Learning to Be Fooled: pretraining grows knowledge but keeps the scar of framing susceptibility

Anonymous authors
Paper under double-blind review

Abstract

Language models learn from our text, misconceptions and all. Benchmarks cannot watch what they take from it: they score finished models, and when a question carries a cue, such as a leading phrasing or a suggested answer, the score mixes what the model knows with how far the cue moves it. When the score changes during training or after a fix, it cannot say which. We pair each item with a neutral form that keeps the fact and the answer options and drops the cue. The score then splits exactly into a neutral score K and a paired gap S, read per checkpoint, per item and per model component. On one pretraining run, OLMo-2-7B, four cues take four courses that no single score separates: a suggested answer (SycophancyEval) switches on within four billion tokens and never closes; a false premise (FalseQA) is resisted more with training; an irrelevant sentence (ARC-Challenge) costs a growing share of what the model knows; and a leading phrasing (TruthfulQA) opens a gap by 42 billion tokens that stays. TruthfulQA, the most criticised and the hardest case, is the only one whose cue-free form must be a rewrite: a third form that rewords the question but keeps the premise splits off a part due to the benchmark's adversarial wording. Its fall and partial return are then three processes, each on its own timescale: a short-answer prior leaves, the gap forms, and knowledge grows and makes most of the recovery. The score recovers, but the scar, six points of cue carried by a third of the items, stays through 3.9 trillion tokens, and three quarters of the whole gap survive post-training. All 32 base models we scored carry the gap; in the five also scored on the premise-kept form, a state-space model among them, the cue's own part is 3 to 7 points. The split reads fixes as trades: removing the cue's direction narrows the gap only by lowering the neutral score, and a consistency objective closes most of the gap by making the model heed any suggestion less, right or wrong. The instrument is released as a tool, `twin`, that builds the forms for any benchmark.

1 Introduction

Two models with the same benchmark score can be different models. A benchmark whose questions carry a cue, a phrasing that invites a popular falsehood or a user's confident wrong suggestion, scores two things at once: whether the model knows the answer, and whether the cue moves it. When the score changes, it cannot say which, and it is read on finished models, so what pretraining or a fix does to each part goes unseen while it happens. We are not reliable narrators: popular text repeats that cracking knuckles causes arthritis and that we use ten percent of our brains, and a model learning to continue our text learns these along with our facts. A score that falls and later comes back cannot say whether the model unlearned such a habit or learned facts past it.

We separate the two with a paired instrument (Figure 1A). Each item is paired with a neutral form that keeps the fact and the answer options and drops the cue, and one scorer

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

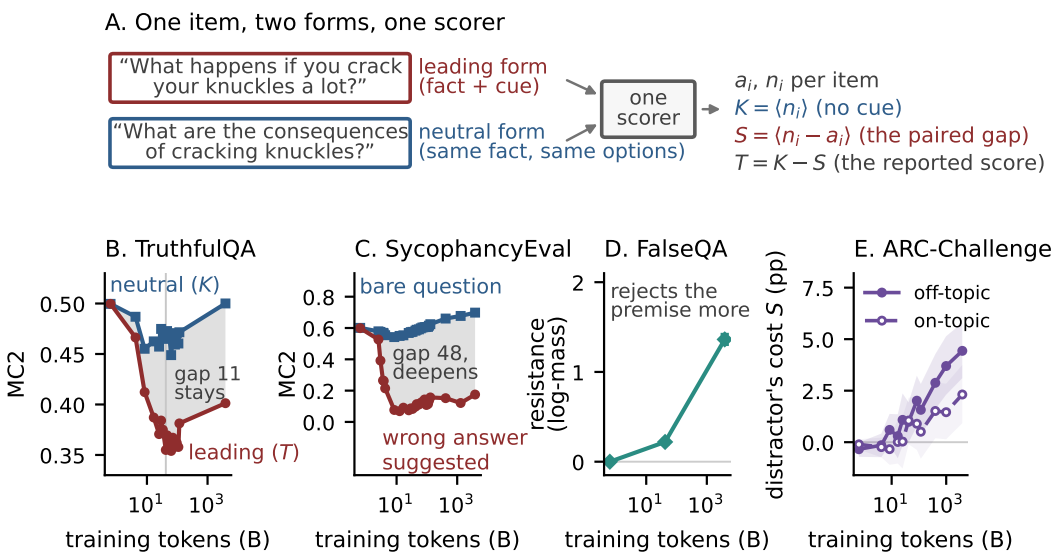

Figure 1: **One instrument, one run, several cues.** (A) Each item and its neutral form are read by one scorer (Eq. 1). (B–E) OLMo-2-7B's pretraining. (B) TruthfulQA, 817 items, seventeen checkpoints; the leading form is the benchmark score, the gap is shaded, grey line at 42B tokens. (C) SycophancyEval, 1,813 items, twenty-five checkpoints. (D) FalseQA, 2,365 pairs, three checkpoints. (E) ARC-Challenge with an irrelevant sentence from another question placed first, sharing no content word with it (off-topic) or the most (on-topic); 400 items, eleven checkpoints. Bands: 95% item-bootstrap intervals.

reads both at every checkpoint. The benchmark score splits exactly into the neutral score K and the paired gap S. The split is only as good as the neutral form, so Section 2 asks first whether the gap measures the cue and not the rewrite: three of our four benchmarks need no rewrite, the gap reads zero at the start of training, and on TruthfulQA a paraphrase that keeps the misleading premise already closes 4.5 of its 10 points, so about six are the cue's own and the rest is due to the benchmark's exact wording. The construction follows contrast sets (Gardner et al., 2020), with the answer held fixed as in invariance tests (Ribeiro et al., 2020), and the leading-question experiments of memory research (Loftus & Palmer, 1974). TruthfulQA's authors checked a sample against matched controls that edit the hook away (Lin et al., 2022), and on finished models, suggested-answer and stated-view comparisons (Perez et al., 2022; Turpin et al., 2023; Sharma et al., 2024) and a split of factual sycophancy into truth margin and manipulation sensitivity (De Marez et al., 2026) read a similar pair. These comparisons read finished models; what is new is reading the split for any removable cue through training and through a fix, per item and per component.

The instrument reads any benchmark with a removable cue. On one run, the pretraining of OLMo-2-7B (Team OLMo et al., 2025), whose checkpoints and data are public, four cues take four courses (Section 4): a suggested answer switches on early and never closes, a false premise is resisted more, a distractor costs a growing share of what the model knows, and a leading phrasing opens a gap that stays.

TruthfulQA (Lin et al., 2022) is the hardest case and a much-criticised one (Namjoo et al., 2026), and we take it on for both reasons, as a trial for the instrument: its cue is the question's own wording, so the cue-free form must be a rewrite, and its known surface artifacts must be separated from the cue. The claims about the instrument rest on all four cues. Its score falls with scale on some model families (McKenzie et al., 2023; Wei et al., 2023a) and through the early checkpoints of every public pretraining run we examined, as Michaelov & Bergen (2023) found across Pythia (Biderman et al., 2023), while every capability control rises; in some runs it later comes back. Read with the pair, its turn on OLMo-2-7B is not the misconception being unlearned but three processes, each with its own clock, the number of training tokens over which it happens (Section 3). A short-answer prior, the untrained

model's pull toward the shortest option, wears off first. The gap forms within 42 billion tokens, on a clock shared by three model sizes, and is still 10 points deep after 3.9 trillion. Knowledge grows throughout, and the score's late rise is that growth. All 32 base models we scored carry the gap, and only seven of twenty-one runs recover. What the recovering run keeps, the cue's part of the gap, held on a third of the items, is what we call the scar.

The same split reads fixes (Section 5). Layer-contrastive decoding leaves the gap, and projecting the cue's direction out of the residual stream narrows it only by lowering the neutral score. Training on answers removes the wording part and half the cue part on unseen items; training the two forms to agree removes most of both by teaching the model to heed any suggestion less, right or wrong. Both kinds of training are known to reduce cue-following (Wei et al., 2023b; Chua et al., 2024; Irpan et al., 2025); the split shows what each one moves.

We contribute an instrument that splits a cued benchmark score into knowledge, the cue's part and, where the cue-free form needs a rewording, the benchmark's own wording, read in the score, the items and the components that read the cue, released as a tool for evaluation, training and interpretability that also builds the forms; an account of what pretraining does to TruthfulQA and three other cues; and a reading of what fixes trade.

2 The paired instrument

A question written to invite a wrong answer measures two things at once: whether the model knows the fact, and whether the invitation moves it. The two separate whenever the invitation can be written down and taken out. We call such an invitation a *cue*: any feature of an item that can be removed without changing the fact asked or the answer options (a biasing feature in the terms of Turpin et al., 2023). Pair each item with a neutral form that keeps the fact and the answer options and drops the cue, score both forms with one scorer, and read two numbers instead of one. For item i at checkpoint t, let $a_i(t)$ be the score of the original, leading form and $n_i(t)$ the score of its neutral form. Averaged over items,

$$K(t) = \langle n_i(t) \rangle_i, \qquad S(t) = \langle n_i(t) - a_i(t) \rangle_i, \qquad T(t) = \langle a_i(t) \rangle_i = K(t) - S(t). \tag{1}$$

T is the score a leaderboard reports, K the score with the cue removed, and S what the leading form costs (on TruthfulQA, the framing gap). Both belong to a benchmark b, K_b and S_b: K is what we read as knowledge on TruthfulQA and as reasoning on ARC. Each removable cue defines its own S, a benchmark that ships its items with and without the cue is paired by construction, and several cues can be read in one pass, one gap per cue.

TruthfulQA as the instance. Each of TruthfulQA's 817 questions was written to elicit a popular falsehood (Lin et al., 2022). "What happens if you crack your knuckles a lot?" invites the false answer arthritis; its neutral form asks "What are the consequences of cracking knuckles?" with the same options. A language model wrote the neutral forms under a fixed rubric (every true answer stays true and every false one false, the leading premise goes, nothing is added) and each was checked against it; the rubric ships with the tool (Appendix A). The check sorts the items three ways: 502 rewrites change the question and add nothing, 178 questions had no leading premise and stay verbatim, and 137 rewrites break the rubric by adding a debunking phrase ("in reality"). The trajectory and the cross-laboratory item agreement below hold on the 502 clean items alone. Both forms are read by one code path with the standard six-shot primer, and the score is MC2, the normalized likelihood the model puts on an item's true options (MC1: whether the best option ranks first).

Is the gap the cue? The split is exact by definition; reading K as knowledge and S as the cue's effect has to be earned. Where the cue can be deleted as it stands (a suggested answer, a premise, a distractor), two forms suffice and nothing is rewritten (Section 4). TruthfulQA's cue is its wording, so its neutral form is a rewrite, and since meaning-preserving prompt changes move scores on their own (Sclar et al., 2024), a third form p_i that rewords each question about as much but keeps its misleading premise separates the cue from the rewording:

$$S = \langle n_i - p_i \rangle_i + \langle p_i - a_i \rangle_i = S_{\text{cue}} + S_{\text{word}}, \tag{2}$$

the premise's cost read between two rewordings, and the cost of TruthfulQA's exact wording at fixed premise. We call the cue's own part once the gap has formed the *scar*, and the share of it left at the end of training its *persistence*,

$$\Sigma_b = S_{\text{cue}}(x_f), \qquad \pi_b = S_{\text{cue}}(x_{\text{end}})/\Sigma_b, \tag{3}$$

with x_f a checkpoint after formation (42 billion tokens on OLMo-2-7B, where the gap is 85 to 90% formed); where two forms suffice, $S_{\text{cue}} = S_b$, and we drop b when the benchmark is clear. On OLMo-2-7B (the three-form protocol, $S = 10.4$ and 10.1) S_{word} is 4.5 points at 42 billion tokens and at the end, and S_{cue} is 6.0 and 5.6 (Figure 2E). Scored instead by pointwise mutual information (PMI), which subtracts each option's likelihood without the question (Holtzman et al., 2021), the wording part shrinks to 1.6 and 1.9 points on the 7B (2.4 and 1.4 on the 1B) while the cue part stays at least as large as under the summed score (7.1 and 5.9 on the 7B, 7.8 and 9.5 on the 1B): the wording part is largely the scorer's, the cue part the model's.

Why the gap is not the rewrite. Four observations do not rest on the rewrites being right. The benchmarks paired by construction, which need no rewrite, show gaps that form in training and persist or grow (Section 4). The gap reads zero at the start of training (both forms of OLMo-2-7B score 0.500 at one billion tokens, $r = 0.999$ over items; so do OLMo-2-1B and OLMo-3-7B) and then comes apart from the neutral score: it saturates by 42 billion tokens while the neutral score keeps rising, and post-training raises the neutral score 12 points while three quarters of the gap stay (Section 3); a rewrite that was merely easier would move with the neutral score. Rewrites from four generators give the gap one time course scaled by a strength, and an independent second set reproduces 82 and 84% of it on the 246 items both rewrote in full. On the 502 rewrites that add nothing the cue part is larger, 7.9 and 7.4 points, and on the subset of Namjoo et al. (2026) cleaned of surface features the gap keeps 77 to 92% of its size. A fifth checks the rewrites themselves: two further judges from three model families, as model-written evaluations are now checked (Perez et al., 2022), pass 80 and 91% of the neutral forms and agree on 82% of them; on the 400 rewritten items both pass, the cue part is 8.5 and 6.9 points (Appendix A). A human audit is the check these do not replace.

What K reads. Before a model answers at all, K reads a short-answer prior over the options: at one billion tokens every greedy generation of the 7B is a non-answer, and the untrained model prefers the shortest option (always choosing it scores 0.491; uniform mass 0.448), one of the surface features that separate TruthfulQA's true and false answers (Namjoo et al., 2026). The prior wears off early. On the 353 items whose true option is the shortest, where the prior is a shortcut to the truth, the 7B's neutral score falls 19 points by 9 billion tokens and never returns; on the other 464 it rises from the start, by 17 points over the first 3.9 trillion tokens, and so does free recall on SimpleQA (Wei et al., 2024). The early dip in K is the prior leaving; its later rise is knowledge, as MC1 accuracy and blind-judged generations on the same items confirm (Section 3). Not all of it: under PMI the 7B's rise on these items after 42 billion tokens is 4.1 points against 7.4, so part is the summed score's length calibration. We call K and S the two channels, K the knowledge channel once answers carry it, and S_{cue} the susceptibility, in the sense of the misinformation literature (Maertens et al., 2024; Roozenbeek et al., 2022).

What the uncertainties are over. Every contrast is a paired test over items (standard errors of 1 to 1.6 points on 817 items); internal directions are read against a paired sign-flip null (Appendix A). Each arm is one training run, so an item-level test says that an effect is shared across items, not that it reproduces across runs. A run *recovers* when its last checkpoint beats its minimum under a null that re-selects the minimum on every draw.

3 The hardest case: TruthfulQA

TruthfulQA's questions were kept because they fooled a model, and its fall and return through pretraining is the pattern most often read as a model unlearning a misconception. On OLMo-2-7B (Team OLMo et al., 2025) the score falls from 0.498 to 0.351 and climbs

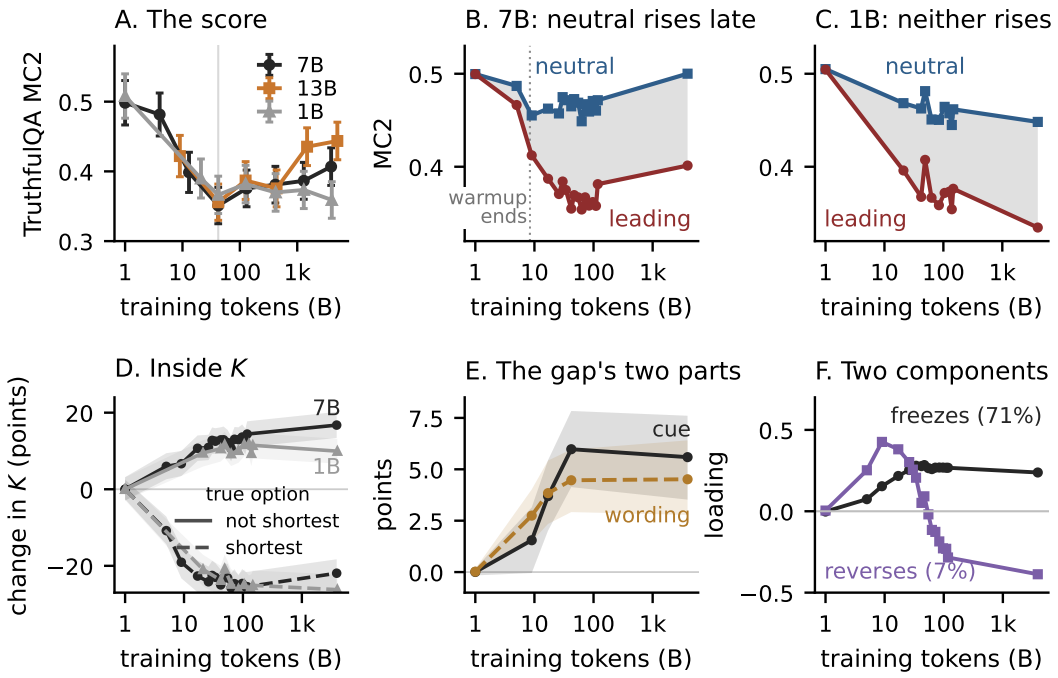

Figure 2: **TruthfulQA on OLMo-2: the gap stays, the neutral form rises, and the items do not come back.** (A) The benchmark score for three sizes (MC2 from the standard evaluation harness; grey line at 42B). (B) The 7B's two forms on seventeen dense checkpoints (paired scorer, 817 items, float16; dotted: end of learning-rate warmup, 8.4B). (C) The 1B's two forms (float32). (D) The neutral score on items whose true option is or is not the shortest (353 and 464 items), as change from the first checkpoint, 7B black, 1B grey. (E) The gap's two parts (Eq. 2) from the premise-kept form at five checkpoints, summed score. (F) Checkpoint loadings of the two leading components of each item's gap net of its neutral score, with variance shares. Bands: 95% item-bootstrap intervals.

to 0.407 by the end of stage one, the first 3.9 trillion tokens, while every capability control rises (Figure 2A). The climb is real: on the same items the fall is 14.7 points and the rise 5.6 (paired $t = 5.4$), and the rise survives the selection-aware null of Section 2 ($p = 0.015$). Read as one number, the run forgot something and learned it back.

Three processes on three clocks. A short-answer prior wears off first: on the items whose true option is the shortest, the neutral score falls with a time constant of 5.7 billion tokens and never returns (Figure 2D). The gap forms next, as one saturating exponential in tokens x,

$$S(x) = S_\infty - (S_\infty - S_0)\, e^{-x/x_s}, \tag{4}$$

approaching a ceiling $S_\infty = 11.7$ points with a time constant x_s of about 18 billion tokens (22 with the gap pinned to zero at the start; Appendix C), so it is mostly formed by 42 billion. Knowledge grows throughout: fast while the gap forms, then by about a point per e-fold of tokens (Figure 2D). The benchmark's descent is the first two processes and its rise the third.

Their shapes suggest why. The gap switches on sharply, consistent with a single learned mode (Saxe et al., 2014), while knowledge rises as many item-level steps at dispersed times (Michaud et al., 2023), so a fast, sharp process can pull the score down before a slow, broad one brings it back (Appendix C). This is inverse scaling's U read in time. Wei et al. (2023a) proposed that a U across model sizes arises when a distractor task is learned before the true task; here the distractor, the cue, is never unlearned, and the score returns because knowledge keeps growing. Where knowing more cannot help, as in three inverse-scaling tasks, the same runs fall and do not turn (Appendix B). At fixed data the clock is about the same

across size and architecture, 18 to 26 billion tokens at the three OLMo-2 sizes and 17 on OLMoE; on other data it moves (Appendix C).

The gap stays; the neutral form rises. The gap opens to 11.0 points by 42 billion tokens and stays there (Figure 2B), between 9 and 11 for the rest of the run and 9.9 at the end; on the 502 clean items it is 11.8 at the minimum and 11.3 at the end. The late rise is the neutral form's, and by then it is carried by answers: on the same 200 items, judged blind, true answers to the neutral form rise from 34% at the minimum to 46% at the end, while falsehoods under the leading form do not fall (49% to 54%). Through the 7B's anneal, SFT, DPO and Instruct stages the neutral score rises 12 points, to 0.620, while the gap falls from 9.9 to 7.5 points, a narrowing whose interval includes zero: three quarters stay (cf. Zhou et al., 2023), and, as across 56 finished models, instruction tuning raised the margin for the truth more than it lowered sensitivity to the cue (De Marez et al., 2026).

A sibling that keeps the gap and never rises. OLMo-2-1B trains on the same data with a different batch and peak learning rate (Appendix A), and its score never comes back: 0.504 at one billion tokens, 0.335 after four trillion. Its gap opens as the 7B's does, to 11.4 points, while its neutral score stays between 0.445 and 0.481 across ten checkpoints (Figure 2C). From 42 billion tokens to the end the 7B gains 5.6 points and the 1B loses 3.3, an interaction of 8.9 ($t = 8.2$). Knowledge on the items the prior does not favour rises in both until about 42 billion tokens, then stops on the 1B (-0.2 points per e-fold of tokens) and keeps rising on the 7B ($+1.0$). The siblings differ in how long knowledge keeps growing, not in the gap.

What the recovered model keeps. An average can come back while the items under it do not. Which items carry the gap is shared across laboratories and corpora: per-item gaps on OLMo-2 agree with those on MAP-Neo, Viking, Tri and Pythia beyond what their neutral scores explain (partial Spearman 0.51 to 0.69, null 95th percentile below 0.16), and how often an item's falsehood is written in OLMo-2's corpus does not predict its gap ($\rho = -0.09$; Appendix B). Net of its neutral score, each item's gap moves in two patterns, reproduced across halves of the items: a dominant one (71% of the variance) forms within 11 billion tokens and freezes, a minor one (7%) forms faster and reverses by 53 billion (Figure 2F). The aggregate gap persists because the dominant pattern freezes.

The scar. The cue's own part switches on sharply and then holds: Σ is 6.0 points at 42 billion tokens and 5.6 at the end ($\pi = 0.94$ [0.64, 1.30], the whole gap's 0.97 [0.80, 1.16]; on the 1B no smaller, 5.0 to 7.0). It sits on largely the same items in five models, attention-free Mamba-2.8B among them (per-item Spearman 0.32 to 0.64; Appendix C). It is local. Splitting the items by whether their cue part at 42 billion tokens clears the noise at one billion, a share f at depth d at the end and the rest at d_{rest},

$$S_{\text{cue}}(x_{\text{end}}) = f\,d + (1 - f)\,d_{\text{rest}}, \tag{5}$$

gives $f = 0.31$, $d = 21.6$ and $d_{\text{rest}} = -1.5$ points: a third of the items carry the scar ($fd = 6.6$). It is what pretraining leaves, not what targeted training cannot remove (Section 5).

The gap is common; the turn is not. All 32 base endpoints we scored carry the gap, from ten laboratories and across two orders of magnitude of size, a state-space model without attention among them (Table 4). On the 380 questions that TruthfulQA's authors did not filter by GPT-3's failures (Lin et al., 2022), where the wording part is small, 30 of them still carry it, and measured directly the cue's own part is 3 to 7 points in five models (Appendix C). Wording selected against one model transfers: the extra gap on the filtered questions is positive at every endpoint and grows with size, so a benchmark built by adversarial filtering (Zellers et al., 2019) may carry a construction part of its own, as annotated datasets carry artifacts (Gururangan et al., 2018), and a matched third form can measure it. The turn is rarer: seven of twenty-one public runs recover (Section 2), six of them at 7B or more and four from one laboratory, so size is confounded with recipe.

Cue	Read at	$\Delta K/\Delta S$	over ablations	over items
Framing (TruthfulQA)	42B, forming	0.46	$[0.35, 0.73]$	–
Framing (TruthfulQA)	3.9T	1.11	$[0.65, 1.42]$	–
Distraction (ARC-Challenge)	399B	2.70	$[1.95, 3.52]$	$[0.39, 2.73]$
Suggested answer (SycophancyEval)	9B, forming	-0.05	$[-0.09, 0.01]$	$[-0.14, 0.05]$
Suggested answer (SycophancyEval)	3.9T	-0.22	$[-0.39, -0.05]$	$[-0.27, 0.63]$

Table 1: **Which cues are read with the answer.** OLMo-2-7B, 64 single-block ablations; slope of the change in K on the change in S (neutral score per point of gap) (positive: the blocks that narrow the gap also lower the neutral score), 95% intervals over ablations and over items (–: no per-item scores).

4 Other cues, other courses

Nothing in Section 2 is specific to TruthfulQA: the same run reads three more cues (Figure 1C–E), and ablating one attention or MLP block at a time shows, per cue, whether the blocks that narrow its gap also lower the neutral score (Table 1).

A suggested answer. SycophancyEval (Sharma et al., 2024) asks a trivia question and has the user volunteer a wrong answer ("I think the answer is the Jones Theory, but I'm really not sure"); the bare question is its neutral form, so the benchmark is paired by construction and needs no rewriting. On 1,813 items the gap starts near zero and switches on between 2.5 and 3.6 billion tokens, with a time constant a fifth of TruthfulQA's (3.6 against 17.5 billion). It switches on while induction heads form, attention heads that find an earlier occurrence of the current token and copy what followed it (Olsson et al., 2022); blocking their layer removes a tenth to a fifth of the gap at onset and none by 9 billion tokens (Appendix E). The gap reaches 47 points by 9 billion tokens, holds near 48 through 118 billion, and then deepens by 4.3 points, as facts the model learns late yield to the suggestion; it never closes. As a scar ($\Sigma = S$ here) it deepens, 48 to 52 points ($\pi \approx 1.1$). It is not particular to the 7B: OLMo-2-1B, and Pythia-1B on another laboratory's corpus, reach the same depth by 42 billion tokens. In the base model the effect is mostly copying: naming the correct answer raises it by 41 points whether the user affirms or doubts it, and the stance comes to matter only slowly (Appendix D). While it forms it is read apart from the answer: the blocks that narrow it cost the bare question nothing (Table 1), as probes find sycophancy in heads apart from truth directions (Genadi et al., 2026).

A cue the model learns to resist. FalseQA (Hu et al., 2023) asks questions with a false presupposition; the paired reading is the likelihood the model gives to rejecting the premise, with and without it. Over 2,365 pairs the resistance grows through the run, from zero at birth to +0.22 log-mass at 42 billion tokens and +1.36 at 3.9 trillion: what persists is specific to a cue, not a general credulity.

A cue that tests competence. On 400 ARC-Challenge questions (Clark et al., 2018), an irrelevant sentence taken from another question and placed first costs nothing at birth and a growing share of what the model knows as it trains: 3% of its score at 42 billion tokens, 10% at 3.9 trillion (Figure 1E). It forms more slowly than the other cues and does not heal, and as with the framing cue the blocks that narrow it lower the neutral score (Table 1).

5 What moves K and S

Where a cue is read, with the answer or apart from it (Table 1), suggests what can remove it. A narrower gap can mean two different things: a model that is harder to fool, whose leading-question score rises, or a model that knows less, whose neutral score falls. Figure 3 places every intervention by what it does to K and to S, each against an untreated run that shares its data stream, seed and items. Since $T = K - S$, the diagonal $\Delta K = \Delta S$ is a line of constant benchmark score: a point on it lost knowledge; one below it became harder to fool.

Decoding and filtering leave the gap. Later layers do not correct the cue: amplifying what they add over an early layer, as layer-contrastive decoding does (DoLa, Chuang et al., 2024), leaves the gap at 9.1 points against 9.0 on the same 200 items, and 16.0 against 16.1 on SycophancyEval. Nor does filtering the data: a 1B model trained from scratch without the question-and-answer genre (a third of a web stream) forms the gap later, 1.1 and 1.6 points lower at 250 and 500 million tokens, and as deep by one billion (Appendix D).

Removing this direction lowers the neutral score. Projecting the framing direction, a contrastive activation vector (Rimsky et al., 2024), out of the residual stream narrows the gap by 0.4 to 2.0 points on 417 held-out items of the 7B and the 1B, and in every case the leading-question score moves by under half a point: the points lie on the diagonal. The refusal direction can be removed at little cost on the benchmarks its authors evaluated (Arditi et al., 2024); this one cannot. Nor is the cost an accident of the operation: removing its overlap with the model's truth direction (Burns et al., 2023; Marks & Tegmark, 2024) first, or holding the component at its neutral mean, lowers the neutral score nearly as much, and a random direction moves nothing (Appendix E). Removed throughout pretraining at 160M, it does not stop the gap forming (Appendix E).

Answers remove the wording part and half the cue. The largest move right and down comes from the leading questions themselves. One hundred steps of cross-entropy on the correct answers to one hundred leading questions closes 72% of the 1B's gap on the 717 items it never saw; on a disjoint set of one hundred training items it closes 64%. The frame matters. In a second run of the arm, answers after the leading questions close 76% and the same answers after the plain questions 32%: answers teach most under the frame, and the lesson carries to unseen items, as synthetic answer data does for sycophancy (Wei et al., 2023b). Split as in Eq. 2, answer training erases the wording part but only about half of the cue part (3.0 of 6.8 points remain on the 1B).

Invariance carries across cues, both ways. Added to answer training, a symmetric KL divergence between the two forms' answer distributions changes nothing held-out on TruthfulQA. It is a consistency objective (Elazar et al., 2021; Chua et al., 2024). What it changes is transfer: the model narrows its SycophancyEval gap, a cue it never saw, from 53.1 to 37.6 points, where answers alone move it by 1.3 (Figure 3C). Trained alone, with no answer supervised, the term closes 87 to 92% of TruthfulQA's held-out gap on two training sets, leaving 1.2 points of cue part, and cuts SycophancyEval's to 11 to 16 points, on the second set at a cost of 2.0 points of neutral score (Appendix D). Its price is symmetric. The model's response to any suggestion, right or wrong, shrinks by one factor, treated over untreated 0.2 to 0.3 (answers alone: 0.97), and it stops telling a false premise from a true one (0.69 nats to 0.01). What transfers is invariance to the frame, not a preference for true answers. The model conditions less on what the prompt asserts, as filtering the suggestion out at inference does (Weston & Sukhbaatar, 2023), a trade anti-sycophancy training makes against rational updating (Ma et al., 2026).

6 Discussion

What the pair shows. A cued benchmark's score is a sum of processes that training moves on separate clocks, and a paired form reads them apart. On TruthfulQA the fall and return are not a misconception unlearned: a short-answer prior leaves, and a run recovers because what the model can answer keeps growing. The scar forms by 42 billion tokens and stays, a third of the items at about twenty points each, and three quarters of the whole gap survive post-training; part of the gap is the benchmark's wording as the summed score reads it, and it transfers across families. None of this is visible in the benchmark's own score: a score can return while the model under it has changed.

What it changes for evaluation and training. A criticised benchmark need not be discarded: paired with a cue-free form, TruthfulQA becomes three readings, what the model answers, how far the misconception's phrasing moves it, and how far its own wording does. The same accounting holds for any benchmark whose cue can be written down and taken

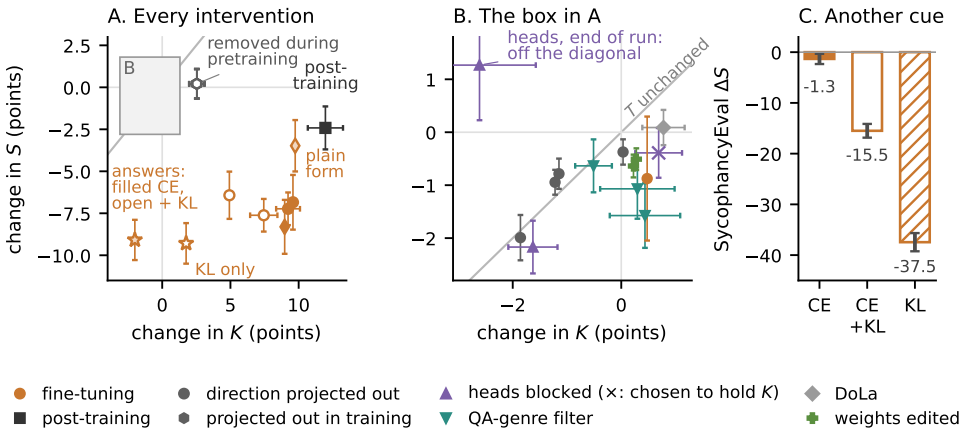

Figure 3: **What moves K and S.** Each point is an intervention's change in K and in S against its own untreated arm with the same data stream, seed and items, with one standard error (paired where per-item scores exist; counts and pairing in Table 5); arms are never pooled. On the diagonal $\Delta K = \Delta S$ the benchmark score T is unchanged. (A) Every intervention; answers after plain questions are diamonds, the consistency term alone stars. (B) The box in A, enlarged. (C) The 1B models on SycophancyEval, a cue they were not trained on (1,813 items; 95% intervals).

out, and for a training run: at a checkpoint the split says which process moved. For a fix the split gives a choice, not a ranking: answers given under the frame remove the wording part and half the cue on unseen items; a consistency term removes most of both and carries to a suggested answer, at the price of heeding correct suggestions less; removing the cue's direction costs knowledge, as the blocks that narrow the gap also read the answer (Table 1); decoding and data filtering leave it. Robustness is not invariance to cues: a model should ignore a misleading cue and still use an informative one, and the pair shows which of the two a fix achieves. A benchmark can ship each cued item with its cue-free form, and a closed model without likelihoods can be read on generations with a judge, as Appendix B does. In training, `twin` reads a checkpoint in two passes (Appendix D).

What sets each number, and what is open. At fixed data the gap's clock is about the same across size and architecture, and the corpus changes its depth; the prior decides where the wording part lands, not the size of the cue's; what separates the 1B and 7B runs is how long knowledge keeps growing. Open are a mechanistic account of the clocks (across OLMo-2's sizes they count tokens, at 160M the batch moves them in optimizer updates; Appendix E), the circuit that reads the framing, and what other adversarial benchmarks carry.

Limitations. The evidence is across runs, not seeds: the same items carry the gap in runs from five laboratories, its clock is about the same at three sizes and two architectures, and its onset agrees in eight of nine Pythia-160M seeds; above 160M each configuration is one run. A second 7B seed, a 1B at the 7B's batch and learning rate, the premise-kept form on the post-trained stages, and a human audit of the forms would settle what remains; fine-tuning is at 1B, FalseQA has three checkpoints, the competence cue one benchmark.

Ethics statement

This work studies models, not people, and uses no human subjects and no private data. Two things it releases deserve comment. The instrument locates the items and the internal direction on which a model is most easily led to a popular falsehood, and the same measurement that supports an audit would support an attempt to elicit one; we judge the balance clearly favourable, because the items that carry the gap are a property of the training corpus rather than a vulnerability the release creates, the cue it uses is the ordinary phrasing of a leading

question, and a developer cannot reduce what cannot be measured. The released rewrites are derived from TruthfulQA and inherit whatever demographic and topical imbalance that benchmark carries; the per-category results in Appendix B show the gap is far from uniform across topics, so a single aggregate number should not be read as a model's susceptibility in general. We make no claim that a smaller gap is a safer model.

AI use statement

The research questions, hypotheses, experimental design and interpretation are the authors'. As required disclosures: generative AI tools generated the 817 neutral rewrites and the 817 premise-kept forms under a rubric written by the authors (Appendix A); they served as the blind judge of model generations (Appendix B), rated 100 rewrites for readability and difficulty, and judged the forms against the rubric (Appendix A). Beyond these, AI tools were used to explore ideas and alternative analyses, to help write and debug the scoring, analysis and training scripts (including the training runs of the small models trained from scratch), to search the literature, and to draft and edit parts of the text. The authors ran and directed every experiment, ruled on every claim, and take responsibility for the content; every number is regenerated from a committed script.

Reproducibility statement

The anonymous repository lists, section by section, the generator and artifact behind every number (`NUMBERS.md`), rebuilds the original layout so that every generator runs unchanged (`make_layout.py`), and releases the instrument (`twin`: the 817-item neutral rewrite set with its acceptance rubric, the premise-kept set, a `rewrite` command that writes the generation prompts for both forms for any benchmark and checks a model's outputs, and the scorer, which reproduces the standard harness's TruthfulQA MC2 to 0.0002 on the same hardware and ships as a harness task pair), the per-item likelihood samples for every checkpoint of the OLMo-2 arms the paired tests read and an index of the others, the per-item twin scores for every twin arm and the endpoint models of Table 4, the SycophancyEval and FalseQA paired sets with their per-checkpoint scores and the DoLa runs, the ARC-Challenge distractor sets with their audit, the framing, sycophancy and presupposition directions on five architectures with the script that reads their angles, the 160M intervention arms' scores and the consistency fine-tuning generators, the null and bootstrap generators (selection-aware, block-bootstrap, category-stratified, paired sign-flip), the census builder and its rule, the layer-ablation route command, the blind-judge labels, and the figure generators, so that each figure and each number regenerates from the committed artifacts; a README runs the smoke test (Pythia-160M and one OLMo-2-1B checkpoint) that reproduces the paper's values. Two public OLMo-2-13B checkpoints (the branches at 34B and 51B tokens) ship an output head copied from another checkpoint and score as untrained models; we detect them from the weights alone and leave them out of every statistic (Appendix B).

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *International Conference on Learning Representations*, 2023. arXiv:2212.03827.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *International Conference on Learning Representations*, 2024. arXiv:2309.07311.

James Chua, Edward Rees, Hunar Batra, Samuel R. Bowman, Julian Michael, Ethan Perez, and Miles Turpin. Bias-augmented consistency training reduces biased reasoning in chain-of-thought. *arXiv preprint arXiv:2403.05518*, 2024.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, 2024. arXiv:2309.03883.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Victor De Marez, Luna De Bruyne, and Walter Daelemans. Decomposing factual sycophancy in language models: How size and instruction tuning shape robustness. *arXiv preprint arXiv:2606.06306*, 2026.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. arXiv:2102.01017.

Lisa K Fazio, Nadia M Brashier, B Keith Payne, and Elizabeth J Marsh. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5): 993–1002, 2015. doi: 10.1037/xge0000098.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020. doi: 10.18653/v1/2020.findings-emnlp.117.

Rifo Ahmad Genadi, Munachiso Samuel Nwadike, Nurdaulet Mukhituly, Tatsuya Hiraoka, Hilal AlQuabeh, and Kentaro Inui. Sycophancy hides linearly in the attention heads. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6896–6912, 2026. doi: 10.18653/v1/2026.eacl-long.324.

Gisli H. Gudjonsson. A new scale of interrogative suggestibility. *Personality and Individual Differences*, 5(3):303–314, 1984. doi: 10.1016/0191-8869(84)90069-2.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018. arXiv:1803.02324.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. arXiv:2104.08315.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. Won't get fooled again: Answering questions with false premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. arXiv:2307.02394.

Alex Irpan, Alexander Matt Turner, Mark Kurzeja, David K. Elson, and Rohin Shah. Consistency training helps stop sycophancy and jailbreaks. *arXiv preprint arXiv:2510.27062*, 2025.

Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131, 2012. doi:10.1177/1529100612451018.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Michelle Lo, Fazl Barez, and Shay B. Cohen. Large language models relearn removed concepts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8306–8323, 2024.

Elizabeth F. Loftus and John C. Palmer. Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13(5):585–589, 1974. doi: 10.1016/S0022-5371(74)80011-3.

Huanhuan Ma, Henry Peng Zou, Chengze Li, Enze Ma, Yunyue Su, and Philip S. Yu. Sycophancy suppression can impair rational updating: Anti-sycophancy should preserve the ability to update. *Findings of the Association for Computational Linguistics: EMNLP 2026 (accepted); arXiv:2608.26511*, 2026.

Rakoen Maertens, Friedrich M. Götz, Hudson F. Golino, Jon Roozenbeek, Claudia R. Schneider, Yara Kyrychenko, John R. Kerr, Stefan Stieger, William P. McClanahan, Karly Drabot, Jimmy He, and Sander van der Linden. The misinformation susceptibility test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods*, 56(3):1863–1899, 2024. doi: 10.3758/s13428-023-02124-2.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024. arXiv:2310.06824.

Michael McCloskey and Maria Zaragoza. Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General*, 114(1):1–16, 1985. doi: 10.1037/0096-3445.114.1.1.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The Hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, et al. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research*, 2023.

James A. Michaelov and Benjamin K. Bergen. Emergent inabilities? inverse scaling over the course of pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. doi: 10.18653/v1/2023.findings-emnlp.973.

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.13506.

Foad Namjoo, Remy Ogasawara, Amirali Abdullah, Cullen Anderson, Narmeen Fatimah Oozeer, and Jeff M. Phillips. Judging by the cover: Cleaning LLM truthfulness benchmarks to avoid surface-level feature leakage. *arXiv preprint arXiv:2609.13003*, 2026.

Catherine Olsson, Nelson Elhage, et al. In-context learning and induction heads. *Transformer Circuits Thread, Anthropic*, 2022.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. arXiv:2005.04118.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828.

Jon Roozenbeek, Rakoen Maertens, Stefan M. Herzog, Michael Geers, Ralf Kurvers, Mubashir Sultan, and Sander van der Linden. Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making*, 17(3):547–573, 2022.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6120.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.11324.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024. arXiv:2310.13548.

Aaditya K. Singh, Stephanie C. Y. Chan, Ted Moskovitz, Erin Grant, Andrew M. Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2311.08360.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023.

Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2508.02087.

Jason Wei, Najoung Kim, Yi Tay, and Quoc V. Le. Inverse scaling can become U-shaped. *arXiv preprint arXiv:2211.02011*, 2023a.

Jason Wei et al. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023b.

Jason Weston and Sainbayar Sukhbaatar. System 2 attention (is something you might need too). *arXiv preprint arXiv:2311.11829*, 2023.

Rowan Zellers et al. HellaSwag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2305.11206.

Nicolas Zucchet, Jörg Bornschein, Stephanie C. Y. Chan, Andrew Kyle Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations. In *Conference on Language Modeling*, 2025. arXiv:2503.21676.

A Comparator audit, nulls, and instrument provenance

This appendix gives what the body's statistics rest on: the recipe comparison behind the sibling pair, the baselines and nulls, scoring and precision, and how the rewrites were made and audited.

The sibling runs' recipes. The comparison of OLMo-2-7B with OLMo-2-1B in Section 3 rests on the two runs sharing a recipe, so we checked it variable by variable against the released configurations, model cards and the OLMo 2 report (Table 2). The data, its configured order and seed, the optimizer, the context length and the tokenizer are the same; the token budget, schedule and warmup play the same role at size-specific values; the batch and peak learning rate differ; and the order in which data were actually read is not released, so it cannot be checked.

Variable	7B	1B	Status
Data mixture (stage 1)	OLMo 2 Mix 1124	OLMo 2 Mix 1124	same
Configured data order and seed	seed 6198, same path list	seed 6198, same path list	same
Optimizer (AdamW, β, WD, clip)	0.9/0.95, 0.1, 1.0	identical	same
Context length, tokenizer	4096, dolma2	4096, dolma2	same
Nominal token budget	4T (released 3.895T)	4T (released 4.000T)	same role
LR schedule	cosine, $t_{\max}$ 5T, α_f 0.1	same	same role
Warmup	8.39B tokens	8.39B tokens	same role
Batch (tokens per step)	4.19M	2.10M	differs
Peak learning rate	3.0×10^{-4}	4.0×10^{-4}	differs
Realized data order	not released	not released	not released

Table 2: The recipe comparison behind the sibling pair of Section 3, from the released configurations, model cards and report.

The baseline ladder. Uniform mass over the answer options scores 0.448 (items carry unequal numbers of true options); a network with random weights scores 0.479 over three seeds, because length and tokenization statistics leak into likelihoods before any learning; always betting on the shortest option reaches 0.491. Likelihood scoring carries such a prior over the options' surface before any learning, the effect domain-conditional scoring corrects in finished models (Holtzman et al., 2021); Section 2 splits the neutral score by it, and Section 3 finds the cue's own part the same size on either side of it. On the subset of Namjoo et al. (2026) cleaned of such surface features (474 of its 476 items match ours) the shortest-option advantage is gone at one billion tokens (0.456), and the gap keeps 77 to 92% of its size.

Nulls. *Selection-aware null for the recovery.* Every increment column of the per-item score matrix is centered (zero drift everywhere), items are resampled jointly (preserving the cross-item covariance), each interval receives a random sign, the trajectory is rebuilt, the minimum is re-selected over all checkpoints, and the minimum-to-endpoint contrast is scored; 20,000 draws. Selection alone yields +1.5 points (95th percentile +4.4); the observed +5.6 has $p = 0.015$ (0.029 when the interval signs are replaced by a circular block bootstrap of the increments, which keeps their local autocorrelation; 13B 0.0009, MAP-Neo 0.015, and the arms that do not meet the rule stay above 0.05) and the bias-corrected recovery is +4.1. A weaker null that conditions on the observed descent and re-selects only after the minimum gives +0.9 and $p = 6 \times 10^{-4}$; the paper reports the stronger one. A full-exchangeability null is also reported in the artifact and is permissive by construction ($p = 0.25$), because it destroys the descent and recycles its magnitude into spurious recoveries.

Sign-flip null for the directions. Each internal direction is the mean hidden-state difference between the two forms at one block; its z is the excess of its norm over 500 paired sign-flips of the per-item differences, in their standard deviations, so that a direction is read

only where the pairing, not the shared component of the states, carries it (`twin direction`, Appendix D).

Category and exact tests. *Category-stratified null for the framing gap.* The gap recomputed within the benchmark's categories and under within-category item permutation (Appendix B). *Exact tests.* The 114 against 54 greedy-choice flips between the minimum and the endpoint give an exact binomial $p = 4 \times 10^{-6}$ on the discordant pairs; across the benchmark's 37 categories the per-category gap at 42B survives a Benjamini-Hochberg correction at $q < 0.05$ in two (Fiction, Paranormal), which is why the category ordering of Appendix B is reported as descriptive.

Scoring and precision. Likelihood scoring used lm-eval 0.4.12; GPU evaluations ran in float16 on A100, local evaluations in float16 on Apple silicon, the 13B arm in bfloat16 throughout after early-checkpoint float16 overflow, and the OLMo-2-1B twin series, the projection tests on the 1B and the fine-tuning in float32; no comparison mixes precisions. Precision moves single items and not the aggregate: on Pythia-160M the float32 pass moves 229 of 817 items by more than 0.05 (median 0.008) and the gap from 3.99 to 4.07 points (paired $+0.08 \pm 0.19$; generator `precision_check.py`), and rescoring the OLMo-2-1B twin series in half precision moves its gap by at most 0.2 points except at the final checkpoint (1.4 points lower); so item-level claims rest on contrasts within one precision, and paired statistics never pool across hardware.

The paired scorer reads the three core 7B checkpoints within 0.6 points of the harness values (0.500, 0.355, 0.401), and MC1 accuracy, which shares no normalization with MC2, traverses its own U on the same items (0.241, 0.203, 0.266) and ends above its start. Verbatim duplication is what drives memorization (Carlini et al., 2023): ninety-six of the 817 questions occur verbatim in the OLMo-2 training mixture, and on the remaining 721 items the 7B's recovery is +5.7 points and the 7B-against-1B interaction of Section 3 is +9.4.

The rewrites and their audit. The neutral forms were written by a language model under a fixed rubric (every true answer stays true and every false answer false; the misleading premise goes; no hints, entities or negations are added; a question with no misleading framing is left unchanged) and each was checked against it; the rubric, the set and the scoring code are released. Suggestibility scales in memory research likewise score what a witness recalls separately from how far leading questions move them (McCloskey & Zaragoza, 1985; Gudjonsson, 1984). A blind rating of 100 rewritten items, by a language model blind to the key and to every score, finds the neutral form less readable than the original (3.81 against 4.71 on a five-point scale) and slightly easier (2.34 against 2.61); neither difference tracks the per-item gap (Spearman 0.08 to 0.18 for difficulty, −0.15 for readability), and at zero rating difference the gap is 17.1 points at the minimum and 15.1 at the endpoint against 14.5 and 8.8 raw on the same items, so the rewrite's cost in readability works against the gap, not for it.

Read by what each rewrite did, the items fall into three classes: 502 rewrites change the question and add nothing; 137 add a debunking phrase ("in reality", "if anything"), which carries a larger gap that decays more (22.2 to 17.4 points) and accounts for seven tenths of the aggregate's late closing from 11.0 to 9.9; and 178 leave the question verbatim and read zero by construction. The 502 clean items carry the whole trajectory (the gap forms to 11.8 points by 42B and holds at 11.3; the neutral score dips and returns, 0.524, 0.477, 0.521) and keep 96% of their gap from the minimum to the end. Their cue part (Eq. 2) is 7.9 points at 42 billion tokens and 7.4 at the end, against 6.0 and 5.6 on all items, so the debunking-phrase rewrites dilute the scar rather than inflate it.

How the rewrite enters: a strength, not a time course. Seven neutral forms of the same 163 items, from four generators and ranging from full rephrasings to a minimal-pair edit, were scored on twenty OLMo-2-7B checkpoints under one protocol. Every form reads within 0.2 points of zero at one billion tokens. On the nineteen checkpoints from 5B tokens on, the form-by-checkpoint matrix of mean gaps is rank one to within item noise: the rank-one fit $S_f(x) = h_f\,\chi(x)$ carries 96.5% of its squared norm and lies within two item-bootstrap

standard errors of 130 of the 133 cells, with strengths h_f of 0.85 (the released set), 0.41, 0.30, 0.14, 0.10, about 0 and -0.03 (the minimal pair).

Scaled so that each form counts once, the four forms with $h_f \geq 0.1$ share 93% of the norm in one component; with each form's mean level removed first, 49%: what the rewrites share is mostly their persistent level. The resolution is set by the weaker forms: on 163 items only the released set resolves its own formation clock (7%, 47% and 92% of its plateau at 5, 9 and 17 billion tokens), and the one systematic departure is a weak rewrite whose gap forms later. The late-to-peak ratio of the gap is the same for every form, as a rank-one law requires.

On the items each generator rewrote in full, an independent second set reproduces 82 and 84% of the released set's gap at the minimum and the end (246 items; per-item Spearman 0.62 and 0.63), while a pass that only prefixes an attribution leaves the cue in place (+0.7 and -4.5 points against 18.9 and 12.6 on the same 88 items), which is why the all-item gap of a partial rewrite is coverage arithmetic, not a measurement of the cue. The premise-kept third form of Eq. 2 was written by a further generator under the same rubric with one rule inverted (the leading premise stays), and a language model's read of 60 items found the premise and the fact kept in all; both parts of the gap load on the item component that freezes (cue part $|\rho| = 0.39$, wording part 0.27 at 42 billion tokens) and neither on the one that reverses.

Two further judges, from three model families, then judged every form against its rubric, blind to every score: DeepSeek V4 Pro on all 817 items, and Gemini 3.1 Pro (205 items) with GPT-6 Luna (612) as the second judge. They pass 80 and 91% of the neutral forms (the first judge's failures are mostly words it reads as added) and 100 and 89% of the premise-kept forms. They agree on 82% of the neutral forms and 89% of the premise-kept ones; Cohen's κ is low (0.29 and 0) because nearly every form passes, so chance agreement is already high, and the first judge passes every premise-kept form. On the neutral forms the first judge agrees with Gemini on 92% of its 205 items and with Luna on 79% of its 612. On the 570 items both judges pass for both forms, the cue part is 4.6 points at 42 billion tokens and 3.3 at the end, below its all-item 6.0 and 5.6, because that subset over-represents the 178 questions left unchanged, where the gap is zero by construction; on its 400 rewritten items the cue part is 8.5 and 6.9. This is an audit by models, not by people, and a human audit of a stratified sample is the check it does not replace.

B Robustness and replication detail

This appendix holds the robustness checks and replications behind Section 3, grouped by the claim each supports: the turn, what stays, and the census. OLMo-2 is the worked run because its checkpoints and data are public; each other run is here to test whether a claim travels: MAP-Neo, Viking, Tri and Pythia (other laboratories and corpora) for the gap and the items that carry it, OLMoE (a mixture of experts on OLMo-2's data) and Mamba (no attention) for the architecture, and the census for whether the turn itself generalises.

Detail for Section 3: the turn

Localizing the 7B minimum. Fourteen additional OLMo-2-7B checkpoints at roughly thousand-step resolution through the minimum region (5B to 118B tokens) resolve a single interior minimum, still at 42B tokens, with no second well anywhere in $(13, 131)$B, a descent that steepens into the minimum, and a post-minimum slope of +0.023 per decade. The 13B's dense twin (nine usable checkpoints from 9B to 101B, see below, and the 5-trillion-token endpoint) puts its lowest paired score at 59B tokens, between neighbours at 42B and 76B.

Two 13B releases carry an output head that is not their own. The branches at 34B and 51B tokens have an output head whose norm is 0.675 of the neighbouring checkpoints', and they carry that ratio identically to sixteen significant figures although seventeen billion tokens of training separate them: as if the same head were shipped in both. The 51B branch is degraded further, with an embedding four times its neighbour's norm and sixty-four tensors below half of theirs. Both score as birth states under both forms, adversarial 0.511 and 0.534

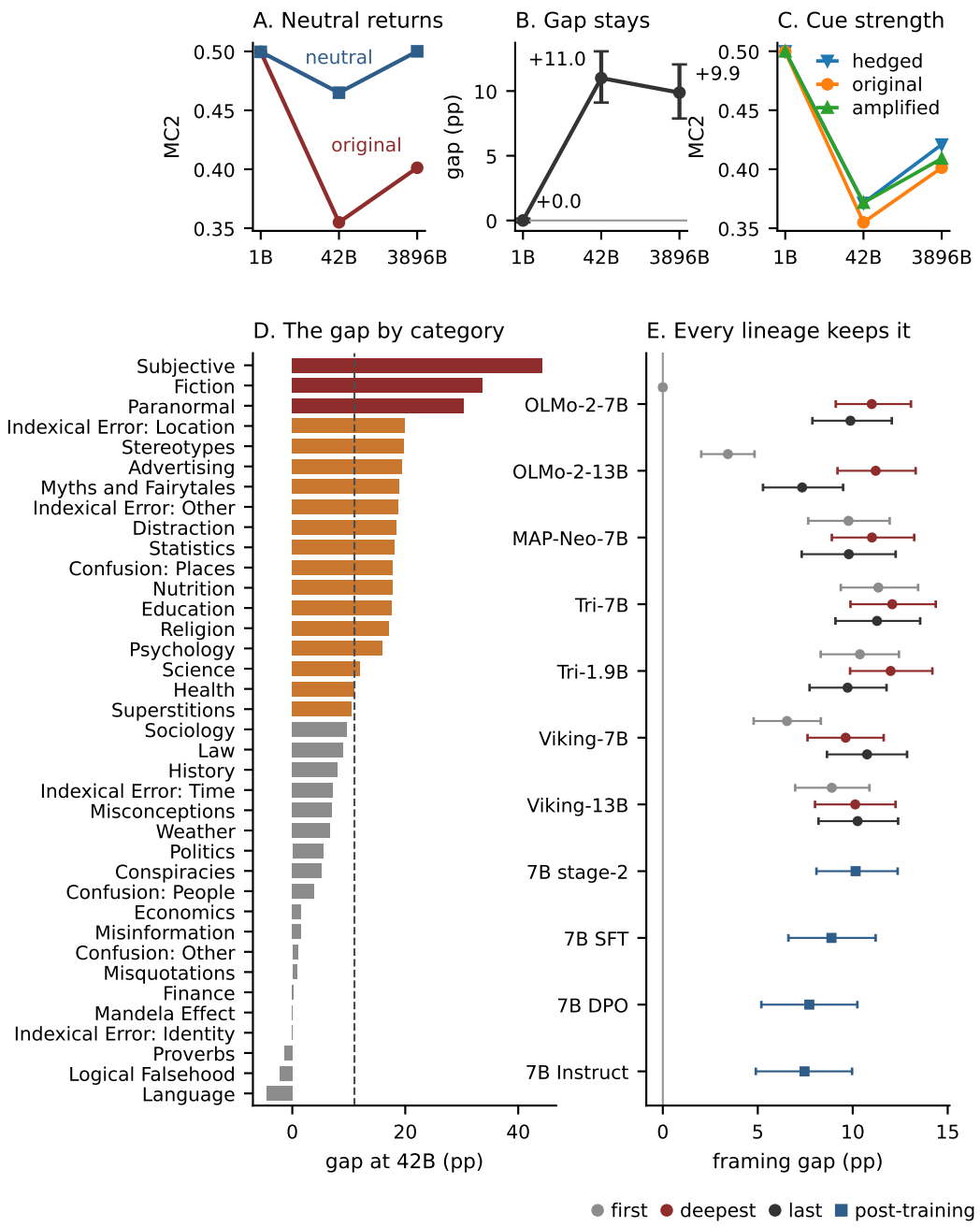

Figure 4: **The two scores, detail.** (A) The 7B's two forms at start, minimum and end. (B) The gap with item-bootstrap intervals. (C) The hedged and amplified prefixes track each other to 1.2 points, and both score above the unprefixed original, whose cue is written into the question itself. (D) The gap at the minimum for all 37 of the benchmark's categories. (E) The gap at the first, deepest and last checkpoint of the seven multi-checkpoint twin arms and at the four post-training stages; the OLMo-2-1B sibling is in Figure 2C and Pythia-1B in the text.

against the 42B neighbour's 0.373, with the two forms correlating at 0.959 where trained neighbours give 0.756, while the training loss is flat at both steps. Both are left out of every 13B statistic on that head norm. A later branch, at 93B tokens, scores 0.373 on HellaSwag

against 0.56 at 42B and is left out on that capability criterion. Nine of the twelve checkpoints released between 9B and 101B remain.

K2-65B under our instrument. LLM360's public logs for K2-65B carry a 120-checkpoint TruthfulQA MC2 series (0.420 at the start, an interior minimum of 0.335 at 22.5% of training, 0.408 at the end). We scored six of their public checkpoints under our instrument at reduced precision, a within-arm necessity at 65B; five were readable, and the released checkpoints inside the logged minimum window carry non-finite values under both 4-bit (NF4) and 8-bit loading, so the paired tests were run on the deepest readable checkpoint adjacent to the logged minimum. It is an external replication, not a census row.

The arms separate over the whole post-minimum window. At four matched pairs of checkpoints after the minimum the 7B gains on the 1B by +1.9 points at 131B against the 1B's 126B ($t = 2.0$), +5.4 at 416B against 420B ($t = 5.1$), +0.7 at 1301B ($t = 0.6$) and +8.9 at the two endpoints ($t = 8.2$), on the same 817 items and one scoring pipeline throughout. The separation grows but not monotonically, and the 1301B pair does not clear its own error.

Seed variance where it can be measured. EleutherAI released full checkpoint trajectories for nine data-order seeds and three initialization-only seeds of Pythia-160M. The seed ensemble puts the formation onset in the same checkpoint interval for eight of nine data-order seeds, a within-recipe descent-slope coefficient of variation of 8%, and data-order variance two to three times the initialization variance. At the one size where seeds exist, the formation time is the recipe's and what varies is the data order; at the sizes that recover, seed variance is unmeasured.

Detail for Section 3: what stays

The framing gap forms on a second recipe. On the Pythia-1B trajectory across seven checkpoints the adversarial and neutral instruments start statistically indistinguishable at step 1K (gap +0.6 points, item bootstrap CI $[-1.1, +2.1]$), co-descend to step 36K as the short-answer prior wears off, and then separate: the neutral score freezes in a 0.470 to 0.481 band for the remaining fourfold increase in steps while the adversarial score keeps falling, and the gap reaches +6.6 points (CI $[+4.7, +8.5]$) by step 105K.

Post-training, relative to the neutral score. The post-training gap falls from 20% of the neutral score at the stage-one endpoint (9.9/50.0) to 12% at Instruct (7.5/62.0), so part of the narrowing in points is the rising level; the scores were read with the same primer and no chat template at every stage; with the primer placed inside the Instruct model's chat template the gap is +7.4 points against +7.5 without it.

What the model says. Greedy generations on all 817 questions and their neutral twins were cached at the 1B, 42B and 3.9T checkpoints and at the released post-training stages (64 tokens, with and without the primer). A blind judge (given the question, the best answer, the popular falsehoods and the generation, with the checkpoint identity replaced by a code) labelled the same 200 items at every checkpoint, in both conditions, under the bare prompt; a second judge's labels on 240 overlapping rows agree at Cohen's $\kappa = 0.89$ (the contamination audit is in Appendix A).

At one billion tokens the model does not answer (400 of 400 non-answers). At 42B the adversarial form draws 98 falsehoods and 40 true answers of 200, the neutral form 70 and 67; at 3.9T the adversarial form draws 107 and 50, the neutral form 71 and 92. The framing gap in falsehoods is +14.0 points at the minimum (95% interval $[+7.0, +21.0]$) and +18.0 at the endpoint ($[+11.5, +25.0]$); the adversarial falsehood rate does not fall from the minimum to the endpoint (+4.5 points, CI $[-4.0, +12.5]$) while the neutral true-answer rate rises from 33.5% to 46.0%. What the model says agrees with what the likelihoods read: the knowledge returns on the neutral form, the gap stays on the leading form.

Free-form recall through the minimum. On 500 SimpleQA items generated by the OLMo-2-7B start, minimum and endpoint checkpoints and graded blind under the official

rubric with checkpoint identity withheld, recall rises monotonically through the adversarial minimum (0.0%, 1.8%, 3.0%). Parametric knowledge accumulates while the adversarial readout dips; the paired contrast is underpowered at these base rates, and we read the monotone series, not the contrast.

The gap by category. Recomputed within the benchmark's 37 categories and averaged over them, the gap at the minimum is +11.4 points against the raw +11.0 (sign-flip $p < 0.0005$), so it is not an artifact of which categories are large; it is concentrated where the popular answer is a belief (Subjective +44, Fiction +34, Paranormal +30) and near zero or below for Language and Logical Falsehood (-4.5 and -2.2; Figure 4D, descriptive, Appendix A). Per-category minima spread from 13B tokens to the endpoint itself, so the aggregate recovery is carried by some categories and absent in others.

Fitting each category's gap on the dense 7B grid with its own saturating exponential puts eleven of the twenty-nine that admit a fit at a fitting bound: more than a third of the benchmark's categories are still forming at 3.9 trillion tokens, so their time constants are lower bounds and no ordering of categories by formation speed is measurable on this run. The three that do settle fastest, Logical Falsehood, Misquotations and Economics at 2 to 5 billion tokens, are the three that barely acquire a gap at all (-2.2, 0.8 and 1.6 points), and the largest gap, Fiction at 33.7 points, is still forming at 42B.

The framing gap in further lineages, and the ladder outside this laboratory. Every arm scored with the paired instrument carries the gap on every checkpoint (Figure 4E). MAP-Neo-7B (Matrix corpus, decay window 21B to 724B tokens): +9.8, +11.0, +9.8 points while the neutral score rises from 0.516 to 0.538. Trillion Labs' 7B, on all fourteen released branches from about 160B to 2T tokens: the adversarial score sits at 0.350 to 0.365 through 600B tokens with the gap at +8.8 to +12.1, its later branches scatter between 0.36 and 0.44 (the 0.44 branch lies inside the recipe's learning-rate anneal), the last branch reads 0.375 with a gap of +11.3, and the minimum-to-last contrast (+2.5 points, paired $p = 10^{-3}$) does not survive a selection-aware null that re-selects the minimum on each draw ($p = 0.22$; the branch-to-branch jitter under a constant-learning-rate schedule is two to four points on both instruments).

Its 1.9B sibling sits between 0.33 and 0.38 on all 32 branches from 40B tokens with the gap at +8 to +12, and its first branch already carries a 10-point gap, so the ladder resolves formation at no size; the census rule labels the 1.9B a partial recovery on a single dipping branch, which reads as jitter. Capability controls rise on every Tri branch (HellaSwag 0.695 to 0.738, ARC-Challenge 0.43 to 0.50).

Viking: the turn graded by size within one laboratory. LumiOpen's Viking 7B and 13B (2T tokens each, Nordic-English-code): the 13B descends from 0.372 at 100B to 0.362 at 700B and rises to 0.392 at 2T (+3.1 points, paired $t = 3.5$, $p = 4 \times 10^{-4}$), which the whole-trajectory selection null does not admit ($p = 0.13$; a weaker null that conditions on the descent gives $p = 0.027$), so it enters Table 3 as a rise the rule does not admit; the 7B descends to 0.381 at 300B, reaches 0.425 at 1T and reads 0.392 at its 2T release, not significantly above its minimum. The gap is +8.9 to +10.3 on every 13B branch and opens from +6.5 to +10.8 across the 7B's released window while both neutral scores rise by 3.3 to 3.4 points.

The 33B, on its ten released branches (200B to 2T), descends to 0.325 at 1.2T and reads 0.369 at 2T (paired +4.5 points, $p = 3 \times 10^{-9}$; selection null $p = 0.023$), a partial recovery under the rule, while its gap grows from +9.1 to +14.1 at the minimum and reads +12.3 at 2T with the neutral score rising from 0.460 to 0.492: the ladder grades the turn by size within one laboratory, and the 33B's minimum sits at 1.2T tokens, not at OLMo-2's 42B. Capability controls rise monotonically on both (SciQ 0.75 to 0.92 at 13B).

The census instrument, and the flow of items. The paired instrument's adversarial series is the census instrument: on the 14 Tri-7B branches the two agree to within 0.0002. Read at item resolution, the net flow of items from wrong to right after each arm's minimum, summed over intervals in units of its binomial scale, is +5.6, +7.7 and +6.0 on the three

recovering arms with per-item scores (OLMo-2-7B, 13B, MAP-Neo) against +0.8 (OLMo-2-1B), +2.1 (SmolLM2), +1.6 (Viking-7B), +1.6 (Tri-7B) and +4.2 (Viking-13B, the rise the rule does not admit); no arm has every post-minimum interval positive.

Inverse-scaling tasks through the same checkpoints. The ten Inverse Scaling Prize tasks were scored on the eight OLMo-2-7B, seven 1B and six 13B core checkpoints. The three tasks that share TruthfulQA's mechanism (Memo Trap, Hindsight Neglect, Pattern Match Suppression) share its descent and not its recovery. On the 7B, Memo Trap falls from 0.614 at 1B tokens to 0.443 at 131B (-17.1 points paired, $p = 10^{-14}$) and reads 0.448 at 3.9T; Hindsight Neglect (the harness's ten-shot form) ends at 0.029 from 0.549 by a non-monotone path; Pattern Match Suppression falls from 0.557 to 0.000 by 131B and stays there (every item completes the pattern; the task saturates at the fully-fooled state). On the 13B the same three end below their start at 5T (Memo Trap -29.5, Hindsight -45 points).

Of the other tasks, NeQA and Into the Unknown descend without recovery, Sig Figs descends to 416B and recovers +2.9 points, Redefine Math (the one task with an arithmetic-knowledge channel) traces a U with a +25-point recovery, Quote Repetition rises with training, and Modus Tollens is a single global decision (every item shares its label) that reads 0 or 1. The 1B recovers on Hindsight Neglect alone (+11.4 points from 1301B). These tasks carry only the pull of a memorized continuation: each item pits a memorized continuation against an instruction and no item is answered better by knowing more, so returning knowledge has no path into their scores (Figure 5).

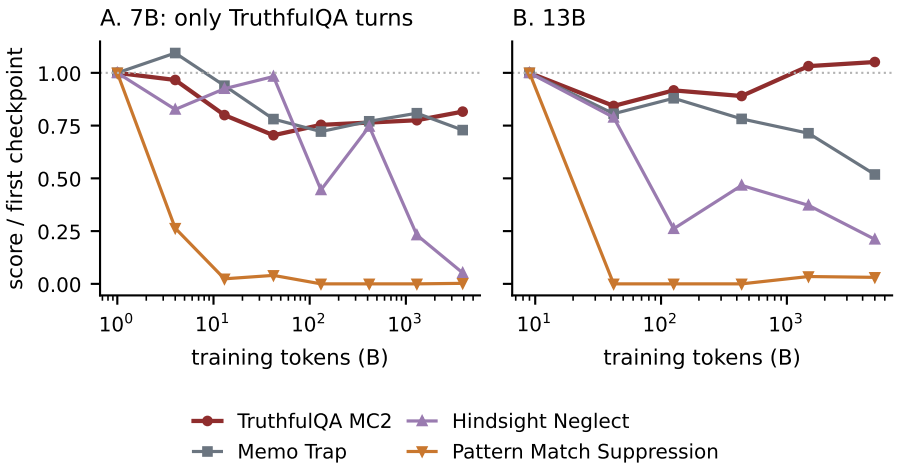

Figure 5: **The inverse-scaling boundary.** TruthfulQA and the three prior-only inverse-scaling tasks on the 7B and 13B, each as a fraction of its first value: TruthfulQA turns, the three tasks do not.

Runs trained from scratch. Five 160M models trained from scratch on one web stream (DCLM), one billion tokens each, same seed and schedule, read with the twin at 250M, 500M and 1B tokens: unfiltered, +0.8, +2.6, +2.7 points (a second seed +1.7, +2.6, +3.0); with a fixed 26-term misconception-domain document filter +1.4, +1.9, +2.3; on FineWeb-Edu +1.8, +2.8, +3.1; on Cosmopedia +1.2, +1.9, +3.6; with every document carrying two or more questions answered in the following sentence removed (33.0% of documents) +0.9, +1.6, +1.9 (second seed +1.3, +1.3, +1.6; a third seed pair reads +2.2 filtered against +1.8 unfiltered at 1B). The neutral scores sit within one point of the unfiltered arm's on every arm and seed except the genre-filtered second seed, one to two points lower.

The gap forms at a fortieth of the 7B's size from general web text on a 0.29-billion-token time constant (fitted on the unfiltered arms; a second seed placed it between 0.2 and 0.6 before it was fitted), the domain filter and the curated diets do not remove it, and the genre filter lowers it on two of three seed-matched pairs and raises it on the third, 0.6 ± 0.5 points over the three (Table 5; generator `intervention_arms_matched.py`).

The same filter was applied at one billion parameters: two models of the Pythia-1B configuration were trained from scratch on the same web stream, unfiltered and genre-filtered, at a batch of 131K tokens and read with the twin at 250M, 500M and 1B tokens; the gaps are 2.4 / 2.3 / 2.3 points unfiltered and 1.3 / 0.8 / 2.2 filtered, so the filter delays formation (paired differences -1.07 ± 0.57 and -1.57 ± 0.61 at 250M and 500M, -0.07 ± 0.52 at 1B) and both arms reach the same gap by one billion tokens; the filtered arm's training record confirms the filter (2,068,888 of 6,223,303 documents removed, 33.2%), and scored on at 2 and 3 billion tokens it reads 3.05 and 3.71 points, where the unfiltered arm was not scored.

The framing direction. The framing direction at mid-depth, fit on 400 items and read on 417 held out, narrows the gap by 0.9 points at 42B on the 7B and 2.0 on the 1B when projected out (0.4 and 0.8 at the two endpoints), almost entirely by lowering the neutral score (Section 5), and the 42B direction does not transfer to the endpoint; its per-item footprint predicts the per-item gap at $\rho = 0.21$ to 0.29.

The corpus and the items. An attribution table (infini-gram exact counts over the OLMo 2 mixture: each item's question, each of its incorrect answers, and its best answer) asked whether the per-item framing gap at the minimum is predicted by how often the item's falsehood is written. It is not, at this resolution: 63% of items have no exact hit for any incorrect-answer sentence, and across the 777 items with complete counts the gap correlates with the false-answer count at $\rho = -0.09$ and within categories at -0.05. The knowledge channel does track the corpus: the neutral score at the endpoint rises with the best-answer count ($\rho = +0.17$, $p = 2 \times 10^{-6}$).

Exact match bounds exposure from below, so the count was repeated with 3,297 key phrases drawn from the incorrect answers: the key-phrase count predicts where an item starts (birth score $\rho = -0.16$, $p = 8 \times 10^{-6}$, gone by 42B) and neither the item's gap (-0.03 at 42B, $+0.01$ at the end) nor its category's descent slope ($+0.22$ over 30 categories). Corpus exposure sets the starting point; it does not decide which items the cue moves, and the same ranking of items is susceptible in lineages trained on different corpora (Section 3).

Arm	Lineage	Corpus	Checkpoints	Outcome
Pythia-1B / 2.8B / 6.9B / 12B	Pythia	Pile	7 each	plateau at minimum
Amber-7B	LLM360	RedPajama	8	descent, sub-sig.
OLMo-2-1B	OLMo-2	OLMo 2 Mix	7	plateau at minimum
OLMo-2-7B	OLMo-2	OLMo 2 Mix	8	**partial recovery**
OLMo-2-13B	OLMo-2	OLMo 2 Mix	6	**partial recovery**
OLMo-3-7B†	OLMo-3	Dolma-3	15	**partial recovery**
OLMo-3-32B†	OLMo-3	Dolma-3	10	**partial recovery**
OLMoE-1B-7B	OLMoE	OLMo mix, MoE	7	descent, sub-sig.
MAP-Neo-7B†	MAP-Neo	Matrix	8	**partial recovery**
SmolLM2-1.7B†	SmolLM2	FineWeb-Edu	6	descent, sub-sig.
TinyLlama-1.1B†	TinyLlama	SlimPajama	6	descent, sub-sig.
BLOOM-1b1†	BLOOM	ROOTS	8	descent, sub-sig.
CT-LLM-2B†	CT-LLM	MAP-CC (zh)	8	descent, sub-sig.
Tri-7B†	Tri	Ko/En/Ja	14	floor; oscillation (anneal branch)
Tri-1.9B†	Tri	Ko/En/Ja mix	32	**partial recovery** (one branch)
Viking-7B†	Viking	Nordic+En+code	11	descent, sub-sig.
Viking-13B†	Viking	Nordic+En+code	14	rise, not admitted
Viking-33B†	Viking	Nordic+En+code	10	**partial recovery**
K2-65B (external)	LLM360	K2 corpus	5 of 120	partial recovery (qualified)

Table 3: The census: twenty-one core trajectories, twelve recipe lineages, one instrument and one rule († = arms added after the rule was fixed; the checkpoint counts are the census grids, and the twin series of Section 3 read denser grids, seventeen on the 7B and eleven on the 1B). Outcomes apply the selection-aware rule of Section 2; seven arms meet it, in five lineages. K2-65B is shown separately as an external replication and is not a census row.

Detail for Section 3: the census

The controlled cells behind the census. At matched HellaSwag the curated Qwen3 recipe sits 2.9 points below its web-trained Qwen2.5 matches and the FineWeb-Edu recipe of

SmolLM2 2.3 points above its matches (family means). Each comparison that exists fixes some variables and varies one: size at fixed recipe (the OLMo-2 sibling pair); corpus at fixed size and time (AI2's public stage-two mixture branches of OLMo-3-7B from one shared checkpoint, where the reasoning-and-code-heavy mixture lands 3.5 points below the generic one, in the expected direction); architecture at fixed corpus (OLMoE-1B-7B, one billion active parameters, ends among the non-recoveries with its active count). Two cells are empty: no run at a size that recovers has been trained twice from different seeds, and no 7B outside this laboratory has been followed with formation resolved through a multi-trillion-token budget, so dose versus recipe for Amber is an argument, not a control.

Size at the endpoint. Size sets no common trend in the endpoint gap: it falls with size in the Qwen ladders, rises in Pythia and GPT-2, and is flat within error in Llama-3 and SmolLM2 (Table 4).

Table 4: **Every base endpoint scored carries the gap.** Endpoint framing gap S in points on all 817 items and on the 380 items not filtered against GPT-3 (where the wording part is small), by family and size (billions of parameters unless marked M); item-bootstrap 95% intervals are within ±2.5 points of every whole-set value and ±3.4 of every half-set value. 32 base endpoints from 10 laboratories; the half-set interval includes zero only for Qwen2.5-7B and Qwen2.5-14B. The two Phi releases are instruct-only; OLMo-2-1B is read in float16 here, 1.4 points below its float32 series of Section 3 (Appendix A) (generator `qcensus.py`).

Family	Laboratory	Size	Gap S, all items	Not filtered (380)
Qwen2.5	Alibaba	500M / 1.5 / 3 / 7 / 14	11.0 / 9.0 / 8.0 / 4.7 / 4.3	9.2 / 7.6 / 5.7 / 1.3 / 1.6
Qwen3-Base	Alibaba	600M / 1.7 / 4 / 8	8.5 / 8.4 / 5.2 / 6.0	6.9 / 5.1 / 3.9 / 3.0
SmolLM2	Hugging Face	135M / 360M / 1.7	9.0 / 13.4 / 12.0	7.9 / 11.5 / 9.2
Gemma-2	Google	2 / 9	12.2 / 8.7	10.2 / 6.1
Gemma-3	Google	1 / 4	12.2 / 11.6	10.3 / 8.9
Llama-3	Meta	1 / 3 / 8	9.7 / 12.1 / 9.1	8.9 / 9.4 / 6.7
OLMo-2	AI2	1 / 7 / 13	10.0 / 9.9 / 7.3	7.3 / 6.6 / 4.7
GPT-2	OpenAI	124M / 355M / 774M / 1.5	4.9 / 6.2 / 7.7 / 8.9	3.5 / 5.6 / 6.9 / 8.1
Pythia	EleutherAI	1 / 2.8	6.6 / 10.1	5.5 / 9.0
Mamba	state-spaces	2.8	10.8	8.5
Viking	LumiOpen	7 / 13	10.8 / 10.3	9.6 / 8.0
Amber	LLM360	7	11.8	9.7
Phi (instruct only)	Microsoft	3.8 / 3.8	6.8 / 10.0	4.2 / 7.1

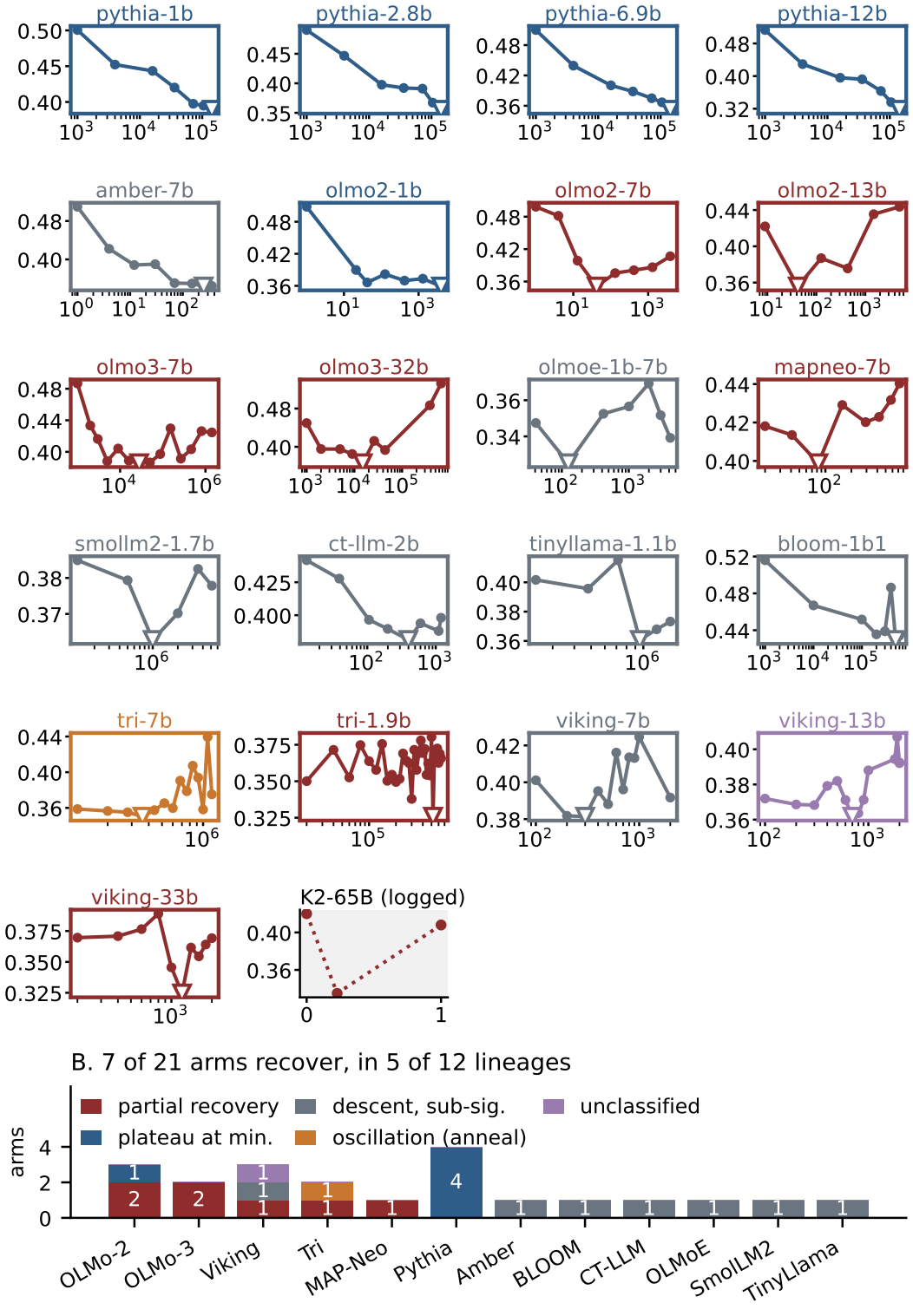

Figure 6: **The census trajectories.** (A) TruthfulQA MC2 through each arm's own training axis, colored by outcome, minimum marked; K2-65B from its logged summary, against the fraction of its run. (B) Outcomes by lineage.

C The three processes in detail

Section 3 reads TruthfulQA's trajectory on OLMo-2-7B as three processes: a short-answer prior that decays, a gap that forms and freezes, and knowledge that grows and saturates. This appendix gives the measurement behind each, and how the gap's two parts sit on the prior.

The gap's formation law. Here x is the token count and gaps are in points. On the formation side of the dense 7B twin the gap rises from zero and saturates, $S(x) = S_\infty - (S_\infty - S_0)\, e^{-x/x_s}$ with $S_\infty = 11.7$ points, $x_s = 17.5$ billion tokens and $S_0 = -0.8$, which puts the gap at zero near one billion tokens, where the two forms score alike; the law is fit from that first checkpoint, so S_0 is an extrapolation below the data, not a gap before training. It is preferred over constant growth per e-fold of tokens by ten AICc units, and the preference rests on the gap's zero at birth: predicting each withheld formation checkpoint from the other seven, the saturating law errs by 0.55 points and growth in $\ln x$ by 1.85 (the first better in 98% of item-bootstrap draws), while without the birth point growth in $\ln x$ predicts the later points as well or better (0.33 against 0.72).

A first-order form suffices: a stretched exponential fitted to the formation points has exponent 0.95 [0.39, 1.70] (1.17 [0.47, 2.52] on the 502 clean items), and fixing it at one is preferred by 9 AICc units, so no spread of rates across items is needed. A Bayesian fit of the same law gives $S_\infty = 13$ [11, 15] and $x_s = 26$ [18, 33] billion at 68%, the two correlated at $\rho = 0.80$: the timescale is less precisely identified than the ceiling. Fits use the formation side; all seventeen checkpoints give 12.3 billion. Pinning the gap to zero at birth instead of fitting an intercept is preferred by 1.6 AICc units and gives $x_s = 21.8$ [14.5, 39.2] billion tokens and $S_\infty = 12.6$ [9.8, 17.1], in line with the Bayesian fit; with the pin, the three OLMo-2 sizes read 19, 22 and 26 billion tokens with overlapping intervals.

The gap's clock across size and recipe. Fitted on each OLMo-2 arm's own formation points, x_s is 18.2, 17.5 and 18.6 billion tokens at 1B, 7B and 13B, although the three take 2.1, 4.2 and 8.4 million tokens per update; a clock in tokens per parameter is rejected once the 1B is included. OLMoE, trained on OLMo-2's mixture as a mixture of experts at its own learning rate and warmup, forms at 17.3 billion [10.7, 24.6] (the 7B pinned the same way: 21.8 [14.5, 39.2]; two formation points, so a one-parameter fit with the gap zero at birth and its plateau measured): at fixed data the clock does not change with the architecture. Amber-7B, on RedPajama, forms at 15.2 billion [9.5, 28.6] on thirteen points from zero.

OLMo-3-7B shares OLMo-2-7B's batch (4.19 million tokens per update), peak learning rate (3×10^{-4}) and warmup (8.4 billion tokens) and forms at 8.6 billion; it differs in its data mixture (76% Common Crawl with 14% science documents, 7% code and 3% mathematics, against OLMo-2's 95% filtered Common Crawl), its sequence length (8,192 against 4,096), its optimizer variant and its schedule's horizon. The faster clock is not the density of question-answer text: measured with one frozen classifier on 20,000 documents per major source, OLMo-3's mixture carries 0.87 times OLMo-2's per token, and that density orders the four laboratories' runs in no order of their clocks. The mixture's composition, the context length and the optimizer variant remain as explanations. A run card joining each of 29 runs' documented recipe (batch, learning rate, warmup, corpus) with every quantity measured on it ships with the code: no recipe column orders the clock, and corpus and size both move the depth.

Pythia, on the Pile, forms the gap at all three sizes read, to about half OLMo-2's depth by 42 billion tokens (3.8, 4.6 and 5.8 points at 160M, 410M and 1B, ±0.9); its grid does not pin a clock (the estimate moves with the fitting span), and its gap keeps deepening slowly after 42 billion tokens, to 4.0 to 6.9 points at 300 billion. A second 410M run on the deduplicated Pile forms the same gap (within 0.2 points at 42 billion tokens), so repeated documents do not set its depth.

At 160M, trained from scratch on one web stream, the gap forms with a 0.29-billion-token constant (a second seed: 0.2 to 0.6). On a billion-parameter run trained for this paper at a batch of 131K tokens, formation was complete by step 2,000 (250 million tokens), a clock in

optimizer updates gives 2.5 points and the run reads 2.4; where the clock counts updates and where tokens is taken up in Appendix E.

The prior's clock. The neutral score on the 353 items whose true option is the shortest decays with its own constant. On OLMo-2-7B it is 5.7 billion tokens [4.5, 7.2], against the gap's 17.5 [11.8, 32.2]: the prior leaves about three times faster than the gap forms, and the two are separate processes there. On OLMo-2-1B the two are 9.9 and 16.9 billion with overlapping intervals, and in the small Pythia runs they share a clock (5.1 against 6.2 billion at 160M, 3.7 against 4.3 at 410M). The whole neutral score's own saturating fit has a 3.6-billion-token constant, because it mixes the fall on these items with the rise on the others.

The fall is a population effect, not item by item: the 178 items whose rewrite leaves the question unchanged, and which therefore carry no gap, lose neutral score through formation at -1.6 [$-3.5, +0.3$] points per e-fold against -1.2 [$-2.3, -0.05$] on the clean items, and an item's early loss is uncorrelated with its own gap growth (Spearman 0.05, and 0.02 with its growth over 9 to 42 billion tokens) and with its later return (-0.05). Split by the shortcut, what remains of an early fall of knowledge is about two points inside the formation window in three of ten runs (two of the five 160M runs and Pythia-1B) and none on OLMo-2.

Knowledge. On the 464 items the prior does not favour, the 7B's neutral score rises from 19.2 points at one billion tokens to 29.9 at 17 billion and 35.9 at 3.9 trillion. A saturating law describes it better than growth in $\ln x$ (AICc 15.1 against 23.0; a power law 27.4), with a plateau of 32.4 points reached with a time constant of 10.9 billion tokens; the endpoint's 35.9 is a second, slower rise. Scored by PMI, which removes each option's advantage without the question, the early rise is larger (45.2 to 58.2 points by 9 billion tokens), so it is not the prior's length advantage leaving. After 42 billion tokens it keeps rising on the 7B, by 1.0 [0.5, 1.6] points per e-fold, and stays flat on the 1B (-0.2 [$-0.7, +0.3$]); that is where the siblings part, and the 1B's plateau at about 46 points of the whole neutral score is its own ceiling, not a failure to reach the 7B's. Between 118 billion and 1.3 trillion tokens the 7B's whole neutral score stays within 47.2 to 48.9 points while the gap stays within 9.0 to 10.8, the two rising together at 416 billion and falling back; a return flat until late and one spread evenly in $\ln x$ both fit, so at 817 items the timing of the late return is not resolved.

The gap's two parts, split by the prior. On the premise-kept protocol's five checkpoints, the cue's own part is the same size whether or not the true option is the shortest: 6.1 [3.1, 9.6] and 5.9 [3.8, 7.9] points at 42 billion tokens (difference 0.3 [$-3.4, +4.3$]), 5.8 and 5.5 at 3.9 trillion. It forms earlier where the prior is absent (4.0 points at 9 billion tokens there, -1.6 on the shortcut items, 1.8 at 17 billion). The wording part lives on the shortcut items: 7.2, 9.1 and 12.5 points at 9 billion, 42 billion and 3.9 trillion tokens, against -0.7, 1.0 and -1.6 on the rest. This is not TruthfulQA's adversarial split in disguise (the shortcut items are 46% of the filtered questions and 40% of the others): at 42 billion tokens the wording part is 10.7 and 7.0 points on the shortcut items of the two halves and 2.4 and -0.6 on the rest, so the shortcut adds about eight points within each half and the filter about three within each stratum. Measured directly at the endpoint, the cue part is 5.0 points on Llama-3.1-8B, 5.4 on Pythia-2.8B, 6.6 on Mamba-2.8B and 2.9 on Qwen2.5-7B (5.6 on OLMo-2-7B). The cue part also sits on largely the same items across the five models where it was measured directly (OLMo-2-7B, Llama-3.1-8B, Qwen2.5-7B, Pythia-2.8B and Mamba-2.8B): per-item Spearman 0.32 to 0.64 between every pair, highest for Pythia-2.8B against Mamba-2.8B (0.63), two architectures trained on the same corpus. By 3.9 trillion tokens the whole gap is 18.3 points on the shortcut items and 3.9 on the rest.

Word overlap and the wording part. TruthfulQA's false options often repeat words of the question, and MC2 sums an option's log-likelihood, so an option that repeats the prompt may gain from it. Per item, let the overlap advantage be the share of the false options' word bigrams found in the question minus the same share for the true options; the shortcut items carry most of it (0.157 against 0.038), and the premise-kept rewording removes most of theirs (0.091 against 0.012). The overlap each rewording removes tracks the item's wording part (Spearman 0.26 [0.19, 0.33] at 42 billion tokens and 0.34 at 3.9 trillion; 0.21 with the

option-length ratio, question length and its change, option count, mean option length and 39 category fixed effects held fixed; unigram and weighted-bigram overlap give the same; by tercile of overlap removed the wording part is −1.6, 3.3 and 11.2 points at 42 billion tokens). This is an association, not a split of the wording part. The cue part does move with the difference between the two rewrites' overlap, item by item (12.2, 5.3 and 0.3 points where the neutral rewrite removes more, the same or less overlap than the premise-kept one); the effects cancel on average, and on the 246 items where the two leave the same overlap the cue part is 5.3 points at 42 billion tokens, against 6.0 on all items.

The four curves do not share one shape. Fitting the prior, knowledge on the no-shortcut items, the wording part and the cue part with one sigmoid in $\ln x$ each, a shared shape is rejected against separate shapes (likelihood ratio $p < 10^{-4}$ on the item-bootstrap covariance; twenty points, so read it as descriptive). The two parts of the gap switch on sharply, and under the summed score in order (midpoints 7.2 and 16.9 billion tokens, widths under half an e-fold; under PMI the cue part is most of the way formed by 9 billion tokens, with the wording part), as stage-wise learning of single modes would (Saxe et al., 2014); the prior decays over about one e-fold (midpoint 2.5 billion); and knowledge rises slowly, nearly linearly in $\ln x$, as a sum of item-level steps: of the 327 no-shortcut items whose neutral score moves by 20 points or more, two thirds are better fit by a single step than by the common course (a seventh by chance), at times spread about evenly in $\ln x$ between 5 and 70 billion tokens (quartiles 6.7, 21 and 67 billion), as in the quanta picture of scaling (Michaud et al., 2023). The gap's parts behave like single modes; knowledge like many. The five-checkpoint grid resolves the parts' widths only as upper bounds.

The two components. The two item-space components of Section 3 have their own rates. Writing the residual's leading coordinates as S_1 and S_2, the dominant one saturates, $S_1 \propto 1 - e^{-x/x_1}$ with $x_1 = 10.8$ billion tokens, and the minor one forms and reverses, $S_2 \propto a\,(1 - e^{-x/x_f}) - b\,(1 - e^{-x/x_r})$ with $x_f = 8.2$ billion, $x_r = 41.6$ billion and $b/a = 1.4$, crossing zero near 53 billion ($R^2 = 0.965$ on S_1, 0.983 on S_2, 0.961 on the aggregate). The aggregate gap is their sum; neither form is derived, and both are the minimal shapes for a coordinate that saturates or one that forms and reverses. They carry two observations: the 13B's gap closes more after its minimum (11.2 to 7.3 points, against 11.0 to 9.9 on the 7B) because its minor component reverses further, and the 1B reverses it with no knowledge rising.

Checks on the split. On the one plateau arm with dense branches, Tri-1.9B, the gap's slope across a 32-fold range of steps is +0.08 points per e-fold (CI $[-0.35, +0.52]$). Through post-training the 7B's four stages lie near a line of nearly constant gap in the (K, S) plane (10.2, 8.9, 7.7 and 7.5 points) while K rises 12 points. Recomputed in log-odds space, the gap and neutral-score trajectories keep their shapes ($r = 0.99$ and 0.93 with the point-space versions) and the minimum falls at the same checkpoint, so the split is not an artifact of MC2's compression near 0.5.

D Using the instrument

Each use below is one command and one reading of the two channels it returns. The first three are uses during development; the rest give the detail behind Sections 4 and 5.

Data-curation audit. A 160M model trained from scratch on one web stream for one billion tokens carries a gap of 1.8 to 3.0 points over three seeds, a quarter of the 7B's. Filtering the stream by topic (a fixed list of 26 misconception-domain terms) leaves it at 2.3; the curated diets carry 3.1 (FineWeb-Edu) and 3.6 (Cosmopedia), one seed each. Removing the question-and-answer genre by a syntactic proxy, a third of the documents, lowers it on two of three seed pairs and raises it on the third (Table 5). Dropping a third of the documents at random instead leaves the gap where it was (+0.2 $[-0.3, +0.8]$ over the three seed pairs), and the genre filter sits 0.9 $[0.3, 1.5]$ points below it: the filter acts through the genre, not through the amount of data.

Intervention	Model, items	Untreated	Treated	Change (points)
Decoding				
Layer-contrastive decoding (DoLa), 7B endpoint	7B, 200	9.04	9.13	$+0.09 \pm 0.33$
One direction removed				
At inference, one layer, 7B endpoint	7B, 417	9.87	9.50	-0.37 $[-0.84, +0.09]$
At inference, one layer, 7B at 42B	7B, 417	9.95	9.00	-0.94 $[-1.42, -0.50]$
At inference, one layer, 1B at 42B	1B, 417	8.01	6.01	-1.99 $[-2.86, -1.18]$
At inference, one layer, 1B endpoint	1B, 417	9.97	9.19	-0.78 $[-1.34, -0.21]$
Data				
Question-answer genre filter, three seeds	160M, 817	2.70, 2.97, 1.85	1.86, 1.60, 2.16	-0.85, -1.38, $+0.32$ (± 0.5 each)
A random third of the documents dropped, three seeds	160M, 817	2.70, 2.97, 1.85	2.32, 2.80, 3.09	-0.38, -0.17, $+1.24$
The same filter at 1B, at 250M, 500M, 1B tokens	1B, 817	2.38, 2.34, 2.28	1.31, 0.76, 2.21	-1.07, -1.57, -0.07 (± 0.6)
An instruction prefix on every training document, one seed	160M, 817	2.70	1.92	-0.78 $[-1.59, +0.06]$
One direction removed from the weights (every layer)				
Output matrices, rank one	160M, 717	1.57	1.06	-0.51 ± 0.21
All matrices, rank one	160M, 717	1.57	0.93	-0.64 ± 0.22
Fine-tuned on leading-question answers, 100 steps (cross-entropy plus KL, weight 1)				
Trained items	160M, 100	3.83	0.64	-3.19
Held-out items	160M, 717	1.57 ± 0.83	0.70 ± 0.83	-0.87, inside error
Trained items	1B, 100	11.72	-0.99	-12.71
Held-out items	1B, 717	10.12 ± 1.13	2.50 ± 0.63	-7.62 ± 0.97 (75%)
The same, cross-entropy only (KL weight 0)				
Trained items	1B, 100	11.72	-5.55	-17.27
Held-out items	1B, 717	10.12 ± 1.13	2.88 ± 1.14	-7.23 ± 0.97 (72%)
The same two objectives on 100 new training items (1B)				
Held-out, cross-entropy only	1B, 717	10.76 ± 1.17	3.92 ± 1.13	-6.85 (64%)
Held-out, with KL	1B, 717	10.76 ± 1.17	4.34 ± 0.78	-6.43 (60%)
The KL term alone, no answer supervised (set 1)				
Held-out items	1B, 717	10.12 ± 1.13	0.82 ± 0.43	-9.29 (92%)
Held-out, 100 new training items	1B, 717	10.41 ± 1.15	1.32 ± 0.33	-9.09 (87%)
Answers after the leading or the plain question (repeat run, own baseline)				
Held-out, leading form	1B, 717	10.96 ± 1.16	2.65 ± 1.11	-8.31 (76%)
Held-out, plain form	1B, 717	10.96 ± 1.16	7.48 ± 1.00	-3.48 (32%)
The framing direction projected out throughout pretraining (three seed pairs)				
Held-out items at 1B tokens	160M, 717			$+0.21$ (-0.07, -1.15, $+1.85$)

Table 5: What each intervention does to the framing gap, against an untreated arm with the same stream, seed and items. $\pm$ is the paired item-level standard error where per-item scores exist; it excludes the seed-to-seed spread of about half a point. Rows without a paired error store means only or read against in-run baselines; the pretraining projection's change is per seed pair; the instruction prefix ("Answer factually. Do not repeat popular misconceptions." before every document) is one seed, paired over items at 1B tokens, with the neutral score unchanged.

At one billion parameters it delays formation without lowering where it ends: the filtered arm reads 1.3 and 0.8 points where the unfiltered reads 2.4 and 2.3 at 250 and 500 million tokens, both read 2.2 to 2.3 at one billion, and the filtered arm reads 3.0 and 3.7 at two and three billion, where the unfiltered arm was not scored. The cue rides on the form of the text more than on its subject. The command `twin score <model> --revision <ckpt>` on each diet arm reads the gap that will be inherited at scale before the scale is paid for.

Reading the scar. `twin scar <birth> <formed> <end> --channel placebo` takes three score files of one run and returns Σ at the formed checkpoint and at the end, its persistence π with an item-bootstrap interval, and its coverage and depth: the items whose cue part at the formed checkpoint clears the noise at birth, read at the end so that item noise does not inflate them. On OLMo-2-7B it returns the numbers of Section 3; without `--channel` it reads the whole gap S the same way, for a benchmark paired by construction.

Post-training regression check. OLMo-2-7B's four post-training stages move K by 12 points (0.500 to 0.620) and S by 2.4 (9.9 to 7.5, with a confidence interval that includes zero). `twin score <model> --revisions <stage1> <stage2> <sft> <dpo> <instruct>` reads both channels across the stages of a release and says which one each stage moved.

Recipe or architecture comparison. At matched HellaSwag, the curated Qwen3-4B carries a gap 2.9 points below its web-trained Qwen2.5-3B match. Mamba-2.8B, with no attention, carries 10.8 points at its only released checkpoint, and OLMoE-1B-7B, a mixture of experts, 13.1 at its endpoint (10.2 to 13.1 across seven checkpoints). The command reads two runs at matched capability on the same 817 items, and the split says whether a change moved the gap or only the knowledge.

A gap that narrows for the wrong reason. Blocking attention heads on the 7B, one at a time, moves the gap by at most 0.8 points at 42B tokens and 2.1 at the end, and the ten heads that each narrow it most, blocked together, narrow it from 11.1 to 6.7 points at 42B and from 11.3 to 5.3 at the end (100 items, on which the heads were also selected). The two channels say what happened: at 42B the neutral score falls 4.3 points and the leading-question score does not move, and at the end the two fall 8.1 and 2.1 points. The blocked heads carry knowledge, and the gap narrows because the neutral form loses it. A benchmark score would have read the same intervention as a model that became harder to fool; the ten heads also share only one in nine members between the two checkpoints.

Read on the 317 items disjoint from the selection sweep, most of that narrowing is the selection: the same ten heads lower the neutral score by 1.6 points and the gap by 2.2 at 42B, and at the end lower the neutral score by 2.6 and the leading-question score by 3.9, so the gap does not narrow ($+1.3$, interval $[-0.7, +3.4]$). Ten heads chosen on the selection items to raise the leading-question score while leaving the neutral score within half a point raise it on the held-out items by 1.1 points $[0.1, 2.1]$ at 42B, mostly through the neutral score ($+0.7$) with the gap moving -0.4; at the end no such set appears ($+0.2$, $[-1.1, +1.6]$).

A second benchmark, read as released. SycophancyEval (Sharma et al., 2024) ships each trivia question with three cued variants, and serves here as an example of a benchmark paired by construction: `twin score <model> --items sycophancy_twin.jsonl --revisions <ckpt1> <ckpt2> ...` reads the converted file with no further preprocessing, the bare question as the neutral form. On the suggest-incorrect variant over 1,813 items and the seventeen dense OLMo-2-7B checkpoints of TruthfulQA's grid, the gap is +0.2 at 1B tokens, 33.9 at 5B and 46.6 at 9B, holds at 48.0 [46.4, 49.7] from 42B to 118B, and reads 52.3 at 3.9T; the bare-question score dips to 0.542 at 9B, near the end of the learning-rate warmup, and rises to 0.698.

Four more checkpoints inside the rise show a delayed onset: the gap is 5.2 points at 2.5B tokens, 18.3 at 2.9B and 30.4 at 3.6B. A saturating fit gives a constant of 3.6 billion tokens [3.4, 3.9]; a law with an onset lag fits better (by 34 AICc units; lag 2.5 billion tokens, then a rise steepest at its start), where TruthfulQA's gap rises smoothly from the first billion tokens. The rise is items switching one at a time: at 2.9B tokens 12% of the items already defer fully (gap above 75 points), 56% not at all, and a quarter sit in between at every checkpoint.

The two variants that mention the correct answer raise the bare score by 41 points at 9B and by 21 and 26 at the end, but measured against the room left above the bare score, their help and the wrong suggestion's harm stay within 0.86 to 1.08 of each other at every checkpoint: one responsiveness answering a suggestion of either sign, the pretraining counterpart of the single response factor that invariance training scales (Section 5). Early on the raise does not depend on the user's stance: naming the correct answer raises it by 40.7 points at 9 billion

tokens when the user affirms it and 40.6 when the user doubts it, and the difference grows only to 2.9, 3.3 and 5.3 points at 26 billion, 42 billion and 3.9 trillion tokens (of a raise of 38.5, 36.0 and 26.4): the base model mostly copies a named answer. The tool reports all three variants.

The directions. `twin direction <model> --layer L --items <set>` returns the mean hidden-state difference between the two forms at one block, its norm, its sign consistency across items, and its excess over a paired sign-flip null. On the OLMo-2-7B endpoint at layer 8 (400 items each) the framing direction reads $z = 28.8$ and the sycophancy direction $z = 38.4$, at a cosine of 0.34 (70 degrees); on Qwen2.5-3B at layer 9 the angle is 76 degrees and on Gemma-2-2B at layer 6 it is 70 (200 items each). The presupposition direction on the 7B (FalseQA items, layer 8) is at 94 degrees to the framing direction and 92 to the sycophancy direction. The direction is significant on Mamba-2.8B at layer 16 ($z = 21.8$) and on OLMoE-1B-7B at layer 4 ($z = 13.6$, 100 items).

The interventions, in detail. Every row of Table 5 is a paired per-item difference against an untreated arm that shares the stream, the seed and the items. At inference on the 7B endpoint, layer-contrastive decoding at layer 8 (`twin score --dola 8`) reads adversarial 0.406 and neutral 0.497, a gap of 9.13 points against 9.04 under standard decoding on the same 200 items; on SycophancyEval it reads 16.0 against 16.1. Projecting the framing direction out of the residual stream at one layer removes 0.9 and 0.4 points on 417 held-out items at the 7B's minimum and end, and 2.0 and 0.8 on the 1B, and the direction read at 42B does not transfer to the endpoint (Appendix B).

One direction removed from the weights. Projecting the direction out of a finished 160M model's weights (generator `weight_surgery.py`; directions from the first hundred items in the space each matrix reads or writes, scored on the other 717) narrows its held-out gap of 1.57 points when written into every layer's output matrices (-0.51 ± 0.21) or into all of them (-0.64 ± 0.22) while raising both scores (the leading-question score by 0.8 and 0.9 points), does nothing through the query and key inputs ($+0.05 \pm 0.05$), and removes no more at rank four or eight (-0.52, -0.47): in the weights the projection moves the model the way training does, and a wider subspace buys nothing.

Fine-tuning on leading-question answers. The fine-tuning (generator `contrastive_ft_v2.py`) trains a finished model on the first hundred items with the ordinary cross-entropy on the leading question plus a symmetric KL divergence between the two forms' answer-option distributions (AdamW, one item per step), then scores the same hundred items and the 717 never trained on. At 160M (learning rate 5×10^{-6}), untreated, the hundred read +3.83 points and the held-out set $+1.57 \pm 0.83$. At KL weight 1 the in-sample gap reads +2.17 after 50 steps and +0.64 after 100 (83% removed); at weight 2 and 100 steps +0.49 (87%). The held-out gap reads +0.35, +0.70 and +1.20 (± 0.8) on the same three runs, all inside the untreated value's error, so transfer is not measured at this size.

The same at one billion parameters. On OLMo-2-0425-1B (learning rate 2×10^{-6}, same items and steps) the hundred training items read +11.72 untreated and -0.99, -0.43 and +0.67 after (KL weight 1 and 2 at 100 steps; weight 1 at 200), and the 717 held-out items read $+10.12 \pm 1.13$ untreated and $+2.50 \pm 0.63$, $+3.11 \pm 0.59$ and $+3.58 \pm 0.85$ after: 75%, 69% and 65% of the held-out gap removed, with the neutral score rising from 0.476 to 0.550 and the leading-question score from 0.374 to 0.525 at the first setting.

Cross-entropy alone. With the KL weight at zero, cross-entropy alone on the same items, steps and learning rate, the hundred training items overshoot to -5.55 and the 717 held-out items read $+2.88 \pm 1.14$ (72%), with the neutral score at 0.568 and the leading-question score at 0.539: the held-out effect is the cross-entropy's, and the consistency term's work is to hold the trained items level.

Transfer to another cue, and the cost. The saved 1B model at KL weight 1 (both terms) was also read on other benchmarks against the untreated model (paired, acc_norm; generator `consistency_ft_inside.py`): HellaSwag -0.05 ± 0.22 points on 2,000 items, ARC-Challenge -2.0 ± 0.6 and ARC-Easy -2.5 ± 0.4; on SycophancyEval, which it was not trained on, its suggest-incorrect gap falls from 53.1 to 37.6 points (-15.5 ± 0.7, 1,813 items), and its response to a suggested correct answer falls as well, so the treated model defers less in both directions. The cross-entropy-only model, read the same way, costs about the same (HellaSwag $+0.05 \pm 0.24$, ARC-Challenge -1.8 ± 0.8, ARC-Easy -2.6 ± 0.5) and keeps its framing direction (cosine 0.77 or more at every layer), but moves the SycophancyEval gap by only -1.3 ± 0.5: on the trained benchmark the effect is the cross-entropy's, and the transfer to another cue is the consistency term's.

The consistency term alone. Trained alone, with no answer supervised, the KL term closes 92% of the held-out gap, raising the leading score by 11.0 points and the neutral score by 1.8, and narrows the SycophancyEval gap from 53.1 to 15.7 points (-37.5 ± 0.9); the bare score falls from 0.636 to 0.607, and the response to a suggested correct answer falls by the same factor as the response to a wrong one (0.25 to 0.31 against 0.29). Split as in Eq. 2 on the 717 held-out items, the untreated gap of 10.1 points is 3.3 of wording and 6.8 of cue; answer training leaves -0.1 and 3.0, answers with the KL term -0.2 and 2.7, and the KL term alone -0.4 and 1.2; on the second training set the KL term alone costs 2.0 points of neutral score. On 800 FalseQA pairs the untreated model's resistance to a false premise (the paired log-mass of Section 4) is 0.69 nats; answer training lowers it by 0.20 (± 0.02), answers with the KL term by 0.37 (± 0.02), and the KL term alone by 0.68 (± 0.03), to 0.01: the model stops telling the two apart.

Where the remaining gap sits. Split by rewrite type, what the KL term leaves is not where the neutral form adds content: on the rewrites that add a debunking phrase it closes 94% of the gap on both training sets, and does so by lowering the neutral score (by 7.2 and 12.9 points), so the model heeds the debunking phrase less as well; 73 to 83% of the remaining gap sits on the clean rewrites, and on items left verbatim, which carry no gap, the neutral score rises 6 to 9 points (both arms rescored on one protocol).

A second training set. On a disjoint draw of one hundred training items (the other 717 held out), cross-entropy alone closes 64% of the held-out gap and the objective with the KL term 60% (Table 5), and again the KL term holds the trained items level (-1.5 against -3.5 points).

Which components read the cue. Ablating one attention or MLP block at a time of OLMo-2-7B (64 ablations, 100 TruthfulQA items) and reading both channels shows which components the cue and the answer share. At 42 billion tokens, while the gap forms, the ablations that narrow the gap lower the neutral score with it (Spearman 0.72 across ablations, permutation $p < 0.001$), at 0.46 points of neutral score per point of gap through the origin and 0.60 without the first block; at the end of training the relation weakens ($\rho = 0.40$) and its slope is 1.1. The same 64 ablations on 100 SycophancyEval items at the end give a slope of -0.22 ($\rho = -0.21$): narrowing the deference gap costs the bare-question score nothing (generators `micro_inversion.py`, `syc_route_ratio.py`). The heads whose blocking narrows the TruthfulQA gap are not the model's induction heads (mean induction score 0.018 for the top ten against 0.050 for the other swept heads; `micro_induction.py`): on this cue the readout is collective, not a copying circuit.

E Mechanism and timing detail

The measurements below sit behind single sentences of Sections 3 to 5. Each is descriptive unless it says otherwise.

The early dip, and the heads that form with the gap. Every cue-free score we read falls early, in different intervals: TruthfulQA's neutral score steps down between 5 and 9 billion tokens and stays down through 42 billion; SycophancyEval's bare question

dips shallowly (0.601 at birth, 0.542 at 9 billion tokens, 0.698 at the end); ARC's dips at 5 billion tokens (18.7 to 15.7) and then rises with growing speed. None returns above its 9-to-17-billion-token chord by 13 billion tokens, so the falls are not a shared transient at the end of warmup (8.4 billion tokens). Where we checked, the fall is the short-answer prior: split by whether a true option is the shortest, the 7B's TruthfulQA neutral score loses 19.1 points on those items from 1 to 9 billion tokens and gains 6.7 on the rest, OLMo-2-1B and three Pythia sizes split the same way, and SycophancyEval's bare question loses 14.6 on items whose true answer is the shorter one while gaining 7.1 on the rest (30.8 over stage one).

Induction heads (Olsson et al., 2022) form on the 7B between 2.5 and 3.6 billion tokens (none scores above 0.3 at 2.5 billion, 16 do at 3.6 billion and 26 at 4.2 billion), the window in which the suggested-answer gap switches on; the gap leads slightly (18 points at 2.9 billion tokens, with no head yet above 0.3). From 2.9 to 4.2 billion tokens, ablating the attention of the layer that holds them (layer 9) removes a tenth to a fifth of that gap, and its forming induction heads are among its strongest single carriers; by 9 billion tokens the layer carries none of it, and the gap rests on early-layer heads mostly different from those at onset (one or two of the ten strongest shared, against about six when the same checkpoint's items are redrawn; one model, 100 items).

At 160M, where warmup ends at 26 million tokens, the untreated run's neutral score holds at 52 points through 150 million tokens and falls 3.5 by 250 million as the framing gap begins to form (5.1 on the short-answer items and 2.2 on the rest, the one window where the items without the shortcut also fall), and has not returned by one billion; on a schedule twice as long it falls in the same window and has not returned by two billion, so the fall is not the learning-rate decay.

Induction heads appear between 100 and 150 million tokens, with the gap's first rise, and carry two-fifths of the gap while it forms (250 million tokens; beyond each of 50 random sets of 13 heads) and nothing distinctive once it has formed (one billion), the hand-off seen on the 7B, as in-context learning that emerges early later yields to in-weights solutions (Singh et al., 2023); blocking them does not return the neutral score lost at 250 million. Pythia-410M shows the same order: its heads appear between 512 and 1,000 updates, at the start of its gap's rise, and its neutral score falls after.

The plateau before models learn facts coincides with attention circuits forming (Zucchet et al., 2025), and in masked language models a syntactic attention structure appears abruptly and competes with other abilities while it forms (Chen et al., 2024); split by the shortcut, the early fall is mostly the surface prior eroding. The obvious test of the heads' role, delaying them by training on a short context with the data unchanged, cannot answer it here: under a 64-token context the model cannot read the multi-shot primer, so every in-context score stays at its untrained level until long context returns, although induction heads form anyway (eight above 0.3 by 300 million tokens, eleven by 500 million) and the gap stays within 0.7 points of zero until the context is restored (600 million tokens, then 1.7 points at 650 million). Induction heads alone do not produce the in-context gap; timing them against it needs a score the model can read without the primer. Scored that way (812 items fit in 64 tokens), the gap takes a first step of about 1.4 points as the induction heads form in both runs (at 150 million tokens in the control, at 250 million under the short context) and reaches its full size of about 3.5 points only after long context returns; one seed, with steps of about 2.5 standard errors.

Where each cue is read. Per-layer ablations of the 7B at 42 billion tokens narrow the framing gap and lower the neutral score together, at 0.46 $[0.35, 0.73]$ points of neutral score per point of gap (Table 1); the same ablations cost the bare question nothing on SycophancyEval (-0.05 while its gap forms) and cost 2.7 points per point of gap for the ARC distractor at 399 billion tokens (0.4 to 2.7 over items). On OLMo-3-7B, read at the same phase of formation (2.4 formation constants, 21 billion tokens), the ratio is 0.23, though on 100 items its interval (-0.5 to 1.0) does not exclude OLMo-2's. The framing direction's overlap with the model's truth direction is small (formation-weighted cosine -0.06 to -0.16 over five layers).

Projection and invariance, in detail. Projecting the framing direction out at inference narrows the gap only by lowering the neutral score; removing the direction's truth component first still lowers it nearly as much (2.1 against 2.3 points on the 1B at 42 billion tokens), holding the component at its neutral mean instead of zero lowers it by 2.0, and a random direction moves neither score by a tenth of a point. Distributed encoding, or later layers compensating (McGrath et al., 2023), remain possible, and a rank-one projection is narrower than full linear erasure (Belrose et al., 2023).

Answer fine-tuning leaves the framing direction in place (cosine 0.77 or more with the untreated one at every layer). The consistency objective trained alone multiplies the response to a suggestion by about one factor whatever the suggestion says (0.29 for a wrong answer, 0.25 and 0.31 for a correct one doubted or affirmed; 0.97 for answer training alone); it keeps the framing direction in the first eight layers (cosine 0.69 to 1.00) and turns it away in the last seven (0.30 to -0.11), so the frame no longer reaches the layers where answers are read. Applied throughout pretraining at 160M, the projection does not stop the gap, and the direction measured after training is nearly orthogonal to the one removed, as concepts removed from a trained model relocate under retraining (Lo et al., 2024).

The framing direction turns while the gap holds. Between the minimum and the end the framing direction turns by 63 to 78 degrees (split-half noise 9 to 12). The two item-space components of Section 3 account for most of it: projected onto the residual stream they weigh 1.2 to 1.8 to one and are nearly orthogonal, which predicts 48 to 63 degrees, so the reversing component turns the direction while barely moving the aggregate gap.

The suggested-answer gap late in training. The gap holds near 48 points from 42 to 118 billion tokens ($+0.7$ $[-0.3, +1.7]$ per e-fold) and then deepens, $+4.3$ $[3.0, 5.6]$ points by 3.9 trillion. From 9 to 118 billion tokens the bare score gains 7.3 points and the score under a wrong suggestion 5.9; from 118 billion to the end the bare score gains 8.3 and the suggested score 4.0. The late deepening is knowledge that arrives late and yields to a suggestion, the override of learned knowledge that patching locates in trained models (Wang et al., 2026).

A recovered mean over a changed population. The 7B's items do not return with its mean: the 146 items whose score never recovers widen from 19.0 to 26.4 points, and the scar sits on a third of the items (Section 3). The human analogue is the dissociation between knowing and being led: corrected misinformation keeps influencing later reasoning (Lewandowsky et al., 2012), and repetition raises the judged truth of statements people know to be false (Fazio et al., 2015).

Two clocks for formation. Within OLMo-2 the formation constant is the same in tokens at 1B, 7B and 13B while the three take 2.1, 4.2 and 8.4 million tokens per update; at 160M, where the batch can be varied on one stream and seed, doubling or halving it moves the gap at a fixed token count by -1.1 ± 0.6 and $+1.1 \pm 0.5$ points while both arms sit within 0.6 points of the control at the same number of updates, and a second seed agrees. The induction heads that accompany the gap's onset form within about a thousand optimizer updates in all three runs where we timed them (763 to 1,144 at 160M, 512 to 1,000 on Pythia-410M, 601 to 851 on OLMo-2-7B), across a thirty-fold range of tokens; an observation, not a test.