after a cue such as "within", "every" or "earlier". It passed 19 of 20 hard negatives and 10 of 10 holdouts. Judges scored it 0.62 for detection and 0.54 for fuzzing. In these prompts the slot is the second clause's number, such as "twelve". The clause alone has the same slot, yet the pole stays below its line there. So here the time reading comes from the first clause's unit, and the pole is silent when that unit isn't a time unit, as with "fingers" and "eggs".

## 9 Discussion

The measures agree on one picture. The facts are weak in this small model: the second clause alone rarely gets its unit (Table 2). The MLPs at the subject don't enrich it. The right unit still reaches the answer position, from elsewhere, and rises to rank 2 to 4 by layer 4 (Figure 2). There the copied unit outvotes it. In Geva et al.'s (2023) terms, enrichment fails, and extraction takes the wrong source. In Yu et al.'s (2023) terms, the context's answer wins over a weak memory, as they found for small models.

The verified route explains how the copy works, but it is not the only path. Switching off its three copy heads, and L4/H1, flips one prompt of six (Table 3). So a repair aimed only at the route's heads would likely fail. A repair that strengthened the fact would face a fact the model barely holds. These are notes for the campaign to repair models, which the maintainer designs. They are not a design for it.

## 10 Limitations

- **One small model.** Pythia-70m has six layers. Larger Pythia models may hold the facts. Yu et al. (2023) found that smaller models take the answer in the context more often (see #4784's synthesis).

- **Six prompts.** The shortlist is small, and its prompts share one template. One of the six copies a noun, not a unit.
- **First-order lens.** The lens reads each state as if the remaining layers did nothing. A rank under it is a hint of what the state holds, not a measure of what the model does with it.
- **Switching off is not mean ablation.** A head's output set to zero is far from its usual value, and may add effects of its own. The random-head control shares that cost.
- **The few-shot context is set.** One context of four facts can't stand for a sampled set of contexts.

## 11 Conclusion

Pythia-70m copies the first clause's unit because the right unit is weak, not because it is missing. The model barely knows these facts, its MLPs don't attach them to the subject, and the copy outvotes what reaches the answer. The copy runs through more heads than the verified route names. These tests would go further:

- Knocking out attention edges (#5059) would show which heads beyond the route carry the copy.
- What L4/H1 does (#5063), now that it is known to read the copied unit. On these prompts it lowers the copied unit, but too weakly to change an answer (`docs/analyses/copy-suppression-l4h1.md`).
- Pythia SAEs (#5065) and a gradient-routed positive control (#5066) would test the copy another way.
- The coverage tail of the " days" explanation, and the three units to revise (#5094).
- The same measures on larger Pythia models would test whether the fact grows strong enough to win.
- Whether the model can tell a wrong copy from a right one (#5074). The route's units can't, but a residual-stream direction after layer 2 does, modestly (`docs/analyses/copy-route-separation.md`).

## References


Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., & Herzig, J. (2024). *Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?* EMNLP 2024. arXiv:2405.05904.

Geva, M., Schuster, R., Berant, J., & Levy, O. (2021). *Transformer Feed-Forward Layers Are Key-Value Memories*. EMNLP 2021. arXiv:2012.14913.

Geva, M., Bastings, J., Filippova, K., & Globerson, A. (2023). *Dissecting Recall of Factual Associations in Auto-Regressive Language Models*. EMNLP 2023. arXiv:2304.14767.

McDougall, C., Conmy, A., Rushing, C., McGrath, T., & Nanda, N. (2023). *Copy Suppression: Comprehensively Understanding an Attention Head*. Preprint. arXiv:2310.04625.

Yu, L., Cao, M., Cheung, J. C. K., & Dong, Y. (2024). *Mechanistic Understanding and Mitigation of Language Model Non-Factual Hallucinations*. Findings of EMNLP 2024. arXiv:2403.18167.

Yu, Q., Merullo, J., & Pavlick, E. (2023). *Characterizing Mechanisms for Factual Recall in Language Models*. EMNLP 2023. ACL Anthology.

# Appendices

## Reproduce

```
uv run python scripts/copy_first_unit.py --out papers/copy-first-unit/results.jsonl
uv run python scripts/paper_figures.py
```

The script reads EleutherAI/pythia-70m twice, once with eager attention for the patterns. It runs in about 10 seconds on a CPU. The random head sets come from seed 0, and `results.jsonl` lists them.